

Large Sample Properties of Matching Estimators for Average Treatment Effects

Alberto Abadie – Harvard University and NBER

Guido Imbens – UC Berkeley and NBER

First version: July 2001

This version: January 2004

A previous version of this paper circulated under the title “Simple and Bias-Corrected Matching Estimators for Average Treatment Effects”. We wish to thank Don Andrews, Joshua Angrist, Gary Chamberlain, Geert Dhaene, Jinyong Hahn, James Heckman, Keisuke Hirano, Hidehiko Ichimura, Whitney Newey, Jack Porter, James Powell, Geert Ridder, Paul Rosenbaum, Ed Vytlačil, a co-editor and two anonymous referees, participants at seminars at Berkeley, Brown, BU, Chicago, Georgetown, Harvard/MIT, McGill, Michigan, Princeton, SMU, Texas, Vancouver, 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).

LARGE SAMPLE PROPERTIES OF MATCHING ESTIMATORS
FOR AVERAGE TREATMENT EFFECTS

ALBERTO ABADIE AND GUIDO W. IMBENS

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. The absence of formal results in this area may be partly due to the fact that standard asymptotic expansions do not apply to simple matching estimators, which are highly nonsmooth functionals of the data. In this article, we develop new methods 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 of stochastic order $N^{-1/k}$ where k is the number of continuous matching variables. As a result, matching estimators are generally not $N^{1/2}$ -consistent when more than two continuous covariates are used for matching. 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. Third, we describe a bias-correction that removes the conditional bias asymptotically, making matching estimators $N^{1/2}$ -consistent. Fourth, we provide new a estimator for the variance that does not require consistent nonparametric estimation of unknown functions. We apply these ideas to the study of the effects of the National Supported Work Demonstration (NSW), a labor market program previously analyzed by Lalonde (1986) and others. We also carry out a small simulation study based on the NSW example where 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, root-mean-squared-error and coverage rates. Software for implementing the proposed estimators in STATA and Matlab is available from the authors on the web.

keywords: *matching estimators, average treatment effects, bias correction, unconfoundedness, selection-on-observables, potential outcomes*

Alberto Abadie
John F. Kennedy School of Government
Harvard University
79 John F. Kennedy Street
Cambridge, MA 02138
and NBER
alberto_abadie@harvard.edu
<http://www.ksg.harvard.edu/fs/aabadie/>

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
<http://elsa.berkeley.edu/users/imbens/>

1. INTRODUCTION

Estimation of average treatment effects is an important goal of much evaluation research, both in academic studies (e.g, Ashenfelter, 1978; 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 assumptions that (i) assignment to treatment is unconfounded or exogenous, that is, based on observable pretreatment variables only, and (ii) there is sufficient overlap in the distributions of the pretreatment variables. Methods for estimating average treatment effects in parametric settings under these assumptions have a long history. See for example Cochran and Rubin (1973), Rubin (1977), Barnow, Cain, and Goldberger (1980), Rosenbaum and Rubin (1983), Heckman and Robb (1984), and Rosenbaum (1995). Recently, a number of nonparametric implementations of this idea have been proposed. Hahn (1998) calculates the efficiency bound and proposes an asymptotically efficient estimator based on nonparametric series estimation. Heckman, Ichimura, and Todd (1997, 1998) and Heckman, Ichimura, Smith, and Todd (1998) focus on the average effect on the treated and consider estimators based on local linear kernel regression methods. 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 (1983), leads to an efficient estimator. Ichimura and Linton (2001) consider higher order expansions of such estimators to analyze optimal choices for smoothing parameters.

Empirical researchers, however, often use simple matching procedures to estimate average treatment effects when assignment for treatment is believed to be unconfounded. Much like nearest neighbor estimators, these procedures match each treated unit to a fixed number of untreated units with similar values for the pretreatment variables. The average effect of the treatment is then estimated by averaging within-match differences in the outcome variable between the treated and the untreated units (see, e.g., Rosenbaum, 1989, 1995; Rubin and Thomas, 1992ab; Gu and Rosenbaum, 1993; Rubin, 1973ab; Dehejia and Wahba, 1999; Zhao, 2001; Becker and Ichino, 2002; Frölich, 2000). Matching estimators have great intuitive appeal, and are widely used in practice, as they do not require the researcher to choose smoothing parameters as functions of

the sample size. However, their formal large sample properties have not been established. Part of the reason may be that simple matching estimators are highly non-smooth functionals of the distribution of the data, not amenable to standard asymptotic methods for smooth functionals. In this article, we study the large sample properties of matching estimators of average treatment effects and establish a number of new results.

Our results show that some of the formal large sample properties of simple matching estimators are not very attractive. First, we show that matching estimators include a conditional bias term of stochastic order $N^{-1/k}$, where k is the number of continuous matching variables. As a result, matching estimators are in general not $N^{1/2}$ -consistent when more than two continuous covariates are used for matching. 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 a single continuous covariate is used to match, we show that the efficiency loss can be made arbitrarily close to zero by allowing a sufficiently large number of matches.

Despite these poor formal properties, matching estimators do have some attractive features that may account for their popularity. First of all, matching estimators are extremely easy to implement, as they do not require the choice of smoothing parameters as functions of the sample size. Second, as they do not exploit existence of high order derivatives, matching estimators are consistent under limited smoothness requirements (a situation where consistency of competing estimators has not been established). This is an appealing feature of simple matching estimators because reliance on properties of higher order derivatives has recently come under some criticism (e.g., Horowitz and Mammen, 2002; Angrist and Hahn, 2003; Robins and Ritov 1997).

We also investigate estimators that combine matching with a nonparametric extension of the bias correction proposed in Rubin (1973b) and Quade (1982). We show that a nonparametric implementation of the bias correction removes the conditional bias asymptotically without affecting the variance. This bias-corrected matching estimator combines some of the advantages and disadvantages of both matching and regression estimators. Compared to the simple matching estimator it has the advantage of being $N^{1/2}$ -consistent and asymptotically normal irrespective of the number of covariates. It has the disadvantage of being more difficult to implement than the simple matching estimator, because it involves the choice of smoothing parameters as functions of the sample size. 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 weighting estimators (Hirano, Imbens, and Ridder, 2000), the bias-corrected matching estimator has the advantage of an additional layer of robustness, because matching ensures consistency for any given value of the smoothing parameters, without accurate approximations to either the regression function or the propensity score. Compared to the nonparametric regression estimators the bias-adjusted matching estimator has the disadvantage of not being fully efficient.

In this paper we also propose a consistent estimator for the variances of matching estimators that does not require the choice of data-dependent smoothing parameters.¹ In addition, we show that matching estimators can be interpreted as estimators of the sample average treatment effect conditional on the covariates. We show that this conditional average treatment effect can be estimated more precisely than the population average treatment effect, at the expense of being specific to the covariate and sample design. We also propose an estimator for its variance.

We apply the estimators investigated in this article to the National Supported Work Demonstration data, analyzed originally by Lalonde (1986) and subsequently by Heckman and Hotz (1989), Dehejia and Wahba (1999), Smith and Todd (2001, 2003), Zhao (2002), and Imbens (2003a). For this data set, 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 by linear least squares is much more robust. Moreover, in a small simulation study designed to mimic the data from the NSW application, we find that the linear implementation of the bias-corrected matching estimator performs well compared to both simple matching estimators and to regression estimators, in terms of both 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 an estimator for the large sample variance of matching estimators. In Section 6 we apply the estimators to the NSW data. In Section 7 we carry out a small simulation study to investigate the properties of the various estimators discussed in this article. Section 8 concludes. The appendix contains proofs.

¹To our knowledge, there are no formal results available on the properties of the bootstrap or other resampling methods to estimate the variance of matching estimators of average treatment effects. Because matching estimators are highly non-smooth functionals of the data such properties may be difficult to derive.

2. NOTATION AND BASIC IDEAS

2.1. NOTATION

We are interested in estimating the average effect of a binary treatment on some outcome. For unit i , with $i = 1, \dots, N$, let $Y_i(0)$ and $Y_i(1)$ denote the two potential outcomes given the control treatment and given the active treatment, respectively. The variable W_i , 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 . Following the literature, our main focus is on 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), Heckman and Robb (1984), and Imbens (2003b) for some discussion of these estimands. In addition, we will present some results for the conditional average treatment effect for the sample,

$$\overline{\tau(X)} = \frac{1}{N} \sum_{i=1}^N \mathbb{E}[Y_i(1) - Y_i(0) | X_i],$$

and its counterpart for the subpopulation of treated units,

$$\overline{\tau(X)}^t = \frac{1}{N_1} \sum_{i=1}^N W_i \cdot \mathbb{E}[Y_i(1) - Y_i(0) | X_i, W_i = 1],$$

where $N_1 = \sum_{i=1}^N W_i$ is the number of treated units. As we will show later, the advantage of these last two estimands is that they can generally be estimated more precisely than the population average treatment effects. A disadvantage is that they are specific to the sample.

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

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

ASSUMPTION 2: For almost every $x \in \mathbb{X}$,

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

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

The dimension of X , denoted by k , will be seen to play an important role in the properties of matching estimators. We assume that all covariates have continuous distributions.² The combination of the two conditions in Assumption 2 is referred to as strong ignorability (Rosenbaum and Rubin, 1983). These conditions are strong, and in many cases may not be satisfied. In many studies, however, researchers have found it useful to consider estimators based on these or similar conditions. See, for example, Cochran (1968), Ashenfelter (1978), Barnow, Cain, and Goldberger (1980), Rosenbaum and Rubin (1983), Heckman and Robb (1984), Lalonde (1986), Heckman, Ichimura, and Todd (1997), Card and Sullivan (1988), Angrist (1998), Hahn (1998), Lechner (1998), Dehejia and Wahba (1999), Hotz, Imbens, and Mortimer (1999), Blundell, Costa Dias, Meghir, and Van Reenen (2001), Becker and Ichino (2002), and Firpo (2003). If Assumption 2(i), 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, 2003), or bounds calculations (Manski, 1990, 1995) may be considered. For a general discussion of such issues, see the surveys in Heckman and Robb (1984), Angrist and Krueger (2000), Heckman, Lalonde, and Smith (2000), and Blundell and Costa Dias (2001). The importance of Assumption 2(ii), the restriction on the probability of assignment, has been discussed in Heckman, Ichimura, and Todd (1997) and Dehejia and Wahba (1999). Compactness and convexity of the support of the covariates are convenient regularity conditions.

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

ASSUMPTION 2': For almost every $x \in \mathbb{X}$,

²Discrete covariates with a finite number of support points can be easily dealt with by analyzing estimation of average treatment effects within subsamples defined by their values. The number of such covariates does not affect the asymptotic properties of the estimators. In small samples, however, matches along discrete covariates may not be exact, so discrete covariates may create the same type of biases as continuous covariates.

- (i) W is independent of $Y(0)$ conditional on $X = x$;
- (ii) $\Pr(W = 1|X = x) < 1 - \eta$, for some $\eta > 0$.

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

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

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

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

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

$$\tau^t(x) = \mathbb{E}[Y(1) - Y(0)|W = 1, X = x] = \mathbb{E}[Y|W = 1, X = x] - \mathbb{E}[Y|W = 0, X = x]. \quad (2)$$

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

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

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

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

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

reasons that will become clear later, to estimate average effects on the treated we will require that the size of the control group is at least of the same order of magnitude as the size of the treatment group.

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

In this paper we focus on matching with replacement, allowing each unit to be used as a match more than once. For $x \in \mathbb{X}$, let $\|x\| = (x'x)^{1/2}$ be the standard Euclidean vector norm.³ Let $j_m(i)$ be the index j that solves $W_j = 1 - W_i$ and

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

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 covariate values, among the units with the treatment opposite to that of unit i . In particular, $j_1(i)$, which will be sometimes denoted by $j(i)$, is the nearest match for unit i . For notational simplicity and because we only consider continuous covariates, we ignore the possibility of ties, which happen with probability zero. Let $\mathcal{J}_M(i)$ denote the set of indices for the first M matches for unit i :⁴

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

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:

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

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

In many analyses of matching methods (e.g., Rosenbaum, 1995), matching is carried out without replacement, so that every unit is used as a match at most once, and $K_M(i) \leq 1$. In this article, however, we focus on matching with replacement, allowing each unit to be used as a match more than once. Matching with replacement produces matches of higher quality than

³Alternative norms of the form $\|x\|_V = (x'Vx)^{1/2}$ for some positive definite symmetric matrix V are also covered by the results below, because $\|x\|_V = ((Px)'(Px))^{1/2}$ for P such that $P'P = V$.

⁴For this definition to make sense, we assume that $N_0 \geq M$ and $N_1 \geq M$. We maintain this assumption implicit throughout.

matching without replacement by increasing the set of possible matches.⁵ In addition, matching with replacement has the advantage that it allows us to consider estimators that match all units, treated as well as controls, so that the estimand is identical to the population average treatment effect that is the focus of the Hahn (1998), Robins and Rotnitzky (1995), and Hirano, Imbens and Ridder (2000) studies.

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 for the average treatment effects we consider impute the expected potential outcomes in some way. The first estimator, the simple matching estimator, uses the following estimates for the expected potential outcomes:

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

and

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

leading to the following estimator for the average treatment effect:

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

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

$$\hat{\tau}_M^{sm,t} = \frac{1}{N_1} \sum_{W_i=1} \left(Y_i - \hat{Y}_i(0) \right) = \frac{1}{N_1} \sum_{i=1}^N \left(W_i - (1 - W_i) \cdot \frac{K_M(i)}{M} \right) \cdot Y_i. \quad (4)$$

It is useful to compare 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

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

⁵As we show below, inexact matches generate bias in matching estimators. Therefore, expanding the set of possible matches will tend to produce smaller biases.

and

$$\bar{Y}_i(1) = \begin{cases} \hat{\mu}_1(X_i) & \text{if } W_i = 0, \\ Y_i & \text{if } W_i = 1. \end{cases} \quad (6)$$

The regression imputation estimators of τ and τ^t are

$$\hat{\tau}^{reg} = \frac{1}{N} \sum_{i=1}^N (\bar{Y}_i(1) - \bar{Y}_i(0)) \quad \text{and} \quad \hat{\tau}^{reg,t} = \frac{1}{N_1} \sum_{W_i=1} (Y_i - \bar{Y}_i(0)). \quad (7)$$

In our discussion we classify as regression imputation estimators those for which $\hat{\mu}_w(x)$ is a consistent estimator of $\mu_w(x)$. The estimators proposed by Hahn (1998) and some of those proposed by Heckman, Ichimura, and Todd (1997) and Heckman, Ichimura, Smith, and Todd (1998) fall into this category.

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

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

$$\tilde{Y}_i(0) = \begin{cases} Y_i & \text{if } W_i = 0, \\ \frac{1}{M} \sum_{j \in \mathcal{J}_M(i)} (Y_j + \hat{\mu}_0(X_i) - \hat{\mu}_0(X_j)) & \text{if } W_i = 1, \end{cases} \quad (8)$$

and

$$\tilde{Y}_i(1) = \begin{cases} \frac{1}{M} \sum_{j \in \mathcal{J}_M(i)} (Y_j + \hat{\mu}_1(X_i) - \hat{\mu}_1(X_j)) & \text{if } W_i = 0, \\ Y_i & \text{if } W_i = 1, \end{cases} \quad (9)$$

with corresponding estimators

$$\hat{\tau}_M^{bcm} = \frac{1}{N} \sum_{i=1}^N \left(\tilde{Y}_i(1) - \tilde{Y}_i(0) \right) \quad \text{and} \quad \hat{\tau}_M^{bcm,t} = \frac{1}{N_1} \sum_{W_i=1} \left(Y_i - \tilde{Y}_i(0) \right). \quad (10)$$

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

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

Our motivation for including the bias-corrected matching estimator in this discussion is twofold. First, it will turn out that the simple matching estimator has unattractive bias properties when more than one covariate is used in the matching, making it difficult to construct confidence intervals. Under sufficient smoothness conditions, the bias-corrected matching estimator will allow for confidence intervals based on a normal limiting distribution. Second, the bias-corrected matching estimator provides a link between matching and regression estimators, highlighting advantages and disadvantages of both.

3. SIMPLE MATCHING ESTIMATORS

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

$$\hat{\tau}_M^{sm} - \tau = \left(\overline{\tau(X)} - \tau \right) + E_M^{sm} + B_M^{sm}, \quad (11)$$

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

$$\overline{\tau(X)} = \frac{1}{N} \sum_{i=1}^N (\mu_1(X_i) - \mu_0(X_i)), \quad (12)$$

E_M^{sm} is a weighted average of the residuals:

$$E_M^{sm} = \frac{1}{N} \sum_{i=1}^N E_{M,i}^{sm} = \frac{1}{N} \sum_{i=1}^N (2W_i - 1) \cdot \left(1 + \frac{K_M(i)}{M} \right) \cdot \varepsilon_i, \quad (13)$$

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

$$B_M^{sm} = \frac{1}{N} \sum_{i=1}^N B_{M,i}^{sm} = \frac{1}{N} \sum_{i=1}^N (2W_i - 1) \frac{1}{M} \sum_{m=1}^M \left(\mu_{1-W_i}(X_i) - \mu_{1-W_i}(X_{j_m(i)}) \right). \quad (14)$$

The first two terms on the right hand side of equation (11), $(\overline{\tau(X)} - \tau)$ and E_M^{sm} , have zero mean. They will be shown to be $N^{1/2}$ -consistent and asymptotically normal. The first term depends only on the covariates, and its variance is $V^{\tau(X)}/N$, where $V^{\tau(X)} = \mathbb{E}[(\tau(X) - \tau)^2]$ is the variance of the conditional average treatment effect $\tau(X)$. Conditional on \mathbf{X} and \mathbf{W} , (the matrix and vector with i -th row equal to X_i' and W_i respectively) the variance of $\hat{\tau}_M^{sm}$ is equal to the conditional variance of the second term, E_M^{sm} . We will analyze the variances of these two terms in Section 3.2. We will refer to the third term on the right hand side of of equation (11), B_M^{sm} , as the bias term, or the conditional bias, and to $\text{Bias}_M^{sm} = \mathbb{E}[B_M^{sm}]$ as the (unconditional) bias. If matching is exact, $X_i = X_{j_m(i)}$ for all i , and the bias term is equal to zero. In general it is not and its properties will be analyzed in Section 3.1.

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

$$\hat{\tau}_M^{sm,t} - \tau^t = \left(\overline{\tau(X)}^t - \tau^t \right) + E_M^{sm,t} + B_M^{sm,t}, \quad (15)$$

where

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

is the average conditional treatment effect for the treated,

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

is the contribution of the residuals, and

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

is the bias term.

3.1. BIAS

Here we investigate the stochastic order of the conditional bias (14) and its counterpart for the average treatment effect for the treated. The conditional bias 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 :

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

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

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

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 matching 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 at interior points of the support of X .

LEMMA 1: (MATCHING DISCREPANCY – ASYMPTOTIC PROPERTIES)

Suppose that $f(z) > 0$ and that f is differentiable in a neighborhood of z . Let $V_m = N^{1/k} \cdot U_m$ and f_{V_m} be the density of V_m . Then, as $N \rightarrow \infty$,

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

where $\Gamma(y) = \int_0^\infty e^{-t} t^{y-1} dt$ (for $y > 0$) is Euler's Gamma Function, so that $U_m = O_p(N^{-1/k})$.

Moreover, the first three moments of U_m are:

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

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

and

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

where I_k is the identity matrix of size k .

(All proofs are given in the appendix.)

This 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. The following lemma shows that for all points in the support, including boundary points, the normalized moments of the matching discrepancies, U_m , are bounded.

LEMMA 2: (MATCHING DISCREPANCY – UNIFORMLY BOUNDED MOMENTS)

If Assumption 1 holds, then all the moments of $N^{1/k}\|U_m\|$ are uniformly bounded in N and $z \in \mathbb{X}$.

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

THEOREM 1: (ORDER OF THE CONDITIONAL BIAS FOR THE AVERAGE TREATMENT EFFECT)

Suppose assumptions 1, 2 and 3, hold. If $\mu_0(x)$ and $\mu_1(x)$ are Lipschitz on \mathbb{X} , then $B_M^{sm} = O_p(N^{-1/k})$.

Consider the implications of this theorem for the asymptotic properties of the simple matching estimator. First notice that, under regularity conditions, $\sqrt{N}(\overline{\tau(X)} - \tau) = O_p(1)$ with a normal limiting distribution, by a standard central limit theorem. Also, it will be shown later that, under regularity conditions $\sqrt{N}E_M^{sm} = O_p(1)$, again with a normal limiting distribution. Now, suppose the covariate is scalar ($k = 1$). In that case $B_M^{sm} = O_p(N^{-1})$. Hence the asymptotic properties of the simple matching estimator will be dominated by those of $\overline{\tau(X)} - \tau$ and E_M^{sm} , and $\sqrt{N}(\hat{\tau}_M^{sm} - \tau)$ will be asymptotically normal. Next, consider the case with $k = 2$. In that case $B_M^{sm} = O_p(N^{-1/2})$, and the asymptotic properties will be determined by all three terms. Next, consider the case with $k \geq 3$. Now the order of B_M^{sm} is $O_p(N^{-1/k})$. In this case, the asymptotic distribution is dominated by the bias term and the simple matching estimator is not $N^{1/2}$ -consistent.⁶

A similar result holds for the average treatment effect on the treated.

THEOREM 2: (ORDER OF THE CONDITIONAL BIAS FOR THE AVERAGE TREATMENT EFFECT ON THE TREATED)

Under assumptions 1, 2' and 3',

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

⁶A similar role for the dimension of the covariates arises in nonparametric differencing methods for regression models (e.g., Estes and Honoré, 2001; Yatchew, 1999).

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

$$\begin{aligned} \text{Bias}_M^{sm,t} = \mathbb{E}[B_M^{sm,t}] &= - \left(\frac{1}{M} \sum_{m=1}^M \Gamma \left(\frac{mk+2}{k} \right) \frac{1}{(m-1)!k} \right) \frac{1}{N_1^{2r/k}} \times \\ &\theta^{2/k} \int \left(f_0(x) \frac{\pi^{k/2}}{\Gamma(1+k/2)} \right)^{-2/k} \left\{ \frac{1}{f_0(x)} \frac{\partial f_0}{\partial x'}(x) \frac{\partial \mu_0}{\partial x}(x) + \frac{1}{2} \text{tr} \left(\frac{\partial^2 \mu_0}{\partial x' \partial x}(x) \right) \right\} f_1(x) dx \\ &+ o \left(\frac{1}{N_1^{2r/k}} \right). \end{aligned}$$

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

In part (ii) of Theorem 2, we show that a general expression of the bias $\text{Bias}_M^{sm,t}$ can be calculated if \mathbb{X}_1 is compact and $\mathbb{X}_1 \subset \text{int } \mathbb{X}_0$ (so that the bias is not affected by the geometric characteristics of the boundary of \mathbb{X}_0). Under these conditions, the bias of the matching estimator is at most of order $N_1^{-2/k}$. This bias is further reduced when μ_0 is constant or when μ_0 is linear and f_0 is constant, among other cases. Note, however, that randomizing the treatment, $f_0 = f_1$ does not reduce the order of $\text{Bias}_M^{sm,t}$.

3.2. VARIANCE

In this section we investigate the variance of the simple matching estimator, $\hat{\tau}_M^{sm}$. We focus on the first two terms of the simple matching estimator, (12) and (13), ignoring for the moment the bias term (14). Conditional on \mathbf{X} and \mathbf{W} , the matrix and vector with i -th row equal to X_i' and W_i respectively, the variance of $\hat{\tau}_M^{sm}$ is

$$\mathbb{V}(\hat{\tau}_M^{sm} | \mathbf{X}, \mathbf{W}) = \frac{1}{N^2} \sum_{i=1}^N \left(1 + \frac{K_M(i)}{M} \right)^2 \sigma^2(X_i, W_i). \quad (16)$$

For $\hat{\tau}_M^{sm,t}$ we obtain:

$$\mathbb{V}(\hat{\tau}_M^{sm,t}|\mathbf{X}, \mathbf{W}) = \frac{1}{N_1^2} \sum_{i=1}^N (W_i - (1 - W_i) \cdot K_M(i)/M)^2 \cdot \sigma^2(X_i, W_i). \quad (17)$$

Let $V^E = N \cdot \mathbb{V}(\hat{\tau}_M^{sm}|\mathbf{X}, \mathbf{W})$ and $V^{E,t} = N_1 \cdot \mathbb{V}(\hat{\tau}_M^{sm,t}|\mathbf{X}, \mathbf{W})$ be the corresponding normalized variances. Ignoring the bias term, B_M^{sm} , the conditional expectation of $\hat{\tau}_M^{sm}$ is $\overline{\tau(X)}$. The variance of this conditional mean is therefore $V^{\tau(X)}/N$, where $V^{\tau(X)} = \mathbb{E}[(\tau(X) - \tau)^2]$. Hence the marginal variance of $\hat{\tau}_M^{sm}$, ignoring the bias term, is

$$\mathbb{V}(\hat{\tau}_M^{sm}) = (\mathbb{E}[V^E] + V^{\tau(X)})/N.$$

For the estimator for the average effect on the treated the marginal variance is, again ignoring the bias term,

$$\mathbb{V}(\hat{\tau}_M^{sm,t}) = (\mathbb{E}[V^{E,t}] + V^{\tau(X),t})/N_1,$$

where $V^{\tau(X),t} = \mathbb{E}[(\tau^t(X) - \tau^t)^2|W = 1]$.

The following lemma shows that the expectation of the normalized variance is finite. The key is that $K_M(i)$, the number of times that unit i is used as a match, is $O_p(1)$ with finite moments.⁷

LEMMA 3: (i) Suppose assumptions 1-3 hold, then $K_M(i) = O_p(1)$ and its moments are bounded uniformly in N . (ii) If, in addition, $\sigma^2(x, w)$ are Lipschitz in \mathbb{X} for $w = 0, 1$, then $\mathbb{E}[V^E + V^{\tau(X)}] = O(1)$. (iii) Suppose Assumptions 1, 2' and 3', then $(N_0/N_1)\mathbb{E}[K_M(i)^q|W_i = 0]$ is uniformly bounded in N for all $q \geq 1$. (iv) If also $\sigma^2(x, w)$ are Lipschitz in \mathbb{X} for $w = 0, 1$, then $\mathbb{E}[V^{E,t} + V^{\tau(X),t}] = O(1)$.

3.3. CONSISTENCY AND ASYMPTOTIC NORMALITY

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

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

⁷Notice that, for $1 \leq i \leq N$, $K_M(i)$ are exchangeable random variables, and therefore have identical marginal distributions.

THEOREM 3: (CONSISTENCY OF THE SIMPLE MATCHING ESTIMATOR)

(i) Suppose assumptions 1-3 and 4(i) hold. Then

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

(ii) Suppose assumptions 1, 2', 3', and 4(i) hold. Then

$$\hat{\tau}_M^{sm,t} - \tau^t \xrightarrow{P} 0.$$

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

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

THEOREM 4: (ASYMPTOTIC NORMALITY FOR THE SIMPLE MATCHING ESTIMATOR)

(i) Suppose assumptions 1-3 and 4 hold. Then

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

(ii) Suppose assumptions 1, 2', 3', and 4 hold. Then

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

Although one generally does not know the bias term, this result is useful for two reasons. First, in some cases the bias term can be ignored because it is of sufficiently low order. Second, as we shall show below, under some conditions, an estimate of the bias term can be used in the statement of Theorem 4 without changing the result.

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

COROLLARY 1: (ASYMPTOTIC NORMALITY FOR SIMPLE MATCHING ESTIMATOR – VANISHING BIAS)

(i) Suppose assumptions 1-3 and 4 hold, and $k = 1$. Then

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

(ii) Suppose assumptions 1, 2', 3', and 4 hold, and $r > k/2$. Then

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

Similar results hold for simple matching estimators viewed as estimators of $\overline{\tau(X)}$ and $\overline{\tau(X)}^t$.

COROLLARY 2: (ASYMPTOTIC NORMALITY FOR THE SIMPLE MATCHING ESTIMATOR AS AN ESTIMATOR OF THE CONDITIONAL AVERAGE TREATMENT EFFECT)

(i) Suppose assumptions 1-3 and 4 hold. Then

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

(ii) Suppose assumptions 1, 2', 3', and 4 hold. Then

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

Note that we can estimate $\overline{\tau(X)}$ and $\overline{\tau(X)^t}$ more precisely than τ and τ^t , because for the former the sampling variation in the covariates does not add to the variance.

3.4. EFFICIENCY

The asymptotic efficiency of the estimators considered here depends on the limit of $\mathbb{E}[V^E]$, which in turn depends on the limiting distribution of $K_M(i)$. It is difficult to work out the limiting distribution of this variable for the general case.⁸ Here we investigate the form of the variance for the special case with a scalar covariate ($k = 1$) and a general M .

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

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

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

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

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

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

⁸The 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 tessellations (Okabe, Boots, Sugihara and Nok Chiu, 2000). The variance of the volume of such objects under uniform $f_0(x)$ and $f_1(x)$, normalized by the mean, has been worked out numerically for the one, two, and three dimensional cases.

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

4. BIAS CORRECTED MATCHING

In this section we analyze the properties of the bias corrected matching estimators, defined in equation (10). In order to establish the asymptotic behavior of the bias-corrected estimator, we consider a nonparametric series estimator for the two regression functions, $\mu_0(x)$ and $\mu_1(x)$, with $K(N)$ terms in the series, where $K(N)$ increases with N . An important disadvantage of this estimator is that it will rely on selecting smoothing parameters as functions of the sample size, something that the simple matching estimator allows one to avoid. An advantage of the bias-corrected matching estimator is that it is root- N consistent for any dimension of the covariates, k . In both these properties the bias-corrected matching estimator is similar to the regression imputation estimator. However, it has the same large sample variance as the simple matching estimator and, therefore, it is in general not as efficient as the regression imputation estimator in large samples. Compared to the regression imputation estimator the bias-corrected matching estimator is more robust in the sense that it is consistent for any fixed value of the smoothing parameter. However, because choosing smoothing parameters as functions of the sample size is precisely what matching estimators allow one to avoid, in the empirical analysis and simulations of sections 6 and 7 we will investigate the performance of a simple implementation of the bias correction by linear least squares.

Let $\lambda = (\lambda_1, \dots, \lambda_k)$ 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 $x^\lambda = x_1^{\lambda_1} \dots x_k^{\lambda_k}$. The λ -th partial derivative of a function $g(x)$ is given by $\partial^{|\lambda|} g(x) / \partial x_1^{\lambda_1} \dots \partial x_k^{\lambda_k}$. Consider a series $\{\lambda(r)\}_{r=1}^\infty$ containing all distinct such vectors such that $|\lambda(r)|$ is nondecreasing. Let $p_r(x) = x^{\lambda(r)}$, where $p^K(x) = (p_1(x), \dots, p_K(x))'$. 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}_M^{sm} be the

estimated bias term:

$$\hat{B}_M^{sm} = \frac{1}{N} \sum_{i=1}^N \left\{ W_i \left(\frac{1}{M} \sum_{j \in \mathcal{J}_M(i)} (\hat{\mu}_0(X_i) - \hat{\mu}_0(X_j)) \right) - (1 - W_i) \left(\frac{1}{M} \sum_{j \in \mathcal{J}_M(i)} (\hat{\mu}_1(X_i) - \hat{\mu}_1(X_j)) \right) \right\},$$

and

$$\hat{B}_M^{sm,t} = \frac{1}{N_1} \sum_{i=1}^N W_i \left(\frac{1}{M} \sum_{j \in \mathcal{J}_M(i)} (\hat{\mu}_0(X_i) - \hat{\mu}_0(X_j)) \right),$$

for the average effect for the treated, so that $\hat{\tau}^{bcm} = \hat{\tau}_M^{sm} - \hat{B}_M^{sm}$, and $\hat{\tau}^{bcm,t} = \hat{\tau}_M^{sm,t} - \hat{B}_M^{sm,t}$. The following theorem shows that the bias correction removes the conditional bias without affecting the asymptotic variance.

THEOREM 6: (BIAS CORRECTED MATCHING ESTIMATOR)

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

$$\sqrt{N} \left(B_M^{sm} - \hat{B}_M^{sm} \right) \xrightarrow{p} 0 \quad \text{and} \quad \left(V^E + V^{\tau(X)} \right)^{1/2} \sqrt{N} (\hat{\tau}_M^{bcm} - \tau) \xrightarrow{d} \mathcal{N}(0, 1).$$

If assumptions 1, 2', 3' and 4 hold in addition to (i)-(iii), we obtain:

$$\sqrt{N_1} \left(B_M^{sm,t} - \hat{B}_M^{sm,t} \right) \xrightarrow{p} 0 \quad \text{and} \quad \left(V^{E,t} + V^{\tau(X),t} \right)^{1/2} \sqrt{N_1} (\hat{\tau}_M^{bcm,t} - \tau^t) \xrightarrow{d} \mathcal{N}(0, 1).$$

Thus, the bias corrected matching estimator has the same normalized variance as the simple matching estimator.⁹

5. ESTIMATING THE VARIANCE

Theorem 4 and Corollary 2 use the square-roots of $V^E + V^{\tau(X)}$ and V^E , respectively, as normalizing factors to attain a limiting normal distribution for matching estimators. In this section, we show how to estimate these two terms separately. This will allow researchers to conduct inference for the conditional or unconditional average treatment effect, as well as to assess the variation in the conditional treatment effect $\tau(X)$ by providing an estimator for $V^{\tau(X)}$.

⁹It is easy to check that the same result holds if $\mu_w(x)$ has a finite series representation which is estimated by least squares. This feature will be used later to construct standard errors for the bias corrected estimator when the bias correction is implemented using a simple linear regression.

5.1. ESTIMATING THE CONDITIONAL VARIANCE

Estimating the conditional variance, $V^E = \sum_{i=1}^N (1 + K_M(i)/M)^2 \sigma_{W_i}^2(X_i)/N$, is complicated by the fact that it involves the conditional outcome variances, $\sigma_w^2(x)$. Estimating $\sigma_w^2(x)$ consistently would require exactly the type of nonparametric regression that the simple matching estimator allows one to avoid. For this reason, we propose a different estimator of the conditional variance of the simple matching estimator which does not require consistent nonparametric estimation of $\sigma_w^2(x)$. Our method uses a matching estimator for $\sigma_w^2(x)$ where instead of the original matching of treated to control units, we now match treated units to treated units and control units to control units. This leads to an approximately unbiased estimator of $\sigma_w^2(x)$, although not a consistent one. However, the average of these variance estimators is consistent for the average of the variances.

To fix ideas, suppose we have two units i and j with the same covariate values, $X_i = X_j = x$, and the same treatment values, $W_i = W_j = w$. Notice that

$$E \left[\left(Y_i - Y_j \right)^2 \middle| X_i = X_j = x, W_i = W_j = w \right] = 2\sigma_w^2(x).$$

In this case we can estimate the variance, $\sigma_{W_i}^2(X_i)$, as $\hat{\sigma}_{W_i}^2(X_i) = (Y_i - Y_j)^2/2$. This estimator is unbiased, but is not consistent because its variance does not go to zero with the sample size. However, consistent estimation of $\sigma_{W_i}^2(X_i)$ is not required for the estimator of V^E to be consistent. In practice, it may not be possible to find different units with the same value of the covariates. Hence let us consider the nearest match to unit i within the same treatment group:

$$l(i) = \operatorname{argmin}_{l:l \neq i, W_l = W_i} \|X_i - X_l\|,$$

and let

$$\hat{\sigma}_{W_i}^2(X_i) = \frac{1}{2} \left(Y_i - Y_{l(i)} \right)^2, \tag{18}$$

be an estimator for the conditional variance $\sigma_{W_i}^2(X_i)$. More generally, one can use J nearest neighbors to estimate the local variances with the same result. Let $l_j(i)$ be the j -th closest unit to unit i among the units with the same value for the treatment. Then, we estimate the conditional variance as

$$\hat{\sigma}_{W_i}^2(X_i) = \frac{J}{J+1} \left(Y_i - \frac{1}{J} \sum_{j=1}^J Y_{l_j(i)} \right)^2. \tag{19}$$

The next theorem establishes consistency of an estimator of the conditional variance based on the estimators of $\sigma_{W_i}^2(X_i)$ defined in equation (19).

THEOREM 7: Let $\widehat{\sigma}_{W_i}^2(X_i)$ be as in equation (19). Let

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

If Assumptions 1 to 4 hold, then $|\widehat{V}^E - V^E| = o_p(1)$. If Assumptions 1 2', 3' and 4 hold, then $|\widehat{V}^{E,t} - V^{E,t}| = o_p(1)$.

5.2. ESTIMATING THE MARGINAL VARIANCE

Here we develop consistent estimators for $V = V^E + V^{\tau(X)}$ and $V^t = V^{E,t} + V^{\tau(X),t}$. These estimators are based on the same matching approach to estimating the conditional error variance $\sigma_w^2(x)$ as in the previous subsection. In addition, these estimator exploit the fact that,

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

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

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

which can then be combined to estimate $V^{\tau(X)}$. This in turn can be combined with the previously defined estimator for V^E to obtain an estimator of V .

THEOREM 8: Let $\widehat{\sigma}^2(X_i, W_i)$ be as in equation (19). Let

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

and

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

If Assumptions 1 to 4 hold, then $|\widehat{V} - V| = o_p(1)$. If Assumptions 1 2', 3' and 4 hold, then $|\widehat{V}^t - V^t| = o_p(1)$.

6. AN APPLICATION TO THE EVALUATION OF A LABOR MARKET PROGRAM

In this section we apply the estimators studied in this article to data from the National Supported Work Demonstration (NSW), an evaluation of a subsidized work program first analyzed

by Lalonde (1986) and subsequently by Heckman and Hotz (1989), Dehejia and Wahba (1999) and Smith and Todd (2001). The specific sample we use here is the one employed by Dehejia and Wahba (1999).¹⁰ The data set contains an experimental sample from a randomized evaluation of the NSW 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 program. We then compute non-experimental matching estimators using the experimental participants and the nonexperimental comparison group from the PSID, and compare them to the experimental estimate.¹¹ Given the relative sizes 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 1 presents summary statistics for the three groups used in our analysis. The first two columns present the summary statistics for the experimental treatment group. The second pair of columns presents summary statistics for the experimental controls. The third pair of columns presents summary statistics for the non-experimental comparison group constructed from the PSID. The last two columns present t-statistics for the hypotheses that the differences in population averages for the experimental treated and controls, and for the experimental treated and the PSID comparison group, respectively, are zero. Panel A contains the results for pretreatment variables and Panel B for outcomes. Notice the large differences in background characteristics between the program participants and the PSID sample. This is what makes drawing causal inferences from comparisons between the PSID sample and the treatment group a tenuous task. From Panel B, we can obtain an unbiased estimate of the effect of the NSW program on earnings in 1978 by comparing the averages for the experimental treated and controls, $6.35 - 4.55 = 1.80$, with a standard error of 0.67 (earnings are measured in thousands of dollars). Using a normal approximation to the limiting distribution of the experimental estimator, we obtain a 95% confidence interval, which is [0.49, 3.10].

Table 2 presents estimates of the causal effect of the NSW program on earnings using various matching and regression adjustment estimators. Panel A reports estimates for the experimental data (treated and controls). Panel B reports estimates based on the experimental treated and the PSID comparison group. The first set of rows in each case reports matching estimates for M equal to 1, 4, 16, 64 and 2490.¹²

¹⁰This data set is available on the website of Dehejia, at <http://www.columbia.edu/~rd247>.

¹¹Software for implementing matching estimators in Matlab and STATA is available from the authors on the web. Becker and Ichino (2001) have implemented a number of alternative estimators for average treatment effects under unconfoundedness.

¹²The matching estimates include simple matching with no bias adjustment, and bias-adjusted matching. All

The experimental estimates range from 1.16 (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, 2.07 and 2.42 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 units in the comparison group 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 3 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 comparison group and the experimental treated. 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. 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 treated and comparison units 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 very small until we get to 16 or more matches. As expected, match quality deteriorates when the number of matches increases. This explains why, as shown in Table 2, the bias-correction matters more for larger M . The last row reports matching differences for logistic estimates of the

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. The bias adjustment uses linear regression on the nine pretreatment covariates in Table 1, panel A, but not higher order terms or interactions. The bias correction is estimated using only the matched units in the comparison group. The confidence intervals are based on variance estimated using the matching method described in Section 5 with $J = 4$. The last three rows of each panel report estimates based on differences in means, linear regression including terms for all covariates, and linear regression including also quadratic terms and a full set of interactions, respectively.

propensity score. Although the matching is not directly on the propensity score, with a single match the average difference in the propensity score is only 0.21, whereas without matching the difference between treated and comparison units is 8.16, almost 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 aim to resemble the NSW data set analyzed in the previous section.

In the simulation we have nine regressors, designed to match the following variables in the NSW data set: age, education, black, hispanic, married, earnings1324, unemployed1324, earnings1975, unemployed1975. For each simulated data set we sampled with replacement 185 observations from the empirical covariate distribution of the experimental treated, and 2490 observations from the empirical covariate distribution of the PSID comparison group. 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 comparison group, where the probability of zero earnings is a logistic function of the covariates with a full set of quadratic terms and interactions.¹³ Conditional on log earnings being positive, the log of earnings is modeled as a linear function of the covariates with again a full set of quadratic terms and interactions. We fix the treatment effect at 2.0 for all units.

We perform 10,000 replications. 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 based on the matching estimator for the variance. As before, we implement an extremely simple version of the bias adjustment, using only linear terms in the covariates. The results are reported in Table 4.

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, in terms of bias or rmse. The pure regression adjustment estimators perform poorly. They have high rmse and substantial bias. The bias-corrected estimator also performs well in terms of coverage rates.

¹³With the nine variables, this gives 54 covariates in the set with all linear terms, quadratic terms and interactions, not including the intercept. Out of these 54 we drop 8 because they are perfectly collinear by definition (e.g., the interaction of earnings 1975 and unemployed 1975).

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 that simple matching estimators may include a conditional bias term which does not disappear in large samples, under the standard $N^{1/2}$ normalization. Therefore, standard matching estimators are not $N^{1/2}$ -consistent in general. 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 a realistic setting 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 in terms of bias and mean-squared error. It also has good coverage rates for 90 and 95% confidence intervals, suggesting it may be a useful estimator in practice.

There are several open avenues for future work in this area. One possible extension concerns the optimal number of matches for matching estimators, when M is chosen as a function of the sample size. Another possible extension is related to missing data problems. The methods developed in this article for matching estimators can be applied to analyze hot-deck imputation procedures for missing data. Hot-deck procedures impute missing data for an individual or family using the value of the same variable for another individual or family with similar observed characteristics (see, e.g., Little and Rubin, 2002). Matching and hot-deck imputation are formally equivalent. This implies that hot-deck imputation methods suffer from the same conditional bias issues as matching estimators, leading to potentially invalid confidence intervals.

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,

$$\int_{S_k} \omega \lambda_{S_k}(d\omega) = 0, \quad \text{and} \quad \int_{S_k} \omega \omega' \lambda_{S_k}(d\omega) = \frac{\int_{S_k} \lambda_{S_k}(d\omega)}{k} I_k = \frac{\pi^{k/2}}{\Gamma(1+k/2)} I_k,$$

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*

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

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

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

PROOF: It follows from Theorem 7.1 in Olver (1997), page 81.

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

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

Because 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:

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

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

$$f_{U_m}(u) = N \binom{N-1}{m-1} f(z+u) (1 - \Pr(\|X - z\| \leq \|u\|))^{N-m} (\Pr(\|X - z\| \leq \|u\|))^{m-1}. \quad (\text{A.1})$$

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

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

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,

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

In addition, it is easy to check that

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

Therefore,

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

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:

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

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),

$$EU_m = N \binom{N-1}{m-1} A_m,$$

where

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

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

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

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

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

and substituting this into the expression for A_m gives:

$$\begin{aligned} A_m &= \int_0^\infty r^{k-1} \left(\int_{S_k} r\omega f(z+r\omega) \lambda_{S_k}(d\omega) \right) \left(1 - \int_0^r s^{k-1} \left(\int_{S_k} f(z+s\omega) \lambda_{S_k}(d\omega) \right) ds \right)^{N-m} \\ &\quad \times \left(\int_0^r s^{k-1} \left(\int_{S_k} f(z+s\omega) \lambda_{S_k}(d\omega) \right) ds \right)^{m-1} dr = \int_0^\infty e^{-Nb(r)} a(r) dr, \end{aligned}$$

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

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

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

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

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

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

Therefore,

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

Similar calculations yield

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

Therefore

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

Now, it is clear that

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

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

$$\begin{aligned} A_m &= \Gamma\left(\frac{mk+2}{k}\right) \frac{a_0}{kb_0^{(mk+2)/k}} \frac{1}{N^{(mk+2)/k}} + o\left(\frac{1}{N^{(mk+2)/k}}\right) \\ &= \Gamma\left(\frac{mk+2}{k}\right) \frac{1}{k} \left(\frac{f(z)}{k} \int_{S_k} \lambda_{S_k}(d\omega) \right)^{-2/k} \frac{1}{f(z)} \frac{df}{dx}(z) \frac{1}{N^{(mk+2)/k}} + o\left(\frac{1}{N^{(mk+2)/k}}\right) \\ &= \Gamma\left(\frac{mk+2}{k}\right) \frac{1}{k} \left(f(z) \frac{\pi^{k/2}}{\Gamma\left(1 + \frac{k}{2}\right)} \right)^{-2/k} \frac{1}{f(z)} \frac{df}{dx}(z) \frac{1}{N^{(mk+2)/k}} + o\left(\frac{1}{N^{(mk+2)/k}}\right). \end{aligned}$$

Now, because

$$\lim_{N \rightarrow \infty} \frac{N^m / (m-1)!}{N \binom{N-1}{m-1}} = 1,$$

we have that

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

which finishes the proof for the second result of the lemma.

To get the result for $E[U_m U'_m]$, notice that

$$\mathbb{E}[U_m U'_m] = N \binom{N-1}{m-1} B_m,$$

where

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

Applying the same techniques as for A_m , we obtain

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

Hence,

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

It follows from a similar argument that

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

Therefore

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

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 $\bar{f}_{V_m}(v)$ which does not depend on N or z , followed by a proof that $\int \|v\|^L \bar{f}_{V_m}(v) dv < \infty$ for any $L > 0$. It is enough to show the result for $N > m$ (the bounded support condition implies uniformly bounded moments of V_m over $z \in \mathbb{X}$ for any given N , and in particular for $N = m$.) Recall from the proof of Lemma 1 that

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

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} ($\bar{u} = \sup_{x, y \in \mathbb{X}} \|x - y\|$) which is finite because \mathbb{X} is bounded by assumption. 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. Notice that, because \mathbb{X} has dimension k , then $0 < c < 1$; and that, because \mathbb{X} is convex, this proportion can only increase for a smaller radius. Let $z \in \mathbb{X}$ and $\|v\| \leq N^{1/k} \bar{u}$.

$$\begin{aligned}
\Pr\left(\|X - z\| \leq \frac{\|v\|}{N^{1/k}}\right) &= \int_0^{\|v\|N^{-1/k}} r^{k-1} \int_{S_k} f(z + r\omega) \lambda_{S_k}(d\omega) dr \\
&= \int_0^{\|v\|N^{-1/k}} r^{k-1} \int_{S_k} f(z + r\omega) 1\{f(z + r\omega) > 0\} \lambda_{S_k}(d\omega) dr \\
&\geq \underline{f} \int_0^{\|v\|N^{-1/k}} r^{k-1} \int_{S_k} 1\{f(z + r\omega) > 0\} \lambda_{S_k}(d\omega) dr \\
&\geq c \underline{f} \int_0^{\|v\|N^{-1/k}} r^{k-1} \int_{S_k} \lambda_{S_k}(d\omega) dr \\
&= c \frac{\|v\|^k}{N} \underline{f} \frac{\pi^{k/2}}{\Gamma(1 + k/2)}.
\end{aligned}$$

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 $0 < b < N$ and $N > m$ we have $(1 - b/N)^{(N-m)} \leq \exp(-b(N-m)/N) \leq \exp(-b/(m+1))$. In addition,

$$N^{1-m} \binom{N-1}{m-1} \leq \frac{1}{(m-1)!}.$$

It follows that

$$\begin{aligned}
f_{V_m}(v) &\leq \bar{f}_{V_m}(v) = \frac{1}{(m-1)!} \bar{f} \exp\left(-c \frac{\|v\|^k}{(m+1)} \underline{f} \frac{2\pi^{k/2}}{\Gamma(k/2)}\right) \left(\|v\|^k \bar{f} \frac{2\pi^{k/2}}{\Gamma(k/2)}\right)^{m-1} \\
&= C_1 \cdot \|v\|^{k(m-1)} \cdot \exp(-C_2 \cdot \|v\|^k),
\end{aligned}$$

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 , $\int \|v\|^L \bar{f}_{V_m}(v) dv$ is finite and thus all moments of $N^{1/k} \cdot U_m$ are uniformly bounded in N and z . \square

PROOF OF THEOREM 1:

For part (i) of the theorem, let the unit-level matching discrepancy $U_{m,i} = X_i - X_{j_m(i)}$. Define the unit-level conditional bias from the m -th match as

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

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

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

Notice that

$$\begin{aligned}
\mathbb{E}[N^{2/k}(B_M^{sm})^2] &= N^{2/k} \cdot \mathbb{E} \left[\frac{1}{N^2 \cdot M^2} \sum_{i,j} \sum_{l,m} B_{m,i} \cdot B_{l,j} \right] \\
&\leq N^{2/k} \cdot \max_{m,i} \mathbb{E} [(B_{m,i})^2] \leq C^2 \cdot \max_{m,i} \mathbb{E} [(N^{1/k} \|U_{m,i}\|)^2] \\
&\leq C^2 \cdot \max_{m,i} \mathbb{E} \left[\left(\frac{N}{N_{1-W_i}} \right)^{2/k} \left(N_{1-W_i}^{1/k} \|U_{m,i}\| \right)^2 \right].
\end{aligned}$$

Assumption 2(ii) implies that there are versions of f_0 and f_1 that are bounded and bounded away from zero on \mathbb{X} . Then, by Lemma 2, for any given m , any moment of $N_{1-W_i}^{1/k} U_{m,i}$ is uniformly bounded over N_{1-W_i} and i . Using Chernoff's Inequality, it can be easily seen that any moment of N/N_{1-W_i} is uniformly bounded in N (with $N_w \geq M$ for $w = 0, 1$). Applying Hölder's Inequality, and because m only takes on M values, we obtain that the second moment of $N^{1/k} B_M^{sm}$ is uniformly bounded as a function of N . The result of the theorem follows now from Markov's Inequality. \square

PROOF OF THEOREM 2:

The proof of first part of theorem 2 is very similar to the proof of Theorem 1, and therefore is omitted. To prove the second part of Theorem 2, the following auxiliary lemma will be useful.

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

PROOF OF LEMMA A.2: Fix $z \in \mathbb{Z}$. From the proof of Lemma 1, we know that:

$$\begin{aligned}
E[U_m] &= N \binom{N-1}{m-1} \int u f(z+u) \left(1 - \Pr(\|X-z\| \leq \|u\|)\right)^{N-m} \\
&\quad \times \left(\Pr(\|X-z\| \leq \|u\|)\right)^{m-1} du \\
&= N \binom{N-1}{m-1} \int_0^\infty r^{k-1} \left(\int_{S_k} r \omega f(z+r\omega) \lambda_{S_k}(d\omega)\right) \left(1 - \Pr(\|X-z\| \leq r)\right)^{N-m} \\
&\quad \times \left(\Pr(\|X-z\| \leq r)\right)^{m-1} dr.
\end{aligned}$$

Let $\bar{r}(z) = \sup\{r > 0 \mid z + r\omega \in \mathbb{X}, \text{ for all } \omega \in S_k\}$. Given the conditions of the lemma, there exists \underline{r} such that $\bar{r}(z) \geq \underline{r} > 0$ for all $z \in \mathbb{Z}$. Let

$$A_m^1 = \int_0^{\bar{r}(z)} \varphi(r) dr, \quad \text{and} \quad A_m^2 = \int_{\bar{r}(z)}^\infty \varphi(r) dr,$$

where

$$\begin{aligned}
\varphi(r) &= N \binom{N-1}{m-1} r^{k-1} \left(\int_{S_k} r \omega f(z+r\omega) \lambda_{S_k}(d\omega)\right) \left(1 - \Pr(\|X-z\| \leq r)\right)^{N-m} \\
&\quad \times \left(\Pr(\|X-z\| \leq r)\right)^{m-1}.
\end{aligned}$$

Then,

$$E[U_m] = A_m^1 + A_m^2.$$

Consider the change of variable $t = rN^{1/k}$, then:

$$\begin{aligned} A_m^1 &= \frac{1}{N^{1/k}} \int_0^{\bar{r}(z)N^{1/k}} \binom{N-1}{m-1} t^k \left(\int_{S_k} \omega f \left(z + \frac{t}{N^{1/k}} \omega \right) \lambda_{S_k}(d\omega) \right) \\ &\quad \times \left(1 - \Pr \left(\|X - z\| \leq \frac{t}{N^{1/k}} \right) \right)^{N-m} \left(\Pr \left(\|X - z\| \leq \frac{t}{N^{1/k}} \right) \right)^{m-1} dt. \end{aligned}$$

Let $\overline{\|\partial f/\partial X\|} = \sup_{x \in \text{int } \mathbb{X}} \|(\partial f/\partial X)(x)\|$. By the Mean Value Theorem, for $\tilde{t} \leq t \leq \bar{r}(z)N^{1/k}$

$$\left| f \left(z + \frac{t}{N^{1/k}} \omega \right) - f(z) \right| = \frac{t}{N^{1/k}} \left\| \omega' \frac{\partial f}{\partial X} (z + \tilde{t}N^{-1/k}\omega) \right\| \leq \frac{t}{N^{1/k}} \overline{\|\frac{\partial f}{\partial X}\|}.$$

As a result, for $t \leq \bar{r}(z)N^{1/k}$, we obtain:

$$\begin{aligned} \left\| \int_{S_k} \omega f \left(z + \frac{t}{N^{1/k}} \omega \right) \lambda_{S_k}(d\omega) \right\| &= \left\| \int_{S_k} \omega \left(f \left(z + \frac{t}{N^{1/k}} \omega \right) - f(z) \right) \lambda_{S_k}(d\omega) \right\| \\ &\leq \frac{t}{N^{1/k}} \overline{\|\frac{\partial f}{\partial X}\|} \int_{S_k} \lambda_{S_k}(d\omega). \end{aligned}$$

Note also that for $t \leq \bar{r}(z)N^{1/k}$ we have:

$$\begin{aligned} \Pr \left(\|X - z\| \leq \frac{t}{N^{1/k}} \right) &= \int_0^{tN^{-1/k}} s^{k-1} \left(\int_{S_k} f(z + s\omega) \lambda_{S_k}(d\omega) \right) ds \\ &\geq \underline{f} \int_0^{tN^{-1/k}} s^{k-1} ds \int_{S_k} \lambda_{S_k}(d\omega) = \frac{1}{N} \frac{t^k}{k} \underline{f} \int_{S_k} \lambda_{S_k}(d\omega). \end{aligned}$$

Similarly:

$$\Pr \left(\|X - z\| \leq \frac{t}{N^{1/k}} \right) \leq \frac{1}{N} \frac{t^k}{k} \bar{f} \int_{S_k} \lambda_{S_k}(d\omega).$$

Therefore

$$\begin{aligned} \|A_m^1\| &\leq \frac{1}{N^{2/k}} \int_0^{\bar{r}(z)N^{1/k}} N^{1-m} \binom{N-1}{m-1} t^{k+1} \overline{\|\frac{\partial f}{\partial X}\|} \int_{S_k} \lambda_{S_k}(d\omega) \\ &\quad \times \left(1 - \frac{1}{N} \frac{t^k}{k} \underline{f} \int_{S_k} \lambda_{S_k}(d\omega) \right)^{N-m} \left(\frac{t^k}{k} \bar{f} \int_{S_k} \lambda_{S_k}(d\omega) \right)^{m-1} dt. \end{aligned}$$

It is easy to see that:

$$N^{1-m} \binom{N-1}{m-1} \leq \frac{1}{(m-1)!}.$$

It is also easily seen that for $0 < b < N$ and $N > m$, $(1 - b/N)^{N-m} \leq \exp(-b/(m+1))$. Therefore, for $z \in \mathbb{Z}$ and $N > m$, we obtain:

$$\begin{aligned} \|A_m^1\| &< \frac{1}{N^{2/k}} \int_0^\infty \frac{1}{(m-1)!} t^{k+1} \overline{\|\frac{\partial f}{\partial X}\|} \int_{S_k} \lambda_{S_k}(d\omega) \\ &\quad \times \exp \left(-\frac{t^k}{(m+1)k} \underline{f} \int_{S_k} \lambda_{S_k}(d\omega) \right) \left(\frac{t^k}{k} \bar{f} \int_{S_k} \lambda_{S_k}(d\omega) \right)^{m-1} dt \\ &\leq C_1 \frac{1}{N^{2/k}}, \end{aligned}$$

for some positive constant, C_1 .

For A_m^2 , notice that, for $N \geq m$:

$$\begin{aligned}
\|A_m^2\| &\leq \left(1 - \Pr(\|X - z\| \leq \bar{r}(z))\right)^{N-m} N \binom{N-1}{m-1} \int_{\bar{r}(z)}^{\infty} r^{k-1} \left\| \int_{S_k} r\omega f(z+r\omega) \lambda_{S_k}(d\omega) \right\| \\
&\quad \times \Pr(\|X - z\| \leq r)^{m-1} dr \\
&< \left(1 - \Pr(\|X - z\| \leq \bar{r}(z))\right)^{N-m} \frac{N^m}{m!} \\
&\quad \times \left(m \int_0^{\infty} r^{k-1} \left\| \int_{S_k} r\omega f(z+r\omega) \lambda_{S_k}(d\omega) \right\| \Pr(\|X - z\| \leq r)^{m-1} dr\right) \\
&\leq N^m \left(1 - \Pr(\|X - z\| \leq \bar{r}(z))\right)^{N-m} (\bar{u}/m!),
\end{aligned}$$

where \bar{u} is the diameter of \mathbb{X} , that is, $\bar{u} = \sup_{x,y \in \mathbb{X}} \|x - y\| < \infty$. The last inequality holds because the last term on the left hand side is equal to the expectation of $\|U_m\|$ when $N = m$ which is bounded by \bar{u} . Consequently, we obtain:

$$N^{2/k} \|A_m^2\| < (\bar{u}/m!) N^{2/k+m} \left(1 - \underline{f} \frac{\bar{r}(z)^k}{k} \int_{S_k} \lambda_{S_k}(d\omega)\right)^{N-m} \quad (\rightarrow 0, \forall z \in \mathbb{Z}).$$

To simplify notation, let $b = \underline{f} \int_{S_k} \lambda_{S_k}(d\omega)/k$. Then:

$$N^{2/k} \|A_m^2\| < (u/m!) N^{2/k+m} (1 - b \bar{r}(z)^k)^{N-m} \leq (u/m!) N^{2/k+m} (1 - b \underline{r}^k)^{N-m},$$

where $0 < b \underline{r}^k < 1$. The right hand side of last equation is maximal at:

$$N = -\frac{(2/k + m)}{\ln(1 - b \underline{r}^k)}.$$

Therefore,

$$\begin{aligned}
N^{2/k} \|A_m^2\| &< (u/m!) \frac{(2/k + m)^{2/k+m}}{(-\ln(1 - b \underline{r}^k))^{2/k+m}} (1 - b \underline{r}^k)^{\left(\frac{-(2/k+m)}{\ln(1 - b \underline{r}^k)} - m\right)} \\
&= (2/k + m)^{(2/k+m)} (u/m!) e^{-(2/k+m)} \frac{(1 - b \underline{r}^k)^{-m}}{(-\ln(1 - b \underline{r}^k))^{2/k+m}} = C_2 < \infty.
\end{aligned}$$

Note that this bound does not depend on N or z . As a result, we obtain

$$N^{2/k} \|E[U_m]\| < C_1 + C_2 < \infty,$$

for all $N > m$ and $z \in \mathbb{Z}$. □

The proof of the second part of Theorem 2 is as follows.

$$\begin{aligned}
\text{Bias}_M^{sm,t} &= \mathbb{E}[B_M^{sm,t}] = \mathbb{E} \left[\frac{1}{N_1 M} \sum_{i=1}^N \sum_{m=1}^M W_i (\mu_0(X_i) - \mu_0(X_{j_m(i)})) \right] \\
&= \frac{1}{M} \sum_{m=1}^M \mathbb{E} [\mu_0(X_i) - \mu_0(X_{j_m(i)}) | W_i = 1].
\end{aligned}$$

Applying a second order Taylor expansion, we obtain:

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

Therefore, because the trace is a linear operator:

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

Lemma 2 implies that the norm of $N_0^{2/k} \mathbb{E}[U_{m,i} U'_{m,i} | X_i = z, W_i = 1]$ and $N_0^{2/k} \mathbb{E}[\|U_{m,i}\|^3 | X_i = z, W_i = 1]$ are uniformly bounded over $z \in \mathbb{X}_1$ and N_0 . Lemma A.2 implies the same result for $N_0^{2/k} \mathbb{E}[U_{m,i} | X_i = z, W_i = 1]$. As a result, $\|N_0^{2/k} \mathbb{E}[\mu_0(X_{j_m(i)}) - \mu_0(X_i) | X_i = z, W_i = 1]\|$ is uniformly bounded over $z \in \mathbb{X}_1$ and N_0 . Applying Lebesgue's Dominated Convergence Theorem along with Lemma 1, we obtain:

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

Now, the result follows easily from the conditions of the theorem. \square

PROOF OF LEMMA 3: Define $\underline{f} = \inf_{x,w} f_w(x)$ and $\bar{f} = \sup_{x,w} f_w(x)$, with $\underline{f} > 0$ and \bar{f} finite. Let $\bar{u} = \sup_{x,y \in \mathbb{X}} \|x-y\|$. Consider 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, because \mathbb{X} is convex, this proportion nonincreasing 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 for the probability that the distance to a match, $\|X_{j_m(i)} - X_i\|$ exceeds some value. Second, we use this to obtain an exponential bound on the volume of the catchment area, $A_M(i)$, defined as the subset of \mathbb{X} such that i is matched to each observation, j , with $W_j = 1 - W_i$ and $X_j \in A_M(i)$. That is, if $W_j = 1 - W_i$ and $X_j \in A_M(i)$, then $i \in J_M(j)$.

$$A_M(i) = \left\{ x \mid \left| \sum_l |_{W_l=W_i} 1\{\|X_l - x\| \leq \|X_i - x\|\} \right| \leq M \right\}.$$

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,

$$\begin{aligned} \Pr \left(\|X_j - X_i\| > u \cdot N_{1-W_i}^{-1/k} \mid W_1, \dots, W_N, W_j = 1 - W_i, X_i = x \right) \\ = 1 - \int_0^{u N_{1-W_i}^{-1/k}} r^{k-1} \int_{S_k} f_{1-W_i}(x + r\omega) \lambda_{S_k}(d\omega) dr \leq 1 - c \underline{f} \int_0^{u N_{1-W_i}^{-1/k}} r^{k-1} \int_{S_k} \lambda_{S_k}(d\omega) dr \\ = 1 - c \underline{f} u^k N_{1-W_i}^{-1} \pi^{k/2} / \Gamma(1+k/2). \end{aligned}$$

Similarly

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

Notice also that

$$\begin{aligned} \Pr \left(\|X_j - X_i\| > u \cdot N_{1-W_i}^{-1/k} \mid W_1, \dots, W_N, X_i = x, j \in J_M(i) \right) \\ \leq \Pr \left(\|X_j - X_i\| > u \cdot N_{1-W_i}^{-1/k} \mid W_1, \dots, W_N, X_i = x, j = j_M(i) \right) \\ = \sum_{m=0}^{M-1} \binom{N_{1-W_i}}{m} \Pr \left(\|X_j - X_i\| > u \cdot N_{1-W_i}^{-1/k} \mid W_1, \dots, W_N, W_j = 1 - W_i, X_i = x \right)^{N_{1-W_i} - m} \\ \times \Pr \left(\|X_j - X_i\| \leq u \cdot N_{1-W_i}^{-1/k} \mid W_1, \dots, W_N, W_j = 1 - W_i, X_i = x \right)^m. \end{aligned}$$

In addition,

$$\begin{aligned} \binom{N_{1-W_i}}{m} \Pr \left(\|X_j - X_i\| \leq u \cdot N_{1-W_i}^{-1/k} \mid W_1, \dots, W_N, W_j = 1 - W_i, X_i = x \right)^m \\ \leq \frac{1}{m!} \left(u^k \bar{f} \frac{\pi^{k/2}}{\Gamma(1+k/2)} \right)^m. \end{aligned}$$

Therefore,

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

Then, for some constant $C_1 > 0$,

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

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

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, \dots, W_N , $i \in \mathcal{J}_M(j)$, $X_i = x$, and $A_M(i)$, the distribution of X_j is proportional to $f_{1-W_i}(x) \cdot \mathbb{1}\{x \in A_M(i)\}$. Notice that a ball with radius $(b/2)^{1/k}/(\pi^{k/2}/\Gamma(1+k/2))^{1/k}$ has volume $b/2$. Therefore for X_i in $A_M(i)$ and $B_M(i) \geq b$, we obtain

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

The last inequality does not depend on $A_m(i)$ (given $B_M(i) \geq b$). Therefore,

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

As a result, if

$$\Pr \left(\|X_j - X_i\| > \frac{(b/2)^{1/k}}{(\pi^{k/2}/\Gamma(1+k/2))^{1/k}} \mid W_1, \dots, W_N, X_i = x, i \in \mathcal{J}_M(j) \right) \leq \delta \frac{f}{2\bar{f}}, \quad (\text{A.2})$$

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

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

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

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

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

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

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

Note that for $n \geq 0$,

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

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

Therefore, conditional on $P_M(i)$, and $X_i = x, W_1, \dots, W_N$, the q -th moment of $K_M(i)$ is

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

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

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

for some positive constant, C . Using Chernoff's bound for binomial tails, it can be easily seen that $E[(N_{1-W_i}/N_{W_i})^n \mid X_i = x, W_i] = E[(N_{1-W_i}/N_{W_i})^n \mid W_i]$ is uniformly bounded in N , for all $n \geq 1$, so the result of the first part of the lemma follows.

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

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

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

Therefore,

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

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

For part (iv) notice that

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

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

PROOF OF THEOREM 3:

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

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

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

PROOF OF THEOREM 4:

We only prove the first assertion in the theorem as the second follows the same argument. We can write $\sqrt{N}(\hat{\tau}_M^{sm} - B_M^{sm} - \tau) = \sqrt{N}(\overline{\tau(X)} - \tau) + \sqrt{N}E_M^{sm}$. First, consider the contribution of $\sqrt{N}(\overline{\tau(X)} - \tau)$. By a standard central limit theorem

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

Second, consider the contribution of $\sqrt{N}E_M^{sm}/\sqrt{V^E} = \sum_{i=1}^N E_{M,i}^{sm}/\sqrt{NV^E}$. Conditional on \mathbf{W} and \mathbf{X} the unit-level terms $E_{M,i}^{sm}$ are independent with zero means and non-identical distributions. The conditional variance of $E_{M,i}^{sm}$ is $(1 + K_M(i)/M)^2 \cdot \sigma^2(X_i, W_i)$. We will use a Lindeberg-Feller central limit theorem for $\sqrt{N}E_M^{sm}/\sqrt{V^E}$. For a given \mathbf{X}, \mathbf{W} , the Lindeberg-Feller condition requires that

$$\frac{1}{N \cdot V^E} \sum_{i=1}^N \mathbb{E} \left[(E_{M,i}^{sm})^2 1_{\{|E_{M,i}^{sm}| \geq \eta_{LF} \sqrt{N \cdot V^E}\}} | \mathbf{X}, \mathbf{W} \right] \rightarrow 0, \quad (\text{A.4})$$

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

$$\begin{aligned} & \mathbb{E} \left[(E_{M,i}^{sm})^2 1_{\{|E_{M,i}^{sm}| \geq \eta_{LF} \sqrt{N \cdot V^E}\}} | \mathbf{X}, \mathbf{W} \right] \\ & \leq \left(\mathbb{E} [(E_{M,i}^{sm})^4 | \mathbf{X}, \mathbf{W}] \right)^{1/2} \left(\mathbb{E} \left[1_{\{|E_{M,i}^{sm}| \geq \eta_{LF} \sqrt{N \cdot V^E}\}} | \mathbf{X}, \mathbf{W} \right] \right)^{1/2} \\ & \leq \left(\mathbb{E} [(E_{M,i}^{sm})^4 | \mathbf{X}, \mathbf{W}] \right)^{1/2} \left(\Pr \left(|E_{M,i}^{sm}| \geq \eta_{LF} \sqrt{NV^E} | \mathbf{X}, \mathbf{W} \right) \right) \\ & \leq \left(\mathbb{E} [(E_{M,i}^{sm})^4 | \mathbf{X}, \mathbf{W}] \right)^{1/2} \frac{\mathbb{E} [(E_{M,i}^{sm})^2 | \mathbf{X}, \mathbf{W}]}{\eta_{LF}^2 \cdot N \cdot V^E}. \end{aligned}$$

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

$$\begin{aligned} & \frac{1}{N \cdot V^E} \sum_{i=1}^N \mathbb{E} \left[(E_{M,i}^{sm})^2 1_{\{|E_{M,i}^{sm}| \geq \eta_{LF} \sqrt{N \cdot V^E}\}} | \mathbf{X}, \mathbf{W} \right] \\ & \leq \frac{1}{N \cdot V^E} \sum_{i=1}^N \left((1 + K_M(i)/M)^4 \mathbb{E} [\varepsilon_i^4 | \mathbf{X}, \mathbf{W}] \right)^{1/2} \cdot \frac{(1 + K_M(i)/M)^2 \cdot \sigma^2(X_i, W_i)}{\eta_{LF}^2 \cdot N \cdot V^E} \\ & \leq \frac{\bar{\sigma} \bar{K}^{1/2}}{\eta_{LF}^2 \underline{\sigma}^2} \cdot \frac{1}{N} \left(\frac{1}{N} \sum_{i=1}^N (1 + K_M(i)/M)^4 \right). \end{aligned}$$

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

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

Finally, $\sqrt{N} E_M^{sm} / \sqrt{V^E}$ and $\sqrt{N}(\tau(\bar{X}) - \tau)$ are asymptotically independent (the central limit theorem for $\sqrt{N} E_M^{sm} / \sqrt{V^E}$ holds conditional on \mathbf{X} and \mathbf{W}). Thus, the fact that both converge to standard normal distributions, boundedness of V^E and $V^{\tau(X)}$, and boundedness away from zero of V^E imply that $(V^E + V^{\tau(X)})^{-1/2} N^{1/2} (\hat{\tau}_M^{sm} - B_M^{sm} - \tau)$ converges to a standard normal distribution. \square

Corollaries 1 and 2 follow directly from Theorem 4 and their proofs are therefore omitted.

PROOF OF THEOREM 5:

The proof of Theorem 5 is long but mechanical, so we omit it here. In this proof, we use Lemma A.1 to characterize the asymptotic behavior of the probability of the events $\{j \in \mathcal{J}_M(1)\}$ and $\{j \in \mathcal{J}_M(1) \cup j \in \mathcal{J}_M(2)\}$ conditional on $X_j = x_j$ and $W_1 = W_2 = 1 - W_j$. These results allow us to establish the asymptotic behavior of $E[K_M(i)|W_i = w, X_i = x]$ and $E[K_M(i)^2|W_i = w, X_i = x]$, which in turn, allows us to establish the limit of $NV(\hat{\tau}_M^{sm})$. The derivation of the asymptotic expansion of the conditional probability of $\{j \in \mathcal{J}_M(1) \cup j \in \mathcal{J}_M(2)\}$ depends crucially on $k = 1$. To see why, assume, without loss of generality, that $X_2 \leq X_1$. Then, the restriction that $k = 1$ allows us to study the asymptotic behavior of the conditional probability of $\{j \in \mathcal{J}_M(1) \cup j \in \mathcal{J}_M(2)\}$ under three exhaustive cases: $X_j \leq X_2 \leq X_1$, $X_2 \leq X_j \leq X_1$, and $X_2 \leq X_1 \leq X_j$. The entire proof is available on the web pages of the authors.

Before proving Theorem 6 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 $\partial^\lambda g(x) = \partial^{|\lambda|} g(x) / \partial x_1^{\lambda_1} \dots \partial x_k^{\lambda_k}$. For $d \geq 0$, define $|g|_d = \max_{|\lambda| \leq d} \sup_x |\partial^\lambda g(x)|$.

LEMMA A.3: (UNIFORM CONVERGENCE OF SERIES ESTIMATORS OF REGRESSION FUNCTIONS, NEWEY 1995)

Suppose the conditions in Theorem 6 hold. Then for any $\xi > 0$ and non-negative integer d ,

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

for $w = 0, 1$.

PROOF: Assumptions 3.1, 4.1, 4.2 and 4.3 in Newey (1995) are satisfied for $\mu_w(x)$ and $N_w \rightarrow \infty$, implying that Newey's Theorems 4.2 and 4.4 apply. The result of the lemma holds because $N/N_w = O_p(1)$ for $w = 0, 1$. \square

LEMMA A.4: (UNIT-LEVEL BIAS CORRECTION)

Suppose the conditions in Theorem 6 hold. Then

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

for $w = 0, 1$.

PROOF: Let $U_{m,i} = X_{j_m(i)} - X_i$. Use a Taylor series expansion around X_i to write

$$\left| \mu_w(X_{j_m(i)}) - \mu_w(X_i) - \sum_{1 \leq l \leq k-1} \frac{1}{l!} \sum_{\lambda \in \Lambda_l} \partial^\lambda \mu_w(X_i) U_{m,i}^\lambda \right| \leq \frac{C^k}{k!} \sum_{\lambda \in \Lambda_k} |U_{m,i}^\lambda| \leq \frac{C^k}{k!} \sum_{\lambda \in \Lambda_k} \|U_{m,i}\|^k.$$

Because all moments of $N_{1-W_i}^{1/k} \|U_{m,i}\|$ and N/N_{1-W_i} are uniformly bounded, applying Bonferroni's and Markov's inequalities, we obtain that for any $\varepsilon > 0$:

$$\max_{i=1, \dots, N} \left| \mu_w(X_{j_m(i)}) - \mu_w(X_i) - \sum_{1 \leq l \leq k-1} \frac{1}{l!} \sum_{\lambda \in \Lambda_l} \partial^\lambda \mu_w(X_i) U_{m,i}^\lambda \right| = o_p(N^{-1+\varepsilon}).$$

Because we can choose $\varepsilon \leq 1/2$, it follows that the left hand side of last equation is $o_p(N^{-1/2})$. Similarly, for any $\varepsilon > 0$:

$$\left| \hat{\mu}_w(X_{j_m(i)}) - \hat{\mu}_w(X_i) - \sum_{1 \leq l \leq k-1} \frac{1}{l!} \sum_{\lambda \in \Lambda_l} \partial^\lambda \hat{\mu}_w(X_i) U_{m,i}^\lambda \right| \leq \frac{1}{k!} \sum_{\lambda \in \Lambda_k} |\hat{\mu}_w - \mu_w|_k \|U_{m,i}\|^k + \frac{C^k}{k!} \sum_{\lambda \in \Lambda_k} \|U_{m,i}\|^k.$$

Therefore, for arbitrary $\xi > 0$ and $\varepsilon > 0$:

$$\begin{aligned} \max_{i=1, \dots, N} \left| \hat{\mu}_w(X_{j_m(i)}) - \hat{\mu}_w(X_i) - \sum_{1 \leq l \leq k-1} \frac{1}{l!} \sum_{\lambda \in \Lambda_l} \partial^\lambda \hat{\mu}_w(X_i) U_{m,i}^\lambda \right| \\ = O_p \left(K^{1+2k} \left((K/N)^{1/2} + K^{-\xi} \right) \right) o_p(N^{-1+\varepsilon}) + o_p(N^{-1+\varepsilon}). \end{aligned}$$

Because $\nu < 2/(4k+3)$, we can choose ξ and ε so that the left hand side of last equation becomes $o_p(N^{-1/2})$. Therefore,

$$\begin{aligned} \max_{i=1, \dots, N} |\hat{\mu}_w(X_{j_m(i)}) - \hat{\mu}_w(X_i) - (\mu_w(X_{j_m(i)}) - \mu_w(X_i))| \\ \leq \max_{i=1, \dots, N} \sum_{1 \leq l \leq k-1} \frac{1}{l!} \sum_{\lambda \in \Lambda_l} |\partial^\lambda \hat{\mu}_w(X_i) - \partial^\lambda \mu_w(X_i)| \cdot |U_{m,i}^\lambda| + o_p(N^{-1/2}) \\ \leq |\hat{\mu}_w - \mu_w|_{k-1} \sum_{1 \leq l \leq k-1} \frac{1}{l!} \sum_{\lambda \in \Lambda_l} \max_{i=1, \dots, N} \|U_{m,i}\|^{|\lambda|} + o_p(N^{-1/2}) \\ = O_p \left(K^{2k-1} \left((K/N)^{1/2} + K^{-\xi} \right) \right) o_p \left(N^{-1/k+\varepsilon} \right) + o_p(N^{-1/2}), \end{aligned}$$

for arbitrary $\xi > 0$ and $\varepsilon > 0$. Now, it can be easily seen that $\nu < 2/(4k^2 - k)$ guarantees that the result of Lemma A.4 holds. \square

PROOF OF THEOREM 6:

We focus on the result for the average treatment effect. The second part of the theorem for the average effect for the treated follows the same pattern. The difference $|\hat{B}_M^{sm} - B_M^{sm}|$ can be written as

$$\begin{aligned} \left| \hat{B}_M^{sm} - B_M^{sm} \right| &\leq \frac{1}{N} \sum_{i=1}^N \frac{1}{M} \sum_{i=1}^M \left| \hat{\mu}_{1-W_i}(X_i) - \hat{\mu}_{1-W_i}(X_{j_m(i)}) - (\mu_{1-W_i}(X_i) - \mu_{1-W_i}(X_{j_m(i)})) \right| \\ &\leq \max_{i=1, \dots, N} \sum_{w=0,1} |\hat{\mu}_w(X_i) - \hat{\mu}_w(X_{j_m(i)}) - (\mu_w(X_i) - \mu_w(X_{j_m(i)}))| = o_p(N^{-1/2}), \end{aligned}$$

by Lemma A.4. \square

Before proving Theorems 7 and 8 we give one preliminary result.

LEMMA A.5: : Let $q \geq 0$. Under assumptions 1-4:

$$\frac{1}{N} \sum_{i=1}^N K_M(i)^q \left(\widehat{\sigma}_{W_i}^2(X_i) - \sigma_{W_i}^2(X_i) \right) = o_p(1).$$

Let $q \geq 1$. Under assumptions 1, 2', 3', and 4:

$$\frac{1}{N_1} \sum_{i=1}^N (1 - W_i) K_M(i)^q \left(\widehat{\sigma}_{W_i}^2(X_i) - \sigma_{W_i}^2(X_i) \right) = o_p(1).$$

PROOF: Take N_1 as given. Notice that $N_1^{1/k} \|X_i - X_{l_j(i)}\|$ is identically distributed for all i with $W_i = 1$. By Lemma 2, all the moments of $N_1^{1/k} \|X_i - X_{l_j(i)}\|$ are uniformly bounded in N_1 for all i with $W_i = 1$. Applying Bonferroni's and Markov's inequalities, we obtain

$$N_1^{-\xi} \max_{W_i=1} N_1^{1/k} \|X_i - X_{l_j(i)}\| = o_p(1),$$

as $N_1 \rightarrow \infty$, for all $\xi > 0$. In particular, making $\xi = 1/k$, we obtain

$$\max_{W_i=1} \|X_i - X_{l_j(i)}\| = o_p(1), \tag{A.5}$$

as $N_1 \rightarrow \infty$. Because $N_1 \rightarrow \infty$ almost surely, and because conditional probabilities are bounded, Lebesgue's Dominated Convergence Theorem implies that the result in equation (A.5) holds also as $N \rightarrow \infty$, without conditioning on N_1 . The analogous result holds for $W_i = 0$. Therefore,

$$\max_{i=1 \dots N} \|X_i - X_{l_j(i)}\| = \max_{w=0,1} \left\{ \max_{W_i=w} \|X_i - X_{l_j(i)}\| \right\} = o_p(1). \tag{A.6}$$

In addition, it can be seen that the maximum of number of times that an observation is used as a match within its own treatment group is bounded by $J \cdot \bar{L}(k)$, where $\bar{L}(k) < \infty$ is the "kissing number" in k dimensions. (See Lemma 3.2.1 in Miller et al. 1997. $\bar{L}(k)$ is defined as the maximum number of non-overlapping unit balls in \mathbb{R}^k that can be arranged to overlap with a unit ball.) Notice that

$$\begin{aligned} \widehat{\sigma}_{W_i}^2(X_i) &= \frac{J}{J+1} \left(Y_i - \frac{1}{J} \sum_{j=1}^J Y_{l_j(i)} \right)^2 \\ &= \frac{J}{J+1} \left(\varepsilon_i - \frac{1}{J} \sum_{j=1}^J \varepsilon_{l_j(i)} + \frac{1}{J} \sum_{j=1}^J \left(\mu_{W_i}(X_i) - \mu_{W_i}(X_{l_j(i)}) \right) \right)^2. \end{aligned}$$

Therefore,

$$\begin{aligned} E[\widehat{\sigma}_{W_i}^2(X_i) | \mathbf{X}, \mathbf{W}] &= \frac{J}{J+1} \left[\sigma_{W_i}^2(X_i) + \frac{1}{J^2} \sum_{j=1}^J \sigma_{W_i}^2(X_{l_j(i)}) + \left(\frac{1}{J} \sum_{j=1}^J \left(\mu_{W_i}(X_i) - \mu_{W_i}(X_{l_j(i)}) \right) \right)^2 \right] \\ &= \frac{J}{J+1} \left[\frac{J+1}{J} \sigma_{W_i}^2(X_i) + \frac{1}{J^2} \sum_{j=1}^J \left(\sigma_{W_i}^2(X_{l_j(i)}) - \sigma_{W_i}^2(X_i) \right) + \left(\frac{1}{J} \sum_{j=1}^J \left(\mu_{W_i}(X_i) - \mu_{W_i}(X_{l_j(i)}) \right) \right)^2 \right]. \end{aligned}$$

Using the Lipschitz conditions on $\mu_w(x)$ and $\sigma_w^2(x)$ (Assumption 4(i)), and the result in equation (A.6), we obtain:

$$\max_{i=1\dots N} \left| E[\hat{\sigma}_{W_i}^2(X_i)|\mathbf{X}, \mathbf{W}] - \sigma_{W_i}^2(X_i) \right| = o_p(1). \quad (\text{A.7})$$

Therefore,

$$\frac{1}{N} \sum_{i=1}^N K_M(i)^q \left(E[\hat{\sigma}_{W_i}^2(X_i)|\mathbf{X}, \mathbf{W}] - \sigma_{W_i}^2(X_i) \right) = o_p(1). \quad (\text{A.8})$$

To obtain the first result of the lemma, it is left to be proven that

$$\frac{1}{N} \sum_{i=1}^N K_M(i)^q \left(E[\hat{\sigma}_{W_i}^2(X_i)|\mathbf{X}, \mathbf{W}] - \hat{\sigma}_{W_i}^2(X_i) \right) = o_p(1). \quad (\text{A.9})$$

Notice first that:

$$\begin{aligned} & \left(\frac{J+1}{J} \right) \frac{1}{N} \sum_{i=1}^N K_M(i)^q \left(\hat{\sigma}_{W_i}^2(X_i) - E[\hat{\sigma}_{W_i}^2(X_i)|\mathbf{X}, \mathbf{W}] \right) = \frac{1}{N} \sum_{i=1}^N K_M(i)^q \left(\varepsilon_i^2 - \sigma_{W_i}^2(X_i) \right) \\ & + \frac{1}{N} \sum_{i=1}^N K_M(i)^q \frac{1}{J^2} \sum_{j=1}^J \left(\varepsilon_{l_j(i)}^2 - \sigma_{W_i}^2(X_{l_j(i)}) \right) + \frac{2}{J^2 N} \sum_{i=1}^N K_M(i)^q \sum_{j=1}^J \sum_{h>j}^J \varepsilon_{l_j(i)} \varepsilon_{l_h(i)} \\ & - \frac{2}{JN} \sum_{i=1}^N K_M(i)^q \varepsilon_i \sum_{j=1}^J \varepsilon_{l_j(i)} + \frac{2}{JN} \sum_{i=1}^N K_M(i)^q \varepsilon_i \sum_{j=1}^J \left(\mu_{W_i}(X_i) - \mu_{W_i}(X_{l_j(i)}) \right) \\ & \quad - \frac{2}{J^2 N} \sum_{i=1}^N K_M(i)^q \sum_{j=1}^J \varepsilon_{l_j(i)} \sum_{j=1}^J \left(\mu_{W_i}(X_i) - \mu_{W_i}(X_{l_j(i)}) \right). \quad (\text{A.10}) \end{aligned}$$

The expectations, conditional on \mathbf{X} and \mathbf{W} , of each term on the right hand side of last equation is equal to zero, so the unconditional expectations are also zero. Applying Bonferroni's Inequality, it is easy to show that for any $\xi > 0$, $E[N^{-\xi} \max_{i=1, \dots, N} K_M(i)^{2q}]$ is uniformly bounded. Using this result, along with finiteness of $\bar{L}(k)$, it can be shown that the variances of all the terms on the right hand side of equation (A.10) are $o(1)$. (This last part of the proof is largely technical and therefore omitted here. It is available on the websites of the authors.)

The proof of the second part of the Lemma follows the same pattern and is omitted here. \square

PROOF OF THEOREM 7: It follows directly from Lemma A.5. \square

PROOF OF THEOREM 8:

For part (i) notice that

$$\begin{aligned} \frac{1}{N} \sum_{i=1}^N \left(\hat{Y}_i(1) - \hat{Y}_i(0) - \hat{\tau}_M^{sm} \right)^2 &= \frac{1}{N} \sum_{i=1}^N \left(\hat{Y}_i(1) - \hat{Y}_i(0) - \tau \right)^2 - (\hat{\tau}_M^{sm} - \tau)^2 \\ &= \frac{1}{N} \sum_{i=1}^N \left(\hat{Y}_i(1) - \hat{Y}_i(0) - \tau \right)^2 + o_p(1). \quad (\text{A.11}) \end{aligned}$$

In addition,

$$\begin{aligned}
& \frac{1}{N} \sum_{i=1}^N \left(\widehat{Y}_i(1) - \widehat{Y}_i(0) - \tau \right)^2 \\
&= \frac{1}{N} \sum_{i=1}^N \left((2W_i - 1) \left(\frac{1}{M} \sum_{m=1}^M \mu_{W_i}(X_i) - \mu_{1-W_i}(X_{j_m(i)}) \right) - \tau \right)^2 + \frac{1}{N} \sum_{i=1}^N \left(\varepsilon_i - \frac{1}{M} \sum_{m=1}^M \varepsilon_{j_m(i)} \right)^2 \\
&+ \frac{2}{N} \sum_{i=1}^N \left((2W_i - 1) \left(\frac{1}{M} \sum_{m=1}^M \mu_{W_i}(X_i) - \mu_{1-W_i}(X_{j_m(i)}) \right) - \tau \right) (2W_i - 1) \left(\varepsilon_i - \frac{1}{M} \sum_{m=1}^M \varepsilon_{j_m(i)} \right). \tag{A.12}
\end{aligned}$$

Because the sample maximum of the norms of the matching discrepancies, $\|X_i - X_{j_m(i)}\|$, is $o_p(1)$, and the regression functions, μ_w , are Lipschitz, we obtain

$$\frac{1}{N} \sum_{i=1}^N \left(\frac{1}{M} \sum_{m=1}^M \mu_{1-W_i}(X_i) - \mu_{1-W_i}(X_{j_m(i)}) \right)^2 = o_p(1). \tag{A.13}$$

Consider the first term on the right hand side of equation (A.12):

$$\begin{aligned}
& \frac{1}{N} \sum_{i=1}^N \left((2W_i - 1) \left(\frac{1}{M} \sum_{m=1}^M \mu_{W_i}(X_i) - \mu_{1-W_i}(X_{j_m(i)}) \right) - \tau \right)^2 \\
&= \frac{1}{N} \sum_{i=1}^N \left((\mu_1(X_i) - \mu_0(X_i) - \tau) + (2W_i - 1) \left(\frac{1}{M} \sum_{m=1}^M \mu_{1-W_i}(X_i) - \mu_{1-W_i}(X_{j_m(i)}) \right) \right)^2 \\
&= \frac{1}{N} \sum_{i=1}^N (\mu_1(X_i) - \mu_0(X_i) - \tau)^2 + \frac{1}{N} \sum_{i=1}^N \left(\frac{1}{M} \sum_{m=1}^M \mu_{1-W_i}(X_i) - \mu_{1-W_i}(X_{j_m(i)}) \right)^2 \\
&+ \frac{1}{N} \sum_{i=1}^N (\mu_1(X_i) - \mu_0(X_i) - \tau) (2W_i - 1) \left(\frac{1}{M} \sum_{m=1}^M \mu_{1-W_i}(X_i) - \mu_{1-W_i}(X_{j_m(i)}) \right) \\
&= \frac{1}{N} \sum_{i=1}^N (\mu_1(X_i) - \mu_0(X_i) - \tau)^2 + o_p(1), \tag{A.14}
\end{aligned}$$

by Hölder's Inequality and equation (A.13). Next, consider the second term on the right hand side of equation (A.12):

$$\frac{1}{N} \sum_{i=1}^N \left(\varepsilon_i - \frac{1}{M} \sum_{m=1}^M \varepsilon_{j_m(i)} \right)^2 = \frac{1}{N} \sum_{i=1}^N \left(\varepsilon_i^2 + \frac{1}{M^2} \left(\sum_{m=1}^M \varepsilon_{j_m(i)}^2 + 2 \sum_{m=1}^M \sum_{n>m} \varepsilon_{j_m(i)} \varepsilon_{j_n(i)} \right) - \frac{2}{M} \sum_{m=1}^M \varepsilon_i \varepsilon_{j_m(i)} \right).$$

Therefore,

$$\begin{aligned}
& \frac{1}{N} \sum_{i=1}^N \left(\varepsilon_i - \frac{1}{M} \sum_{m=1}^M \varepsilon_{j_m(i)} \right)^2 - \frac{1}{N} \sum_{i=1}^N \left(1 + \frac{K_M(i)}{M^2} \right) \sigma^2(X_i, W_i) \\
&= \frac{1}{N} \sum_{i=1}^N \left(1 + \frac{K_M(i)}{M^2} \right) (\varepsilon_i^2 - \sigma^2(X_i, W_i)) \\
&+ \frac{1}{N} \sum_{i=1}^N \frac{2}{M^2} \left(\sum_{m=1}^M \sum_{n>m} \varepsilon_{j_m(i)} \varepsilon_{j_n(i)} \right) - \frac{1}{N} \sum_{i=1}^N \frac{2}{M} \sum_{m=1}^M \varepsilon_i \varepsilon_{j_m(i)}. \tag{A.15}
\end{aligned}$$

The expectations conditional on \mathbf{X} and \mathbf{W} of each of the three terms on the right hand side of last expression are zero, so the unconditional expectations are also zero. Because the fourth conditional moments of ε_i are uniformly bounded, we obtain:

$$\begin{aligned} E \left[\left(\frac{1}{N} \sum_{i=1}^N \left(1 + \frac{K_M(i)}{M^2} \right) (\varepsilon_i^2 - \sigma^2(X_i, W_i)) \right)^2 \right] \\ = \frac{1}{N} E \left[\frac{1}{N} \sum_{i=1}^N \left(1 + \frac{K_M(i)}{M^2} \right)^2 (\varepsilon_i^2 - \sigma^2(X_i, W_i))^2 \right] = o(1). \end{aligned}$$

The variance of the second term divided by $4/M^4$ is

$$\begin{aligned} E \left[\left(\frac{1}{N} \sum_{i=1}^N \sum_{m=1}^M \sum_{n>m} \varepsilon_{j_m(i)} \varepsilon_{j_n(i)} \right)^2 \right] &= \frac{1}{N} E \left[\frac{1}{N} \sum_{i=1}^N \left(\sum_{m=1}^M \sum_{n>m} \varepsilon_{j_m(i)} \varepsilon_{j_n(i)} \right)^2 \right] \\ &+ \frac{2}{N} E \left[\frac{1}{N} \sum_{i=1}^N \sum_{j>i} \left(\sum_{m=1}^M \sum_{n>m} \varepsilon_{j_m(i)} \varepsilon_{j_n(i)} \right) \left(\sum_{m=1}^M \sum_{n>m} \varepsilon_{j_m(j)} \varepsilon_{j_n(j)} \right) \right] \\ &\leq \frac{1}{N} E \left[\frac{1}{N} \sum_{i=1}^N \frac{(M-1)M}{2} \bar{\sigma}^4 \right] \\ &+ \frac{1}{N} E \left[\frac{1}{N} \sum_{i=1}^N (M-1)K_M(i)(K_M(i)-1)\bar{\sigma}^4 \right] = o(1). \end{aligned}$$

(Last inequality holds because there are $K_M(i)(K_M(i)-1)/2$ combinations of two different sets of matches, $J_M(i_1)$ and $J_M(i_2)$ with $1 \leq i_1 < i_2 \leq N$, with observation i in both sets. Moreover, for each of these combinations, there are at most $M-1$ terms of the form $\varepsilon_{j_{m_1}(i_1)} \varepsilon_{j_{n_1}(i_1)} \varepsilon_{j_{m_2}(i_2)} \varepsilon_{j_{n_2}(i_2)}$, containing ε_i^2 and with non-zero expectations.)

The variance of the third term divided by $4/M^2$ is

$$\begin{aligned} E \left[\left(\frac{1}{N} \sum_{i=1}^N \sum_{m=1}^M \varepsilon_i \varepsilon_{j_m(i)} \right)^2 \right] &= \frac{1}{N} E \left[\frac{1}{N} \sum_{i=1}^N \left(\sum_{m=1}^M \varepsilon_i \varepsilon_{j_m(i)} \right)^2 \right] \\ &+ \frac{2}{N} E \left[\frac{1}{N} \sum_{i=1}^N \left(\sum_{m=1}^M \varepsilon_i \varepsilon_{j_m(i)} \right) \sum_{j>i} \left(\sum_{m=1}^M \varepsilon_j \varepsilon_{j_m(j)} \right) \right] \\ &\leq \frac{1}{N} E \left[\frac{1}{N} \sum_{i=1}^N M \bar{\sigma}^4 \right] \\ &+ \frac{2}{N} E \left[\frac{1}{N} \sum_{i=1}^N M \bar{\sigma}^4 \right] = o(1). \end{aligned}$$

As a result, we obtain

$$\frac{1}{N} \sum_{i=1}^N \left(\varepsilon_i - \frac{1}{M} \sum_{m=1}^M \varepsilon_{j_m(i)} \right)^2 - \frac{1}{N} \sum_{i=1}^N \left(1 + \frac{K_M(i)}{M^2} \right) \sigma^2(X_i, W_i) = o_p(1).$$

Finally, consider the last term on the right hand side of equation (A.12). Let

$$\Psi_{M,i} = \left((2W_i - 1) \left(\frac{1}{M} \sum_{m=1}^M \mu_{W_i}(X_i) - \mu_{1-W_i}(X_{j_m(i)}) \right) - \tau \right).$$

Notice that there is a finite bound $\bar{\Psi}$, such that $|\Psi_{M,i}| \leq \bar{\Psi}$ for all i . The conditional expectation of the last term of equation (A.12) is zero, so the unconditional expectation is also zero. The conditional variance of this term (divided by 4) is:

$$\begin{aligned} & E \left[\frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N \Psi_{M,i} \Psi_{M,j} \left(\varepsilon_i - \frac{1}{M} \sum_{m=1}^M \varepsilon_{j_m(i)} \right) \left(\varepsilon_j - \frac{1}{M} \sum_{m=1}^M \varepsilon_{j_m(j)} \right) \right] \\ & \leq \left| E \left[\frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N \Psi_{M,i} \Psi_{M,j} \varepsilon_i \varepsilon_j \right] \right| + 2 \left| E \left[\frac{1}{MN^2} \sum_{i=1}^N \sum_{j=1}^N \Psi_{M,i} \Psi_{M,j} \varepsilon_i \sum_{m=1}^M \varepsilon_{j_m(j)} \right] \right| \\ & \quad + \left| E \left[\frac{1}{M^2 N^2} \sum_{i=1}^N \sum_{j=1}^N \Psi_{M,i} \Psi_{M,j} \sum_{m=1}^M \varepsilon_{j_m(i)} \sum_{m=1}^M \varepsilon_{j_m(j)} \right] \right|. \end{aligned}$$

Notice that:

$$\left| E \left[\frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N \Psi_{M,i} \Psi_{M,j} \varepsilon_i \varepsilon_j \right] \right| = \left| E \left[\frac{1}{N^2} \sum_{i=1}^N \Psi_{M,i}^2 \varepsilon_i^2 \right] \right| \leq \frac{\bar{\Psi}^2}{N} E \left[\frac{1}{N} \sum_{i=1}^N \varepsilon_i^2 \right] = o(1),$$

$$\begin{aligned} \left| E \left[\frac{1}{MN^2} \sum_{i=1}^N \sum_{j=1}^N \Psi_{M,i} \Psi_{M,j} \varepsilon_i \sum_{m=1}^M \varepsilon_{j_m(j)} \right] \right| & \leq E \left[\frac{\bar{\Psi}}{MN^2} \sum_{i=1}^N |\Psi_{M,i}| K_M(i) \varepsilon_i^2 \right] \\ & \leq \frac{\bar{\Psi}^2}{MN} E \left[\frac{1}{N} \sum_{i=1}^N K_M(i) \varepsilon_i^2 \right] = o(1), \end{aligned}$$

$$\begin{aligned} \left| E \left[\frac{1}{M^2 N^2} \sum_{i=1}^N \sum_{j=1}^N \Psi_{M,i} \Psi_{M,j} \sum_{m=1}^M \varepsilon_{j_m(i)} \sum_{m=1}^M \varepsilon_{j_m(j)} \right] \right| & \leq \left| E \left[\frac{1}{M^2 N^2} \sum_{i=1}^N \Psi_{M,i}^2 \left(\sum_{m=1}^M \varepsilon_{j_m(i)} \right)^2 \right] \right| \\ + 2 \left| E \left[\frac{1}{M^2 N^2} \sum_{i=1}^N \sum_{j>i}^N \Psi_{M,i} \Psi_{M,j} \sum_{m=1}^M \varepsilon_{j_m(i)} \sum_{m=1}^M \varepsilon_{j_m(j)} \right] \right| & \leq \frac{\bar{\Psi}^2}{N} E \left[\frac{1}{MN} \sum_{i=1}^N \sigma^2 \right] \\ & \quad + \frac{\bar{\Psi}^2}{N} E \left[\frac{1}{M^2 N} \sum_{i=1}^N K_M(i) (K_M(i) - 1) \varepsilon_i^2 \right] = o(1). \end{aligned}$$

As a result, we obtain:

$$\begin{aligned} \frac{1}{N} \sum_{i=1}^N \left(\hat{Y}_i(1) - \hat{Y}_i(0) - \hat{\tau}_M^{sm} \right)^2 & = \frac{1}{N} \sum_{i=1}^N (\mu_1(X_i) - \mu_0(X_i) - \tau)^2 \\ & \quad + \frac{1}{N} \sum_{i=1}^N \left(1 + \frac{K_M(i)}{M^2} \right) \sigma^2(X_i, W_i) + o_p(1) \end{aligned}$$

Applying the previous lemma:

$$\left| \frac{1}{N} \sum_{i=1}^N \left(1 + \frac{K_M(i)}{M^2} \right) \sigma^2(X_i, W_i) - \frac{1}{N} \sum_{i=1}^N \left(1 + \frac{K_M(i)}{M^2} \right) \hat{\sigma}^2(X_i, W_i) \right| = o_p(1).$$

Therefore,

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

From the lemma, we know that

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

Putting the two pieces together:

$$\begin{aligned} \widehat{V} &= \widehat{V}^E + \widehat{V}^{\tau(X)} \\ &= \frac{1}{N} \sum_{i=1}^N \left(\widehat{Y}_i(1) - \widehat{Y}_i(0) - \widehat{\tau}_M^{sm} \right)^2 \\ &\quad + \frac{1}{N} \sum_{i=1}^N \left[\left(\frac{K_M(i)}{M} \right)^2 + \left(\frac{2M-1}{M} \right) \left(\frac{K_M(i)}{M} \right) \right] \widehat{\sigma}^2(X_i, W_i) \xrightarrow{p} V^E + V^{\tau(X)}, \end{aligned}$$

finishing the proof for the first part of the theorem.

The proof for the second part follows the same pattern and is omitted here. It is available from the authors on the web. \square

REFERENCES

- ABADIE, A. (2003), "Semiparametric Instrumental Variable Estimation of Treatment Response Models," *Journal of Econometrics*, 113(2), 231-263.
- 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., AND J. HAHN, (2003) "When to Control for Covariates? Panel-Asymptotic Results for Estimates of Treatment Effects," *Review of Economics and Statistics*, forthcoming.
- 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. (1978), "Estimating the Effect of Training Programs on Earnings," *Review of Economics and Statistics*, 60, 47-57.
- 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.
- FIRPO, S. (2003), "Efficient Semiparametric Estimation of Quantile Treatment Effects," PhD Thesis, Chapter 2, Department of Economics, University of California, Berkeley.
- 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, AND P. TODD, (1998), "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies* 65, 261-294.
- 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, (2003), "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*.
- HOTZ J., G. IMBENS, AND J. MORTIMER (1999), "Predicting the Efficacy of Future Training Programs Using Past Experiences" NBER Working Paper.
- HOROWITZ, J., AND E. MAMMEN, (2002), "Nonparametric Estimation of an Additive Model with a Link Function," unpublished manuscript.
- ICHIMURA, H., AND O. LINTON, (2001), "Trick or Treat: Asymptotic Expansions for some Semiparametric Program Evaluation Estimators." unpublished manuscript, London School of Economics.
- IMBENS, G. (2003a), "Sensitivity to Exogeneity Assumptions in Program Evaluation," *American Economic Review Papers and Proceedings*, 93(2), 126-132.
- IMBENS, G. (2003b), "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Survey," forthcoming, *Review of Economics and Statistics*.
- JOHNSON, N., S. KOTZ, AND A. KEMP, (1992), *Univariate Discrete Distributions*, 2nd Edition, Wiley, New York.
- LALONDE, R.J., (1986), "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *American Economic Review*, 76, 604-620.
- LECHNER, M, (1999), "Earnings and Employment Effects of Continuous Off-the-job Training in East Germany After Unification," *Journal of Business and Economic Statistics*, 17(1), 74-90.
- LITTLE, R.J.A., AND D.B. RUBIN, (2002): *Statistical Analysis with Missing Data*, John Wiley & Sons, Hoboken, NJ.
- 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.
- MILLER, G.L., S. TENG, W. THURSTON, AND S.A. VAVASIS, (1997), "Separators for Sphere-Packings and Nearest Neighbor Graphs," *Journal of the ACM*, 44, 1-29.

- NEWHEY, W.K., (1995) "Convergence Rates for Series Estimators," in G.S. Maddala, P.C.B. Phillips and T.N. Srinivasan 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., AND Y. RITOV, (1997), "Towards a Curse of Dimensionality Appropriate (CODA) Asymptotic Theory for Semi-parametric Models," *Statistics in Medicine* 16, 285-319.
- 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., (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., AND D. RUBIN, (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects", *Biometrika*, 70, 41-55.
- 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.
- RUBIN, D., AND N. THOMAS, (1992a), "Affinely Invariant Matching Methods with Ellipsoidal Distributions," *Annals of Statistics* 20 (2) 1079-1093.
- RUBIN, D., AND N. THOMAS, (1992b), "Characterizing the effect of matching using linear propensity score methods with normal distributions," *Biometrika* 79 797-809.
- 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.
- SMITH, J. A. AND P. E. TODD, (2003), "Does Matching Address LaLonde's Critique of Nonexperimental Estimators," forthcoming in *Journal of Econometrics*.
- 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: SUMMARY STATISTICS

| | Experimental Data | | | | PSID | | T-statistic | |
|---------------------------------|----------------------------|--------|-----------------------------|--------|---------------------|---------|-----------------|----------------|
| | Treated (185 obs.) mean | (s.d.) | Controls (260 obs.) mean | (s.d.) | (2490 obs.) mean | (s.d.) | Treat/ Contr | Treat/ 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 13-24 | 2.10 | (4.89) | 2.11 | (5.69) | 19.43 | (13.41) | [-0.0] | [-38.6] |
| Unemployed 13-24 | 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] |
| Unemployed '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] |
| Unemployed '78 | 0.24 | (0.43) | 0.35 | (0.48) | 0.11 | (0.32) | [-2.7] | [4.0] |

Note: Earnings data are in thousands of 1978 dollars. Earnings 13-24 and Unemployed 13-24 refers to earnings and unemployment during the period 13 to 24 months prior to randomization.

TABLE 2: EXPERIMENTAL AND NON-EXPERIMENTAL ESTIMATES FOR THE NSW DATA

| | $M = 1$ | | $M = 4$ | | $M = 16$ | | $M = 64$ | | $M = 2490$ | |
|-------------------------------------|---------|--------|---------|--------|----------|--------|----------|--------|------------|--------|
| | est | (s.e.) | est | (s.e.) | est | (s.e.) | est | (s.e.) | est | (s.e.) |
| Panel A: Experimental Estimates | | | | | | | | | | |
| simple matching | 1.22 | (0.84) | 1.99 | (0.74) | 1.75 | (0.74) | 2.20 | (0.70) | 1.80 | (0.67) |
| bias-adjusted matching | 1.16 | (0.84) | 1.84 | (0.74) | 1.54 | (0.75) | 1.74 | (0.71) | 1.72 | (0.68) |
| Regression Estimates | | | | | | | | | | |
| mean difference | 1.80 | (0.67) | | | | | | | | |
| linear | 1.72 | (0.65) | | | | | | | | |
| quadratic | 2.27 | (0.73) | | | | | | | | |
| Panel B: Non-experimental Estimates | | | | | | | | | | |
| simple matching | 2.07 | (1.13) | 1.62 | (0.91) | 0.47 | (0.85) | -0.11 | (0.75) | -15.20 | (0.61) |
| bias-adjusted matching | 2.42 | (1.13) | 2.51 | (0.90) | 2.48 | (0.83) | 2.26 | (0.71) | 0.84 | (0.63) |
| Regression Estimates | | | | | | | | | | |
| mean difference | -15.20 | (0.66) | | | | | | | | |
| linear | 0.84 | (0.86) | | | | | | | | |
| quadratic | 3.26 | (0.98) | | | | | | | | |

Note: The outcome is earnings in 1978 in thousands of dollars.

TABLE 3
MEAN COVARIATE DIFFERENCES IN MATCHED GROUPS

| | Average | | $M = 1$ | | $M = 4$ | | $M = 16$ | | $M = 64$ | | $M = 2490$ | |
|------------------|---------|---------|---------|--------|---------|--------|----------|--------|----------|--------|------------|--------|
| | PSID | Treated | mean | (s.d.) | mean | (s.d.) | mean | (s.d.) | mean | (s.d.) | mean | (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 13-24 | 0.09 | -1.18 | -0.01 | (0.10) | -0.01 | (0.12) | -0.05 | (0.17) | -0.15 | (0.30) | -1.26 | (0.36) |
| Unemployed 13-24 | -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) |
| Unemployed '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 treated and the PSID comparison units. 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 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 4
SIMULATION RESULTS (10,000 REPLICATIONS)

| M | Estimator | mean | median | rmse | mae | s.d. | mean | coverage rate | |
|-----|----------------------|--------|--------|-------|-------|------|------|---------------|------------|
| | | bias | bias | | | | s.e. | (nom. 95%) | (nom. 90%) |
| 1 | simple matching | -0.49 | -0.44 | 0.87 | 0.53 | 0.73 | 0.88 | 0.94 | 0.90 |
| | linear bias-adjusted | 0.05 | 0.08 | 0.73 | 0.47 | 0.73 | 0.89 | 0.96 | 0.94 |
| 4 | simple matching | -0.83 | -0.82 | 1.03 | 0.84 | 0.59 | 0.63 | 0.75 | 0.62 |
| | linear bias-adjusted | 0.05 | 0.07 | 0.61 | 0.39 | 0.60 | 0.64 | 0.95 | 0.91 |
| 16 | simple matching | -1.81 | -1.74 | 1.88 | 1.79 | 0.57 | 0.53 | 0.08 | 0.04 |
| | linear bias-adjusted | 0.19 | 0.18 | 0.63 | 0.41 | 0.60 | 0.53 | 0.90 | 0.84 |
| 64 | simple matching | -3.17 | -3.24 | 3.33 | 3.25 | 0.61 | 0.52 | 0.00 | 0.00 |
| | linear bias-adjusted | 0.17 | 0.17 | 0.67 | 0.44 | 0.65 | 0.52 | 0.87 | 0.80 |
| | mean difference | -19.06 | -19.09 | 19.06 | 19.09 | 0.61 | 1.63 | 0.00 | 0.00 |
| | linear regression | -2.04 | -2.06 | 2.26 | 2.06 | 1.00 | 0.98 | 0.44 | 0.33 |
| | quadratic regression | 2.72 | 2.65 | 3.01 | 2.65 | 1.35 | 1.24 | 0.40 | 0.27 |