

# The Variance of Non-Parametric Treatment Effect Estimators in the Presence of Clustering\*

Samuel G. Hanson

Adi Sunderam

Harvard University

Harvard University

shanson@fas.harvard.edu

sunderam@fas.harvard.edu

February 28, 2011

## Abstract

Non-parametric estimators of treatment effects are often applied in settings where clustering may be important. We provide a general methodology for consistently estimating the variance of a large class of non-parametric estimators, including the simple matching estimator, in the presence of clustering. Software for implementing our variance estimator is available in Stata.

*Key Words:* Treatment effects, matching estimators, clustering.

*JEL Classification:* C14, C21

---

\*We would like to thank Sergey Chernenko, Judd Kessler, participants at the Harvard Econometrics Seminar, Alberto Abadie (the editor), an anonymous referee, and especially Guido Imbens for extremely helpful comments and suggestions.

# 1 Introduction

Average treatment effects (ATEs) can be estimated using a variety of non-parametric techniques, including matching and propensity score based estimators. These methods are usually applied in settings, such as program evaluation, where cross-sectional data have been collected. With such data, geographic shocks or omitted common factors may induce correlation across observations, even after controlling for treatment status and the covariates used in the research design. In other words, there may be a clustering problem. While a large literature has studied the consistency of non-parametric ATE estimators (see Imbens 2004), there has been little discussion of the effects of clustering on the variance of these estimators. This is surprising given the significant attention devoted to clustering in parametric settings. Given the popularity of non-parametric estimators, and matching estimators in particular, the ability to compute cluster-robust standard errors for them is important.

In this note we consider a setting where both residuals and treatment effects may be correlated within clusters. We provide a methodology for estimating the variance of matching estimators in this setting. There are two main challenges. First, matching estimators are highly non-smooth functionals of the data and, as discussed further below, standard asymptotic arguments for smooth functionals (e.g. method of moments estimators) cannot be applied (Abadie and Imbens 2006). Second, since they do not rely on consistent estimation of the underlying regression functions, matching methods do not generate estimated residuals, which are crucial to standard clustering adjustments (Moulton 1990, Wooldridge 2002). We surmount these challenges by extending the approach of Abadie and Imbens (2006 and 2008). Our methodology generates “quasi-residuals,” which we use to compute a cluster-robust variance estimator that is consistent as the number of clusters grows large. While we focus on the matching estimator in this note, our methodology can easily be extended to a broader class of non-parametric treatment effect estimators.

The remainder of the paper is organized as follows. Section 2 defines the relevant class of matching estimators for the ATE and derives the clustering correction. Section 3 details our methodology for cluster-robust variance estimation. In Section 4 we explore the finite-sample behavior of our variance estimator using a short Monte Carlo study. Section 5 briefly discusses extensions of our basic methodology and Section 6 concludes.

## 2 Preliminaries

### 2.1 Setup and Notation

We consider matching estimators for the average effect of a binary treatment on some outcome. Let  $j = 1, \dots, J$  index clusters,  $i = 1, \dots, I_j$  index individuals in cluster  $j$ , and  $N = \sum_j I_j$  be the total number of individuals. Let  $X_{ij}$  denote the vector of covariates and  $W_{ij} \in \{0, 1\}$  indicate the treatment status of individual  $ij$ . Also, let  $\mathbf{X}$  denote the matrix of covariates and  $\mathbf{W}$  the vector of treatment indicators for all  $N$  individuals. Let  $(Y_{ij}(0), Y_{ij}(1))$  denote the potential outcomes given the control treatment or the active treatment, respectively, for individual  $ij$ . Of course, only  $Y_{ij} = Y_{ij}(W_{ij})$  is observed. Under the standard unconfoundedness assumption that  $W_{ij}$  is independent of  $(Y_{ij}(0), Y_{ij}(1))$  conditional on  $X_{ij}$ , we can write the conditional expectation of the outcome given treatment  $w$  and covariates  $X = x$  as  $\mu_w(x) = E[Y(w) | X = x]$ . The average treatment effect for the subpopulation with  $X = x$  is  $\tau(x) = E[Y(1) - Y(0) | X = x] = \mu_1(x) - \mu_0(x)$ . The population average treatment effect (PATE) is  $\tau = E[\tau(X)]$  and the sample average treatment effect (SATE), conditional on  $\mathbf{X}$ , is  $\overline{\tau(\mathbf{X})} = N^{-1} \sum_{i,j} \tau(X_{ij})$ .

The matching estimator imputes unobserved potential outcomes for each individual by matching that individual with  $M$  individuals of the opposite treatment status. Specifically, let  $\mathcal{J}_M(ij)$  be the set of indices of the  $M$  closest matches to unit  $ij$  with *the opposite treatment status* (i.e.  $\#\mathcal{J}_M(ij) = M$  and for all  $st \in \mathcal{J}_M(ij)$ ,  $W_{st} = 1 - W_{ij}$  and  $\|X_{ij} - X_{st}\| \leq \|X_{ij} - X_{s't'}\|$  for all  $s't' \notin \mathcal{J}_M(ij)$ ). The estimator imputes missing outcomes as

$$\widehat{Y}_{ij}(0) = \begin{cases} Y_{ij} & \text{if } W_{ij} = 0 \\ \frac{1}{M} \sum_{st \in \mathcal{J}_M(ij)} Y_{st} & \text{if } W_{ij} = 1 \end{cases} \quad \text{and} \quad \widehat{Y}_{ij}(1) = \begin{cases} \frac{1}{M} \sum_{st \in \mathcal{J}_M(ij)} Y_{st} & \text{if } W_{ij} = 0 \\ Y_{ij} & \text{if } W_{ij} = 1 \end{cases}.$$

Defining  $K_M(ij) = \sum_{st} \mathbf{1}\{i, j \in \mathcal{J}_M(st)\}$  as the number of times observation  $ij$  is used as a match, the simple matching estimator is

$$\widehat{\tau}_M = N^{-1} \sum_{i,j} \left( \widehat{Y}_{ij}(1) - \widehat{Y}_{ij}(0) \right) = N^{-1} \sum_{i,j} (2W_{ij} - 1) \left( 1 + \frac{K_M(ij)}{M} \right) Y_{ij}.$$

We consider a random-effects type setting in the sense that individual treatment effects,  $Y_{ij}(1) - Y_{ij}(0)$ , are assumed to be independent of the cluster-level shocks. Specifically, we assume that all members of cluster  $j$  are subject to the same cluster-level shock regardless

of treatment status. Thus, potential outcomes can be decomposed as  $Y_{ij}(w) = \mu_w(X_{ij}) + \eta_j + \omega_{ij}(w)$  and we have  $Y_{ij}(1) - Y_{ij}(0) = \tau(X_{ij}) + \omega_{ij}(1) - \omega_{ij}(0)$  which is independent of  $\eta_j$ . In what follows, we suppress the  $\omega_{ij}(w)$  notation and simply write observed outcomes as  $Y_{ij} = \mu_{W_{ij}}(X_{ij}) + \varepsilon_{ij} = \mu_{W_{ij}}(X_{ij}) + \eta_j + \omega_{ij}$ . For simplicity, we assume that the cluster-level shocks are homoskedastic but the individual-specific shocks may be heteroskedastic. That is, we assume  $\varepsilon_{ij} = \eta_j + \omega_{ij}$ , where  $\eta_j \stackrel{iid}{\sim} (0, \sigma_\eta^2)$  and  $\omega_{ij} \stackrel{iid}{\sim} (0, \sigma_\omega^2(X_{ij}, W_{ij}))$ .<sup>1</sup> In addition to cluster-level correlation of the residuals, in many empirical settings it is likely that the individual treatment effects are correlated within clusters. Therefore, we also allow the covariates  $X_{ij}$  and hence the treatment effects  $\tau(X_{ij})$  to be correlated within clusters, but we assume they are independent across clusters.

Many empirical settings fit the assumptions laid out here. For instance, it would be sensible to assume clustering of both treatment effects and shocks at the school level for program evaluation exercises with school level pay-for-grades treatments within schools districts (e.g. Fryer 2010). Clustering at the county level would be expected when evaluating individual specific job-training treatments within counties (e.g. Hotz, Imbens, and Klerman 2006).

## 2.2 The Variance of Matching Estimators with Clustering

Following Abadie and Imbens (2006) we write the difference between  $\hat{\tau}_M$  and  $\tau$  as  $\hat{\tau}_M - \tau = \overline{(\tau(\mathbf{X}) - \tau)} + E + B$  where  $\overline{(\tau(\mathbf{X}) - \tau)} = N^{-1} \sum_{i,j} \tau(X_{ij})$  is the sample average treatment effect conditional on  $\mathbf{X}$  and

$$\begin{aligned}
 E &= N^{-1} \sum_{i,j} (2W_{ij} - 1) \left( 1 + \frac{K_M(ij)}{M} \right) \varepsilon_{ij} \\
 B &= N^{-1} \sum_{i,j} \left[ (2W_{ij} - 1) \left( 1 + \frac{K_M(ij)}{M} \right) \mu_{W_{ij}}(X_{ij}) - (\mu_1(X_{ij}) - \mu_0(X_{ij})) \right].
 \end{aligned}$$

Here  $E$  is a weighted sum of the residuals, and  $B$  is a conditional bias term.

As pointed out by Imbens (2004), the variance of the matching estimator depends on the quantity we are trying to estimate. If the SATE is the estimand of interest, then the normal-

---

<sup>1</sup>While we assume  $\eta_j \stackrel{iid}{\sim} (0, \sigma_\eta^2)$  for simplicity, our approach can accommodate certain forms of heteroskedasticity for the  $\eta_j$ . For instance, suppose that  $X_{ij} = (X_j^1, X_j^2)'$  where  $X_j^1$  is a vector of covariates that is constant within clusters and  $X_j^2$  is a vector of covariates that varies within clusters. Our approach is robust to forms of heteroskedasticity where the variance of the group-level shock is a function of the  $X_j^1$  and the variance individual specific errors depends on both  $X_{ij}$  and  $W_{ij}$  (i.e. where  $\sigma_\varepsilon^2(X_{ij}, W_{ij}) = \sigma_\eta^2(X_j^1) + \sigma_\omega^2(X_{ij}, W_{ij})$ ).

ized variance of the estimator is given by the *conditional variance*:  $V^E \equiv \text{Var}[\sqrt{N}\widehat{\tau}_M|\mathbf{X}, \mathbf{W}] = \text{Var}[\sqrt{N}E|\mathbf{X}, \mathbf{W}]$ . If the PATE (i.e.  $\tau = E[\tau(X)]$ ) is the estimand of interest, then the normalized variance of the estimator is the *marginal variance*  $V = V^E + V^{\tau(X)}$  where  $V^{\tau(X)} \equiv \text{Var}[\sqrt{N}\overline{\tau}(\mathbf{X})]$  is the normalized variance of the SATE. In this note, we provide cluster-robust estimators for both the conditional and marginal variance.

Under our assumed error structure, we can write the conditional variance as

$$V^E = N^{-1} \sum_{i,j} \left(1 + \frac{K_M(ij)}{M}\right)^2 \sigma_\varepsilon^2(X_{ij}, W_{ij}) \quad (1)$$

$$+ \underbrace{N^{-1} \sum_j \left[ \sum_i \sum_{i' \neq i} (2W_{ij} - 1)(2W_{i'j} - 1) \left(1 + \frac{K_M(ij)}{M}\right) \left(1 + \frac{K_M(i'j)}{M}\right) \sigma_\eta^2 \right]}_{\text{Clustering correction}},$$

where  $\sigma_\varepsilon^2(X_{ij}, W_{ij}) = \sigma_\eta^2 + \sigma_\omega^2(X_{ij}, W_{ij})$ . The first term represents the conditional variance if we ignore the impact of clustering. The second term is the contribution of error clustering to the conditional variance. From (1) we see that the clustering correction is largest when all units in a given cluster have the same treatment status (when  $W_{ij} = W_{i'j}$  for all  $i$  and  $i'$  in cluster  $j$ ) so that all of the terms in the correction are positive. This parallels the linear OLS case where clustering matters more when covariates are more highly correlated within clusters (Greenwald (1983) and Moulton (1986)).

The marginal variance of the matching estimator also depends on  $V^{\tau(X)} = \text{Var}[\sqrt{N}\overline{\tau}(\mathbf{X})]$ . If the  $X_{ij}$  and hence the  $\tau(X_{ij})$  were drawn independently within each cluster, then we would have  $V^{\tau(X)} = E[(\tau(X_{ij}) - \tau)^2]$ . However, under our assumption that the  $X_{ij}$  are correlated within cluster, the identity  $\sqrt{N}(\overline{\tau}(\mathbf{X}) - \tau) = N^{-1/2} \sum_j \sum_i (\tau(X_{ij}) - \tau)$  implies that  $V^{\tau(X)} = N^{-1} J \cdot E[(\sum_i (\tau(X_{ij}) - \tau))^2]$  where the expectation is taken across clusters.<sup>2</sup>

### 3 Cluster-Robust Variance Estimators

#### 3.1 Why Standard Variance Estimation Methods Fail

Before outlining our estimation approach, we first explain why traditional techniques for estimating cluster-robust variances, such as those from the literature on the Generalized

---

<sup>2</sup>Although we do not explore such an extension here for simplicity, our framework could easily be extended to allow for multi-way clustering as in Cameron, Gelbach, and Miller (2010).

Method of Moments (GMM) (see, e.g., Bhattacharya (2005) and Wooldridge (2006)), are not applicable in the present setting. GMM is based on the assumption that we have a population moment condition such that  $E[g(z_i, \theta)] = 0 \Leftrightarrow \theta = \theta_0$  for some *known* function  $g(\mathbf{z}_i, \theta)$  that is specified *a priori* and does not depend on the sample under consideration. The GMM estimator is then defined using the sample analog of the population moment condition:  $N^{-1} \sum_i g(\mathbf{z}_i, \hat{\theta}) = 0$ . Traditional GMM asymptotics and cluster-robust variance estimators are based on the assumption that  $g(z, \theta)$  is continuously differentiable in  $\theta$ . The literature has extended these results to settings where  $E[g(z, \theta)]$  is continuously differentiable even though  $g(z, \theta)$  may not be. In these cases, the estimator will typically have an asymptotically linear representation and existing techniques can be used to consistently estimate a cluster-robust variance.

Can we use these methods to estimate a cluster-robust variance for the matching estimator? For instance, one might reason that  $\hat{\tau}_M$  satisfies the “sample moment condition”  $N^{-1} \sum_{i,j} [(2W_{i,j} - 1)(1 + M^{-1}K_M(ij|\mathbf{X}, \mathbf{W}))Y_{i,j} - \hat{\tau}_M] = 0$ . Here we write  $K_M(ij|\mathbf{X}, \mathbf{W})$  to emphasize that  $K_M(ij)$  depends on *all* the covariates and treatment assignments, not just those for unit  $ij$ . As we add observations  $K_M(ij|\mathbf{X}, \mathbf{W})$  can rise or fall discretely, so the  $K_M(ij|\mathbf{X}, \mathbf{W})$  and hence  $\hat{\tau}_M$  are highly non-smooth functionals of the data (Abadie and Imbens (2006)).<sup>3</sup> As a result, this condition is not based on the sample average of some *known* function, so matching estimators are not traditional GMM estimators.

However, one might still wonder whether existing techniques for asymptotically linear estimators might be used in this setting. Applied researchers often favor matching estimators with small, fixed  $M$  due to concerns about the conditional bias of estimators with large  $M$ . When the number of matches  $M$  is fixed, there is no evidence that matching estimators become asymptotically linear, which may explain the failure of standard bootstrapping methods for inference (Abadie and Imbens (2008)). Thus, in our setting, standard techniques may not be valid and hence should not be used.

### 3.2 Estimating the Conditional Variance

The main difficulty in estimating the conditional variance (1) of the matching estimator is that we do not have estimated residuals since we are not directly estimating the regression

---

<sup>3</sup>This should be contrasted with non-parametric Kernel regressions considered by Bhattacharya (2005), which are typically smooth functionals of the data.

functions  $\mu_0(x)$  and  $\mu_1(x)$ . We follow the approach used by Abadie and Imbens (2006 and 2008) who generate quasi-residuals by matching each individual to the most similar individual *with the same treatment status*. Specifically, they define  $\tilde{\varepsilon}_{ij} = Y_{ij} - Y_{g(ij)h(ij)}$  where  $(g(ij), h(ij)) = \arg \min_{g,h|W_{gh}=W_{ij}} \|X_{ij} - X_{gh}\|$ . Note that this is a different matching from that used to compute  $\hat{\tau}_M$  above. Abadie and Imbens then show that these quasi-residuals can be used to compute a heteroskedasticity-robust variance estimator in the absence of clustering.

The problem is more difficult in our setting because the presence of clustering means that we also need a consistent estimate of  $\sigma_\eta^2$ . Furthermore, following the clustering literature, we seek an estimator that only requires  $J \rightarrow \infty$  for consistency. Therefore, we consider matching *across* clusters. Specifically, let  $l(ij)$  and  $k(ij)$  index the individual and cluster, respectively, of the closest match to  $ij$  with the same treatment status in a different cluster:  $(l(ij), k(ij)) = \arg \min_{l,k \neq j|W_{lk}=W_{ij}} \|X_{ij} - X_{lk}\|$ . Define the quasi-residuals

$$\hat{\varepsilon}_{ij} = Y_{ij} - Y_{l(ij)k(ij)} = \underbrace{\mu_{W_{ij}}(X_{ij}) - \mu_{W_{ij}}(X_{l(ij)k(ij)})}_{\text{Matching discrepancy}} + \omega_{ij} - \omega_{l(ij)k(ij)} + \eta_j - \eta_{k(ij)}. \quad (2)$$

The matching discrepancy,  $\mu_{W_{ij}}(X_{ij}) - \mu_{W_{ij}}(X_{l(ij)k(ij)})$ , will vanish as the number of potential matches grows. Thus, we can ignore these terms as  $J \rightarrow \infty$ .<sup>4</sup>

Expanding a within-cluster cross-product of these quasi-residuals,  $\hat{\varepsilon}_{ij}\hat{\varepsilon}_{i'j}$ , and ignoring the matching discrepancy terms yields

$$\begin{aligned} \hat{\varepsilon}_{ij}\hat{\varepsilon}_{i'j} &= \eta_j^2 + \eta_j (\omega_{ij} - \omega_{l(ij)k(ij)} - \eta_{k(ij)} + \omega_{i'j} - \omega_{l(i'j)k(i'j)} - \eta_{k(i'j)}) \\ &\quad + (\omega_{ij} - \omega_{l(ij)k(ij)} - \eta_{k(ij)}) (\omega_{i'j} - \omega_{l(i'j)k(i'j)} - \eta_{k(i'j)}). \end{aligned}$$

Since we match across clusters and the  $\eta$ s are independent of the  $\omega$ s, it follows that

$$E[\hat{\varepsilon}_{ij}\hat{\varepsilon}_{i'j}|X_{ij}, X_{i'j}] = \sigma_\eta^2 + E[\eta_{k(ij)}\eta_{k(i'j)}] + E[\omega_{l(ij)k(ij)}\omega_{l(i'j)k(i'j)}] + E[\omega_{ij}\omega_{i'j}].$$

If  $i = i'$ , we have  $E[\hat{\varepsilon}_{ij}\hat{\varepsilon}_{i'j}|X_{ij}] = 2(\sigma_\eta^2 + \sigma_\omega^2(X_{ij}, W_{ij})) = 2\sigma_\varepsilon^2(X_{ij}, W_{ij})$ . For  $i \neq i'$ ,

---

<sup>4</sup>Note that if we were to allow matches within clusters (i.e. if  $k(ij) = j$ ), the  $\eta$  terms would drop out leaving  $\hat{\varepsilon}_{ij} = \mu_{W_{ij}}(X_{ij}) - \mu_{W_{ij}}(X_{l(ij)k(ij)}) + \omega_{ij} - \omega_{l(ij)k(ij)}$ . Although such matches would generally vanish and could be ignored as  $J \rightarrow \infty$ , they would impart a downward bias on the variance estimator in small samples. As a result, we would want to keep track of and correct for the occurrence of within-cluster matching, which would unnecessarily complicate our methodology.

$E [\eta_{k(ij)}\eta_{k(i'j)}|X_{ij}, X_{i'j}] = 0$  if  $i$  and  $i'$  are matched to units in distinct clusters (i.e.  $k(ij) \neq k(i'j)$ ); by contrast,  $E [\eta_{k(ij)}\eta_{k(i'j)}|X_{ij}, X_{i'j}] = \sigma_\eta^2$  if  $k(ij) = k(i'j)$ . Similarly,  $E [\omega_{l(ij)k(ij)}\omega_{l(i'j)k(i'j)}] = 0$  unless  $i$  and  $i'$  are matched to the exact same unit in which case  $E [\omega_{l(ij)k(ij)}\omega_{l(i'j)k(i'j)}|X_{ij}, X_{i'j}] = \sigma_\omega^2 (X_{l(ij)k(ij)}, W_{l(ij)k(ij)})$ . Thus, in the absence of these ‘‘duplicative matchings,’’ the cross-product  $\widehat{\varepsilon}_{ij}\widehat{\varepsilon}_{i'j}$  for  $i \neq i'$  is an unbiased estimator for  $\sigma_\eta^2$ .

We can correct for duplicative matchings by defining

$$\widehat{\sigma}^2(X_{ij}, X_{i'j}) = \begin{cases} \widehat{\varepsilon}_{ij}\widehat{\varepsilon}_{i'j} & \text{if } k(ij) \neq k(i'j) \\ \widehat{\varepsilon}_{ij}\widehat{\varepsilon}_{i'j}/2 & \text{if } k(ij) = k(i'j) \text{ and if } l(ij) \neq l(i'j) \\ \widehat{\varepsilon}_{ij}\widehat{\varepsilon}_{i'j} - (\widehat{\varepsilon}_{i'j})^2/2 & \text{if } k(ij) = k(i'j) \text{ and if } l(ij) = l(i'j) \end{cases} \quad (3)$$

If we ignore the matching discrepancy (i.e. assume that both  $ij$  and  $i'j$  are perfectly matched), it follows that

$$E [\widehat{\sigma}^2(X_{ij}, X_{i'j}) | X_{ij}, X_{i'j}] = \begin{cases} \sigma_\varepsilon^2(X_{ij}, W_{ij}) & \text{if } i = i' \\ \sigma_\eta^2 & \text{if } i \neq i' \end{cases},$$

so we sometimes write  $\widehat{\sigma}^2(X_{ij}, X_{ij}) = (\widehat{\varepsilon}_{ij})^2/2 = \widehat{\sigma}_\varepsilon^2(X_{ij}, W_{ij})$ .

Let  $P = \lim_{J \rightarrow \infty} J^{-1} \sum_j [(I_j(I_j - 1))^{-1} \sum_i \sum_{i' \neq i} 1 \{k(ij) = k(i'j)\}]$  be the probability of duplicative matchings as  $J \rightarrow \infty$ . It is worth noting that under many sampling arrangements  $P = 0$ . In such circumstances, correcting for duplicative matchings is unnecessary asymptotically. However, the probability of duplicative matchings need not vanish as  $J \rightarrow \infty$ .<sup>5</sup> By defining  $\widehat{\sigma}^2(X_{ij}, X_{i'j})$  as we have above, we ensure that our estimator is robust for any limiting probability of duplicative matchings.

Formally, we need the following assumptions:

**Assumption 1** (*Unconfoundedness*)  $W$  is independent of  $(Y(1), Y(0))$  conditional on  $X$ .  
*(Overlap)*  $0 < \Pr(W = 1|X) < 1$ .

**Assumption 2** *The  $X_{ij}$  are chosen from some bounded set  $\mathbb{X} \subset \mathbb{R}^m$ , and there is an upper bound  $\bar{I}$  on cluster size.*

---

<sup>5</sup>For instance, suppose that there are a finite number of cluster types each associated with a non-degenerate distribution of continuous covariates. If clusters are sampled *iid* from this set of types, then as  $J \rightarrow \infty$  there will be many clusters of each type and the probability of duplicative matchings will vanish (i.e.  $P = 0$ ).

On the other hand, suppose that there is a single continuous covariate that is constant within clusters. Ignoring ties, all units in a cluster  $j$  will be matched to units in a single cluster  $k$  (i.e.  $P = 1$ ).

**Assumption 3** *The conditional expectation and conditional variance functions are Lipschitz on  $\mathbb{X}$ :  $|\mu_W(X) - \mu_W(X')| \leq C_{\mu,W} \|X - X'\|$  and  $|\sigma_\omega^2(X, W) - \sigma_\omega^2(X', W)| \leq C_{\sigma,W} \|X - X'\|$  for  $W \in \{0, 1\}$ .*

**Assumption 4**  *$E[\eta^4]$  and  $E[\omega^4]$  are bounded.*

Assumption 1 is needed to ensure that  $\hat{\tau}_M \xrightarrow{p} \tau$ . As shown in the Appendix, Assumptions 2 and 3 ensure that the average matching discrepancy vanishes as the number of clusters increases. Assumption 4 ensures that the variances of  $\eta^2$  and  $\omega^2$  are defined.

**Proposition 1** *Suppose assumptions 1 through 5 hold and let*

$$\hat{V}^E = N^{-1} \sum_j \left[ \sum_i \sum_{i'} (2W_{ij} - 1)(2W_{i'j} - 1) \left(1 + \frac{K_M(ij)}{M}\right) \left(1 + \frac{K_M(i'j)}{M}\right) \hat{\sigma}^2(X_{ij}, X_{i'j}) \right] \quad (4)$$

*Then, holding fixed cluster sizes, as  $J \rightarrow \infty$ , we have  $\hat{V}^E \xrightarrow{p} V^E$ .*

While both the clustering correction term in  $V^E$  in (1) and its estimate  $\hat{V}^E$  in (4) are similar in form to the standard case (i.e., a weighted average of the cross-product of “residuals”), it is worth emphasizing that the construction of these “residuals” is quite distinct. As a result, the consistency proof underlying Proposition 1 is quite different from the proof in the standard case. However, we believe it is a strength of our approach that similar formulae are shown to apply in this non-standard setting.

### 3.3 Estimating the Marginal Variance

In this section we show how to estimate  $V^{\tau(X)} = N^{-1} J \cdot E[(\sum_i (\tau(X_{ij}) - \tau))^2]$ . Obviously, if we knew the  $\tau(X_{ij})$ , an estimate of  $V^{\tau(X)}$  could be based on  $N^{-1} \sum_j [\sum_i (\tau(X_{ij}) - \overline{\tau(\mathbf{X})})]^2$ . For the matching estimator, the imputed outcome  $\hat{Y}_{ij}(1) - \hat{Y}_{ij}(0)$  can serve as an estimator of  $\tau(X_{ij})$ . As  $J \rightarrow \infty$ , the matching discrepancy vanishes and we have

$$\begin{aligned} E[(\sum_i (\hat{Y}_{ij}(1) - \hat{Y}_{ij}(0) - \tau))^2] &\approx E[(\sum_i (\tau(X_{ij}) - \tau))^2] + E[(\sum_i (2W_{ij} - 1) \varepsilon_{ij})^2] \\ &- 2E[(\sum_i (2W_{ij} - 1) \varepsilon_{ij})(M^{-1} \sum_{i'} \sum_{st \in \mathcal{J}_M(i'j)} (2W_{i'j} - 1) \varepsilon_{st})] + E[(M^{-1} \sum_i \sum_{st \in \mathcal{J}_M(ij)} (2W_{ij} - 1) \varepsilon_{st})^2]. \end{aligned}$$

In the presence of clustering, this expectation is somewhat more complicated than the case considered in Abadie and Imbens (2006). First and most importantly, the  $\varepsilon_{ij}$  within a

given cluster will be correlated due to the shared component,  $\eta_j$ . In addition, the issue of duplicative matchings we saw above also arises in estimating  $V^{\tau(X)}$ .<sup>6</sup> Specifically, the  $\varepsilon_{ij}$  will be correlated with the  $\varepsilon_{st}$  if units in cluster  $j$  are used to impute missing outcomes for other units in cluster  $j$  (i.e. if any element of  $\mathcal{J}_M(i'j)$  is in cluster  $j$  for some  $i'$ ). Moreover, the  $\varepsilon_{st}$  will be correlated with each other if multiple units in some different cluster  $j' \neq j$  are used to impute missing outcomes for units in cluster  $j$ .

If  $P = 0$ , the last two sources of correlation will vanish asymptotically, but the first will always be present. Below we present a simple estimator of  $V^{\tau(X)}$  that is consistent in this case. However, as stated above, for certain empirical applications it will be important to have an estimator that is valid even if the probability of duplicative matchings does not vanish as  $J \rightarrow \infty$ . This estimator, which contains additional terms to correct for duplicative matchings, is given by equation (8) in the Appendix.

**Proposition 2** *Suppose assumptions 1 through 4 hold and that  $P = 0$ , and let*

$$\begin{aligned} \hat{V}^{\tau(X)} = & N^{-1} \sum_j \left[ \sum_i \left( \hat{Y}_{ij}(1) - \hat{Y}_{ij}(0) - \hat{\tau}_M \right) \right]^2 - N^{-1} \sum_{i,j} \frac{K_M(ij)}{M^2} \hat{\sigma}_\varepsilon^2(X_{ij}, W_{ij}) \quad (5) \\ & - N^{-1} \sum_j \left[ \sum_i \sum_{i'} (2W_{ij} - 1)(2W_{i'j} - 1) \hat{\sigma}^2(X_{ij}, X_{i'j}) \right]. \end{aligned}$$

*Then, holding fixed cluster sizes, as  $J \rightarrow \infty$ ,  $\hat{V}^{\tau(X)} \xrightarrow{p} V^{\tau(X)}$ . An estimate of the marginal variance can then be computed as  $\hat{V} = \hat{V}^{\tau(X)} + \hat{V}^E$ .*

Combining the results in Propositions 1 and 2, we have

$$\begin{aligned} \hat{V} = & N^{-1} \sum_{i,j} \left( \hat{Y}_{ij}(1) - \hat{Y}_{ij}(0) - \hat{\tau}_M \right)^2 + N^{-1} \sum_{i,j} \left( \left( \frac{K_M(ij)}{M} \right)^2 + \frac{2M-1}{M} \frac{K_M(ij)}{M} \right) \hat{\sigma}_\varepsilon^2(X_{ij}, W_{ij}) \\ & + N^{-1} \sum_{i,j} \sum_{i' \neq i} \left( \hat{Y}_{ij}(1) - \hat{Y}_{ij}(0) - \hat{\tau}_M \right) \left( \hat{Y}_{i'j}(1) - \hat{Y}_{i'j}(0) - \hat{\tau}_M \right) \quad (6) \\ & + N^{-1} \sum_{i,j} \sum_{i' \neq i} (2W_{ij} - 1)(2W_{i'j} - 1) \left( \frac{K_M(ij) + K_M(i'j)}{M} + \frac{K_M(ij)K_M(i'j)}{M^2} \right) \hat{\sigma}^2(X_{ij}, X_{i'j}) \end{aligned}$$

The first two terms are the estimator of  $V^E + V^{\tau(X)}$  given in Abadie and Imbens (2006) which is valid in the absence of clustering. The second pair of terms are the combined clustering correction which is valid in the case where  $P = 0$ .

<sup>6</sup>Note that here we are referring to the probability of duplicative matchings for the matching used in the matching estimator (indexed  $st$ ), not the matching used to compute the quasi-residuals (indexed  $lk$ ). However, the asymptotic probability of duplicative matchings is a function of the distribution of covariates, so the probability of duplicative matchings will generally either be zero or non-zero for both matchings.

## 4 A Short Monte Carlo Study

To get a sense of the finite sample properties of our variance estimator, we examine the confidence intervals it generates. We assume there is a single covariate,  $X_{ij} \stackrel{iid}{\sim} N(0, 5)$ , and that  $\mu_0(x) = 0$  and  $\mu_1(x) = x$ . It follows that  $\tau(x) = x$ , and the observed outcome is  $Y_{ij} = W_{ij}X_{ij} + \varepsilon_{ij}$ . We assume that half the clusters are treated and that  $\sigma_\eta^2 = 1.5$  and  $\sigma_\omega^2 = 1.5$ . For each replication, we draw a new set of covariates  $X_{ij}$ , as well as error components  $\eta_j$  and  $\omega_{ij}$ . We then estimate the PATE using the simple matching estimator,  $\hat{\tau}_1$  with  $M = 1$  matches to impute unobserved outcomes. Clustering-corrected 95% confidence intervals for the estimator are computed as  $(\hat{\tau}_1 - 1.96(\hat{V}/N)^{1/2}, \hat{\tau}_1 + 1.96(\hat{V}/N)^{1/2})$ , where  $\hat{V} = \hat{V}^{\tau(X)} + \hat{V}^E$ ,  $\hat{V}^E$  is given by (4), and  $\hat{V}^{\tau(X)}$  is given by equation (8) in the Appendix. We carry out  $R = 1000$  replications and report the fraction of replications where  $E[\tau(x)] = 0$  is covered by the resulting confidence intervals.

Our marginal variance estimator has very good finite sample performance. While it shows slight under-coverage in very small samples, its coverage reaches 95% quickly as  $J$  increases. Furthermore, while the coverage of an estimator that did not correct for clustering would decline to approximately 0.70 for  $I = 10$  and 0.40 for  $I = 50$ , the performance of our estimator is nearly constant as a function of  $I$ . Similar coverage probabilities obtain if we allow the  $X_{ij}$  to be correlated within clusters.

Coverage prob. of 95% CI				
		J		
		10	20	50
I	2	0.90	0.93	0.94
	10	0.91	0.92	0.95
	50	0.92	0.94	0.95

## 5 Other Non-Parametric Treatment Effect Estimators

First, as discussed in the Appendix, it is straightforward to use our approach to compute cluster-robust variances for matching estimators of the average treatment effect for the treated (ATT). Second, while this note has focused on the matching estimators, the methodology described above can be extended to a broader class of non-parametric treat-

ment effect estimators. A number of estimators for average treatment effects, including propensity score based estimators, can be written as

$$\hat{\tau} = \sum_{W_{ij}=1} \gamma_{ij}(\mathbf{X}, \mathbf{W}) Y_{ij} - \sum_{W_{ij}=0} \gamma_{ij}(\mathbf{X}, \mathbf{W}) Y_{ij} = \sum_{i,j} (2W_{ij} - 1) \gamma_{ij}(\mathbf{X}, \mathbf{W}) Y_{ij} \quad (7)$$

where  $\gamma_{ij}(\mathbf{X}, \mathbf{W})$  is a set of data-dependent weights. For estimators of the form (7), the conditional variance can be estimated using Eq. (4) and replacing  $N^{-1}(1 + K_M(ij)/M)$  with  $\gamma_{ij}$ . Similarly, for estimators that impute treatment effects  $\hat{Y}_{ij}(1) - \hat{Y}_{ij}(0)$  for each unit, we can estimate  $V^{\tau(X)}$  using an expression analogous to Eq. (5) above.

## 6 Conclusion

In this note, we develop a methodology to estimate the conditional and marginal variances of the matching estimator in the presence of clustering. Our cluster-robust variance estimators are consistent as the number of clusters grows large, holding the size of clusters fixed. Furthermore, our methodology can easily be extended to a broader class of non-parametric treatment effect estimators.

## References

- [1] Abadie, A. and G. Imbens (2006) “Large Sample Properties of Matching Estimators,” *Econometrica* 74, 235-276.
- [2] Abadie, A. and G. Imbens (2008) “On the Failure of the Bootstrap for Matching Estimators,” *Econometrica* 76, 1537-1557.
- [3] Abadie, A. and G. Imbens (2010) “Estimation of the Conditional Variance in Paired Experiments,” *Annales d’Economie et de Statistique*, forthcoming.
- [4] Bhattacharya (2005) “Asymptotic Inference from Multi-stage Samples,” *Journal of Econometrics*, May 2005, pp. 145-171.
- [5] Cameron C., J. Gelbach, and D. Miller (2010) “Robust Inference with Multi-way Clustering”, *Journal of Business and Economic Statistics*, forthcoming.
- [6] Fryer, R. (2010) “Financial Incentives and Student Achievement: Evidence from a Randomized Trial,” unpublished paper, Harvard University.
- [7] Greenwald, B. (1983) “A General Analysis of Bias in the Estimated Standard Errors of Least Squares Coefficients,” *Journal of Econometrics*, 323-338.
- [8] Heckman, J., H. Ichimura, and P. Todd (1997), “Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme,” *Review of Economic Studies*, 64, 605-654.
- [9] Hirano, K., G. Imbens, and G. Ridder (2003) “Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score,” *Econometrica* 71, 1161-1189.
- [10] Hotz, V., G. Imbens, and J. Klerman (2006) “Evaluating the Differential Effects of Alternative Welfare-to-Work Training Components: A Re-Analysis of the California GAIN Program,” *Journal of Labor Economics* 24(3), 521-566.
- [11] Kloek, T. (1981) “OLS Estimation in a Model where a Microvariable is Explained by Aggregates and Contemporaneous Disturbances are Equicorrelated,” *Econometrica* 49, 205-207.
- [12] Imbens, G. (2004) “Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review,” *Review of Economics and Statistics*, 4-29.
- [13] Moulton, B. (1986) “Random Group Effects and the Precision of Regression Estimates,” *Journal of Econometrics*, 385-397.
- [14] Moulton, B. (1990) “An Illustration of a Pitfalls in Estimating the Effects of Aggregate Variables on Micro Units,” *Review of Economics and Statistics*, 334-338.
- [15] Wooldridge, J. (2002) *Econometric Analysis of Cross Section and Panel Data*, The MIT Press, Cambridge, MA.
- [16] Wooldridge, J. (2006) “Cluster-sample Methods in Applied Econometrics: An Extended Analysis,” unpublished paper, Michigan State University

## A Appendix: Extensions

### A.1 A Cluster-Robust Variance for Matching Estimators of the Treatment Effect for the Treated

Recall that the population average treatment effect for the treated (PATT) is  $\tau^t = E[\tau(X) | W = 1]$  and, conditional on  $\mathbf{X}$  and  $\mathbf{W}$ , the sample average treatment effect for the treated (SATT) is  $\overline{\tau^t(\mathbf{X})} = N_1^{-1} \sum_{ij:W_{ij}=1} \tau(X_{ij})$  where  $N_1$  is the number of treated units in the sample. The simple matching estimator for the ATT is

$$\widehat{\tau}_M^t = N_1^{-1} \sum_{ij:W_{ij}=1} \left( Y_{ij}(1) - \widehat{Y}_{ij}(0) \right) = N_1^{-1} \sum_{ij} \left( W_{ij} - (1 - W_{ij}) \frac{K_M(ij)}{M} \right) Y_{ij}.$$

The normalized conditional variance of  $\widehat{\tau}_M^t$ ,  $V^{E,t} = \text{Var}[\sqrt{N_1} \widehat{\tau}_M^t | \mathbf{X}, \mathbf{W}]$ , in the presence of residual clustering is

$$\begin{aligned} V^{E,t} &= N_1^{-1} \sum_{ij} \left( W_{ij} - (1 - W_{ij}) \frac{K_M(ij)}{M} \right)^2 \sigma_\varepsilon^2(X_{ij}, W_{ij}) \\ &\quad + N_1^{-1} \sum_{i,j} \sum_{i' \neq i} \left( W_{ij} - (1 - W_{ij}) \frac{K_M(ij)}{M} \right) \left( W_{i'j} - (1 - W_{i'j}) \frac{K_M(i'j)}{M} \right) \sigma_\eta^2 \end{aligned}$$

and the normalized variance of the SATT (the conditional mean) is  $V^{\tau(X),t} = N_1^{-1} J \cdot E[(\sum_{i:W_{ij}=1} (\tau(X_{ij}) - \tau^t))^2]$ . The conditional variance can be estimated using

$$\widehat{V}^{E,t} = N_1^{-1} \sum_j \left[ \sum_i \sum_{i'} \left( W_{ij} - (1 - W_{ij}) \frac{K_M(ij)}{M} \right) \left( W_{i'j} - (1 - W_{i'j}) \frac{K_M(i'j)}{M} \right) \widehat{\sigma}^2(X_{ij}, X_{i'j}) \right]$$

and, assuming  $P = 0$ , the marginal variance can be estimated using  $\widehat{V}^{E,t} + \widehat{V}^{\tau(X),t}$  where

$$\begin{aligned} \widehat{V}^{\tau(X),t} &= N_1^{-1} \sum_j \left[ \sum_i W_{ij} \left( Y_{ij}(1) - \widehat{Y}_{ij}(0) - \widehat{\tau}_M^t \right) \right]^2 - N_1^{-1} \sum_{ij} (1 - W_{ij}) \frac{K_M(ij)}{M^2} \widehat{\sigma}_\varepsilon^2(X_{ij}, W_{ij}) \\ &\quad - N_1^{-1} \sum_j \sum_i \sum_{i'} W_{ij} W_{i'j} \widehat{\sigma}^2(X_{ij}, X_{i'j}). \end{aligned}$$

Analogous expressions can be derived for estimators of the average treatment for controls.

### A.2 Computer Code

The program `nnmatch_wcluster.ado` computes two estimates of  $V$  for the simple matching estimator. The first is the estimator given above which fully adjusts for the effects of duplicative matching. Specifically, the estimator of the normalized conditional variance is given by (4) and the estimator of the normalized marginal variance is  $\widehat{V} = \widehat{V}^E + \widehat{V}^{\tau(X)}$  where  $\widehat{V}^{\tau(X)}$  is given by (8) below.

The second estimator ignores all consequences of duplicative matches. Specifically, letting

$$\widetilde{\sigma}^2(X_{ij}, X_{i'j}) = \begin{cases} \widehat{\varepsilon}_{ij} \widehat{\varepsilon}_{i'j} & \text{if } i \neq i' \\ (\widehat{\varepsilon}_{ij})^2 / 2 & i = i' \end{cases},$$

the estimator of the marginal variance is

$$\hat{V}_{dp}^E = N^{-1} \sum_j \left[ \sum_i \sum_{i'} (2W_{ij} - 1) (2W_{i'j} - 1) \left( 1 + \frac{K_M(ij)}{M} \right) \left( 1 + \frac{K_M(i'j)}{M} \right) \tilde{\sigma}^2(X_{ij}, X_{i'j}) \right],$$

and the estimator of the normalized marginal variance is  $\hat{V}_{nd} = \hat{V}_{nd}^E + \hat{V}_{nd}^{\tau(X)}$ , where

$$\begin{aligned} \hat{V}_{dp}^{\tau(X)} &= N^{-1} \sum_j \left[ \sum_i \left( \hat{Y}_{ij}(1) - \hat{Y}_{ij}(0) - \hat{\tau}_M \right) \right]^2 - N^{-1} \sum_{i,j} \frac{K_M(ij)}{M^2} \tilde{\sigma}_\varepsilon^2(X_{ij}, W_{ij}) \\ &\quad - N^{-1} \sum_j \left[ \sum_i \sum_{i'} (2W_{ij} - 1) (2W_{i'j} - 1) \tilde{\sigma}^2(X_{ij}, X_{i'j}) \right]. \end{aligned}$$

The program also reports the sample frequency of within-cluster duplicative matches,  $\hat{P} = J^{-1} \sum_j [(I_j(I_j - 1))^{-1} \sum_i \sum_{i' \neq i} 1 \{k(ij) = k(i'j)\}]$ , which drives the differences between these two estimators.

The program can be used to compute clustering corrections for the average treatment effect (SATE or PATE), the average treatment effect for the treated (SATF or PATF), and the average treatment effect for the controls (SATC or PATC).

## B Appendix: Proofs

### B.1 Proof of Proposition 1

The consistency proof closely follows those given in Abadie and Imbens (2006 and 2008). Defining  $\gamma_{ij} = (2W_{ij} - 1) (1 + M^{-1}K_M(ij))$  to simplify the expressions, recall that our estimator is

$$\hat{V}^E = N^{-1} \sum_j \left[ \sum_i \sum_{i'} \gamma_{ij} \gamma_{i'j} \hat{\sigma}^2(X_{ij}, X_{i'j}) \right]$$

where  $\hat{\sigma}^2(X_{ij}, X_{i'j})$  is defined in (3). Also, recall that

$$V^E = N^{-1} \sum_j \left[ \sum_i \sum_{i'} \gamma_{ij} \gamma_{i'j} \sigma^2(X_{ij}, X_{i'j}) \right]$$

where

$$\sigma^2(X_{ij}, X_{i'j}) = \begin{cases} \sigma_\varepsilon^2(X_{ij}, W_{ij}) & \text{if } i = i' \\ \sigma_\eta^2 & \text{if } i \neq i' \end{cases}.$$

**Proof sketch:** The strategy of the proof is first to show that the conditional expectation of  $\hat{V}^E$ ,  $E[\hat{V}^E | \mathbf{X}, \mathbf{W}]$ , converges to  $V^E$  as  $J$  increases and the matching discrepancy vanishes. The second step is to show that  $Var[\hat{V}^E - E[\hat{V}^E | \mathbf{X}, \mathbf{W}] | \mathbf{X}, \mathbf{W}] \rightarrow 0$  which implies convergence in mean square and, thus, convergence in probability.

First, we require the following two lemmas.

**Lemma 1** *If the  $X_{ij}$  are chosen from some bounded set  $\mathbb{X} \subset \mathbb{R}^k$  and there is an upper bound  $\bar{I}$  on group size, then (1)  $N^{-1} \sum_{i,j} \|X_{ij} - X_{l(ij)k(ij)}\| \rightarrow 0$ , (2)  $N^{-1} \sum_{i,j} \|X_{ij} - X_{l(ij)k(ij)}\|^2 \rightarrow 0$ , and (3)  $N^{-1} \sum_j \left( \sum_i \|X_{ij} - X_{l(ij)k(ij)}\| \right)^2 \rightarrow 0$ .*

**Proof.** The proof is similar to that in Abadie and Imbens (2008). Note that

$$\begin{aligned} N^{-1} \sum_{i,j} \|X_{ij} - X_{l(ij)k(ij)}\|^2 &\leq \bar{I} \cdot J^{-1} \sum_j \max_i \|X_{ij} - X_{l(ij)k(ij)}\|^2 \\ N^{-1} \sum_j (\sum_i \|X_{ij} - X_{l(ij)k(ij)}\|)^2 &\leq \bar{I}^2 \cdot J^{-1} \sum_j \max_i \|X_{ij} - X_{l(ij)k(ij)}\|^2 \end{aligned}$$

Furthermore, we have

$$N^{-1} \sum_{i,j} \|X_{ij} - X_{l(ij)k(ij)}\| \leq \bar{I} \cdot J^{-1} \sum_j \max_i \|X_{ij} - X_{l(ij)k(ij)}\|$$

Thus, it suffices to show that  $J^{-1} \sum_j \max_i \|X_{ij} - X_{l(ij)k(ij)}\| \rightarrow 0$ .<sup>7</sup> Embed  $\mathbb{X}$  in a hypersphere radius with  $\text{diam}(\mathbb{X})/2$ . WLOG, assume that  $\text{diam}(\mathbb{X}) = 2$ . Suppose there are  $H$  clusters whose maximal matching discrepancy is greater than  $2\varepsilon$ . Construct an open ball of radius  $\varepsilon$  around each such point. These  $H$  balls do not intersect. The volume of each such ball is equal to  $(\varepsilon^k/k) S_k$  where  $S_k$  is the surface area of a unit hypersphere in  $k$  dimensions. All of these balls are embedded in hypersphere with radius  $(1 + \varepsilon)$  so that  $H (\varepsilon^k/k) S_k < ((1 + \varepsilon)^k/k) S_k$  which implies that  $H < ((1 + \varepsilon)/\varepsilon)^k$ . Therefore, we have

$$J^{-1} \sum_j \max_i \|X_{ij} - X_{l(ij)k(ij)}\| < 2(H/J) + 2\varepsilon J^{-1}(J - H) < 2(H/J + \varepsilon) < 2 \left( J^{-1} [(1 + \varepsilon)/\varepsilon]^k + \varepsilon \right).$$

Now consider  $\varepsilon = J^{-1/(1+k)}$ . As  $J \rightarrow \infty$  we have  $\varepsilon \rightarrow 0$  and

$$\frac{1}{J} \left( \frac{1 + \varepsilon}{\varepsilon} \right)^k = \frac{1}{J} \left( \frac{1 + J^{-1/(1+k)}}{J^{-1/(1+k)}} \right)^k = \frac{J^{k/(1+k)}}{J} (1 + J^{-1/(1+k)})^k = \frac{J^{k/(1+k)}}{J} O(1) = o(1).$$

As a result, we have  $J^{-1} \sum_j \max_i \|X_{ij} - X_{l(ij)k(ij)}\| \rightarrow 0$  which proves the lemma. ■

As in Abadie and Imbens (2006 and 2008), these calculations make use of the fact that the number of times a given observation can be used as a match for the purposes of variance estimation is bounded above.

**Lemma 2** *Suppose dimension of  $X$  is equal to  $k$ . The maximum number of times that an observation can be used as a match is bounded above by  $\bar{L}(k)$ , the kissing number in  $\mathbb{R}^k$ . The maximum number of times that an observation can be used as a match, given the requirement that all observations are matched across clusters, is bounded above by  $\bar{I} \cdot \bar{L}(k)$ . Since each individual can be matched at most  $\bar{I} \cdot \bar{L}(k)$  times, each cluster (which consists of at most  $\bar{I}$  units) can be matched at most  $\bar{I}^2 \cdot \bar{L}(k)$  times.*

**Proof.** Recall that  $\bar{L}(k)$  is maximum number of non-overlapping unit balls in  $\mathbb{R}^k$  than can touch a unit ball. If we do not require matches to be across clusters and instead only required that a unit be matched to another with the same treatment status, then  $\bar{L}(k)$  would be an upper bound on the number of times a unit can be matched. However, adding this requirement increases the bound by a factor of  $\bar{I}$ . To see the intuition, note that for  $k = 1$  we have  $\bar{L}(k) = 2$ . If we require matches to be across clusters, the maximum number of

<sup>7</sup>Since  $\max_i \|X_{ij} - X_{l(ij)k(ij)}\| \leq \text{diam}(\mathbb{X}) < \infty$ ,  $J^{-1} \sum_j \max_i \|X_{ij} - X_{l(ij)k(ij)}\| \rightarrow 0$  implies  $J^{-1} \sum_j \max_i \|X_{ij} - X_{l(ij)k(ij)}\|^2 \rightarrow 0$ .

times a units can be matched is  $2\bar{I}$  and is obtained when a unit is surrounded by  $\bar{I}$  units from one cluster on the left, by  $\bar{I}$  units from another cluster on the right, and is the nearest “across-cluster” neighbor of each of these  $2\bar{I}$  units.

**Step 1:** As discussed above, the first step is show that  $E[\hat{V}^E|\mathbf{X}, \mathbf{W}] - V^E \rightarrow 0$  as  $J \rightarrow \infty$ . Note that

$$\begin{aligned} E[\hat{V}^E|\mathbf{X}, \mathbf{W}] - V^E &= N^{-1} \sum_j \sum_i (\gamma_{ij})^2 E[\hat{\sigma}_\varepsilon^2(X_{ij}, W_{ij}) - \sigma_\varepsilon^2(X_{ij}, W_{ij})|\mathbf{X}, \mathbf{W}] \\ &\quad + N^{-1} \sum_j \left[ \sum_i \sum_{i' \neq i} \gamma_{ij} \gamma_{i'j} E[\hat{\sigma}_\eta^2(X_{ij}, X_{i'j}) - \sigma_\eta^2|\mathbf{X}, \mathbf{W}] \right] \end{aligned}$$

Consider the first term. By Assumption 3 we have  $|\mu_W(X) - \mu_W(X')| \leq C_{\mu,W} \|X - X'\|$  and  $|\sigma_\omega^2(X, W) - \sigma_\omega^2(X', W)| \leq C_{\sigma,W} \|X - X'\|$  for  $W \in \{0, 1\}$ . Let  $C = \max\{C_{\mu,0}, C_{\mu,1}, C_{\sigma,0}, C_{\sigma,1}\}$ . First note that

$$\begin{aligned} &E[\hat{\sigma}_\varepsilon^2(X_{ij}, W_{ij}) - \sigma_\varepsilon^2(X_{ij}, W_{ij})|\mathbf{X}, \mathbf{W}] \\ &= (\mu_{W_{ij}}(X_{ij}) - \mu_{W_{ij}}(X_{l(ij)k(ij)}))^2/2 + (\sigma_\omega^2(X_{l(ij)k(ij)}, W_{ij}) - \sigma_\omega^2(X_{ij}, W_{ij}))/2 \\ &\leq (C^2/2) \|X_{ij} - X_{l(ij)k(ij)}\|^2 + (C/2) \|X_{ij} - X_{l(ij)k(ij)}\|. \end{aligned}$$

For any  $i' \neq i$  we have

$$\begin{aligned} \hat{\varepsilon}_{ij} \hat{\varepsilon}_{i'j} &= (\mu_{W_{ij}}(X_{ij}) - \mu_{W_{ij}}(X_{l(ij)k(ij)}))(\mu_{W_{i'j}}(X_{i'j}) - \mu_{W_{i'j}}(X_{l(i'j)k(i'j)})) \\ &\quad + (\mu_{W_{ij}}(X_{ij}) - \mu_{W_{ij}}(X_{l(ij)k(ij)}))((\omega_{i'j} + \eta_j) - (\omega_{l(i'j)k(i'j)} + \eta_{k(i'j)})) \\ &\quad + (\mu_{W_{i'j}}(X_{i'j}) - \mu_{W_{i'j}}(X_{l(i'j)k(i'j)}))((\omega_{ij} + \eta_j) - (\omega_{l(ij)k(ij)} + \eta_{k(ij)})) \\ &\quad \eta_j^2 + \eta_j(\omega_{ij} - \omega_{l(ij)k(ij)} - \eta_{k(ij)} + \omega_{i'j} - \omega_{l(i'j)k(i'j)} - \eta_{k(i'j)}) \\ &\quad + (\omega_{ij} - \omega_{l(ij)k(ij)} - \eta_{k(ij)}) \times (\omega_{i'j} - \omega_{l(i'j)k(i'j)} - \eta_{k(i'j)}). \end{aligned}$$

If  $k(ij) \neq k(i'j)$ ,  $E[\hat{\sigma}_\eta^2(X_{ij}, X_{i'j}) - \sigma_\eta^2|\mathbf{X}, \mathbf{W}] = (\mu_{W_{ij}}(X_{ij}) - \mu_{W_{ij}}(X_{l(ij)k(ij)}))(\mu_{W_{i'j}}(X_{i'j}) - \mu_{W_{i'j}}(X_{l(i'j)k(i'j)}))$ . If  $k(ij) = k(i'j)$  and  $l(ij) \neq l(i'j)$ ,  $E[\hat{\sigma}_\eta^2(X_{ij}, X_{i'j}) - \sigma_\eta^2|\mathbf{X}, \mathbf{W}] = (\mu_{W_{ij}}(X_{ij}) - \mu_{W_{ij}}(X_{l(ij)k(ij)}))(\mu_{W_{i'j}}(X_{i'j}) - \mu_{W_{i'j}}(X_{l(i'j)k(i'j)}))/2$ . Finally, if  $k(ij) = k(i'j)$  and  $l(ij) = l(i'j)$ , we have  $E[\hat{\sigma}_\eta^2(X_{ij}, X_{i'j}) - \sigma_\eta^2|\mathbf{X}, \mathbf{W}] = (\mu_{W_{ij}}(X_{ij}) - \mu_{W_{ij}}(X_{l(ij)k(ij)}))(\mu_{W_{i'j}}(X_{i'j}) - \mu_{W_{i'j}}(X_{l(i'j)k(i'j)})) - (\mu_{W_{ij}}(X_{ij}) - \mu_{W_{ij}}(X_{l(ij)k(ij)}))^2 + (\sigma_\omega^2(X_{l(ij)k(ij)}, W_{ij}) - \sigma_\omega^2(X_{ij}, W_{ij}))/2$ . In any case, we have

$$|E[\hat{\sigma}_\eta^2(X_{ij}, X_{i'j}) - \sigma_\eta^2|\mathbf{X}, \mathbf{W}]| \leq C^2 \|X_{ij} - X_{l(ij)k(ij)}\| \|X_{i'j} - X_{l(i'j)k(i'j)}\| + (C/2)^2 \|X_{ij} - X_{l(ij)k(ij)}\|.$$

Since  $K_M(ij) \leq M \cdot \bar{L}(k)$ ,  $\gamma_{ij} \gamma_{i'j} \leq (1 + \bar{L}(k))^2$ . By Assumption 2 and the previous Lemma we have  $N^{-1} \sum_{i,j} \|X_{ij} - X_{l(ij)k(ij)}\| \rightarrow 0$  and  $N^{-1} \sum_j (\sum_i \|X_{ij} - X_{l(ij)k(ij)}\|)^2 \rightarrow 0$ , so we have

$$\begin{aligned} |E[\hat{V}^E|\mathbf{X}, \mathbf{W}] - V^E| &\leq (1 + \bar{L}(k))^2 C^2 \cdot N^{-1} \sum_j (\sum_i \|X_{ij} - X_{l(ij)k(ij)}\|)^2 \\ &\quad + (1 + \bar{L}(k))^2 (C/2) \cdot N^{-1} \sum_{i,j} \|X_{ij} - X_{l(ij)k(ij)}\| \rightarrow 0 \end{aligned}$$

as  $J \rightarrow \infty$ .

**Step 2:** The second step is to show that  $Var \left[ \hat{V}^E - E \left[ \hat{V}^E | \mathbf{X}, \mathbf{W} \right] | \mathbf{X}, \mathbf{W} \right] \rightarrow 0$  which implies convergence in mean square and, thus, convergence in probability. Note that

$$\begin{aligned} \hat{V}^E - E[\hat{V}^E | \mathbf{X}, \mathbf{W}] &= N^{-1} \sum_j \sum_i (\gamma_{ij})^2 (\hat{\sigma}_\varepsilon^2(X_{ij}, W_{ij}) - E[\sigma_\varepsilon^2(X_{ij}, W_{ij}) | \mathbf{X}, \mathbf{W}]) \\ &\quad + N^{-1} \sum_j \left[ \sum_i \sum_{i' \neq i} \gamma_{ij} \gamma_{i'j} (\hat{\sigma}^2(X_{ij}, X_{i'j}) - E[\hat{\sigma}^2(X_{ij}, X_{i'j}) | \mathbf{X}, \mathbf{W}]) \right] \end{aligned}$$

Consider the first term in  $\hat{V}^E - E[\hat{V}^E | \mathbf{X}, \mathbf{W}]$ , we have

$$\begin{aligned} &N^{-1} \sum_{i,j} (\gamma_{ij})^2 (\hat{\sigma}_\varepsilon^2(X_{ij}, W_{ij}) - E[\sigma_\varepsilon^2(X_{ij}, W_{ij}) | \mathbf{X}, \mathbf{W}]) \\ = &(2N)^{-1} \sum_{i,j} (\gamma_{ij})^2 [(\omega_{ij} + \eta_j)^2 - (\sigma_\omega^2(X_{ij}, W_{ij}) + \sigma_\eta^2)] \quad \text{(1)} \\ &+ (2N)^{-1} \sum_{i,j} (\gamma_{ij})^2 [(\omega_{l(ij)k(ij)} + \eta_{k(ij)})^2 - (\sigma_\omega^2(X_{l(ij)k(ij)}, W_{ij}) + \sigma_\eta^2)] \quad \text{(2)} \\ &- N^{-1} \sum_{i,j} (\gamma_{ij})^2 [(\omega_{ij} + \eta_j)(\omega_{l(ij)k(ij)} + \eta_{k(ij)})] \quad \text{(3)} \\ &+ N^{-1} \sum_{i,j} (\gamma_{ij})^2 [(\omega_{ij} + \eta_j)(\mu_{W_i}(X_{ij}) - \mu_{W_i}(X_{l(ij)k(ij)}))] \quad \text{(4)} \\ &- N^{-1} \sum_{i,j} (\gamma_{ij})^2 [(\omega_{l(ij)k(ij)} + \eta_{k(ij)})(\mu_{W_i}(X_{ij}) - \mu_{W_i}(X_{l(ij)k(ij)}))] \quad \text{(5)}. \end{aligned}$$

The conditional expectation of each term on the right-hand side is zero, so it suffices to show that the conditional variance of each of these five terms converges to zero. If all five of these variances vanishes as  $J \rightarrow \infty$ , then conditional variance of sum must also converge to zero (the Cauchy-Schwarz inequality guarantees that if all the variance terms vanish then so do all the covariance terms), so that this term converges in mean square and hence in probability. Since conditional expectation of each term is zero, the conditional variances are simply the expectations of each term squared. As in Abadie and Imbens (2006 and 2008), these calculations make use of the fact that the number of times a given observation can be used as a match for the purposes of variance estimation is bounded above.

As an example of the general proof technique, we show here that the conditional variance of term (2) in the previous expansion converges to zero. The other terms follows by similar arguments.

Since  $N = \sum_j I_j > J$ , we have

$$\begin{aligned} &E[(2N)^{-1} \sum_{i,j} (\gamma_{ij})^2 [\varepsilon_{l(ij)k(ij)}^2 - \sigma_\varepsilon^2(X_{l(ij)k(ij)}, W_{ij})]^2 | \mathbf{X}, \mathbf{W}] \\ = &(2N)^{-1} \sum_{i,j} (\gamma_{ij})^4 E[\varepsilon_{l(ij)k(ij)}^4 - \sigma_\varepsilon^4(X_{l(ij)k(ij)}, W_{ij}) | \mathbf{X}, \mathbf{W}] \\ &+ (2N)^{-1} \sum_{i,j} \sum_{i',j' \neq i,j} \gamma_{ij}^2 \gamma_{i'j'}^2 E[\varepsilon_{l(ij)k(ij)}^2 \varepsilon_{l(i'j')k(i'j')}^2 - \sigma_\varepsilon^2(X_{l(ij)k(ij)}, W_{ij}) \sigma_\varepsilon^2(X_{l(i'j')k(i'j')}, W_{i'j'}) | \mathbf{X}, \mathbf{W}] \\ \leq &(4J)^{-1} (1 + \bar{L}(k))^4 Var[\varepsilon^2] \cdot [1 + \bar{L}(k)] \rightarrow 0 \end{aligned}$$

as  $J \rightarrow \infty$ . We have used the fact that  $E[\varepsilon_{l(ij)k(ij)}^2 \varepsilon_{l(i'j')k(i'j')}^2 - \sigma_\varepsilon^2(X_{l(ij)k(ij)}, W_{ij}) \sigma_\varepsilon^2(X_{l(i'j')k(i'j')}, W_{i'j'})]$  is equal to 0 if  $ij$  and  $i'j'$  are matched to individuals in different clusters,  $Var[\varepsilon^2]$  if  $ij$  and  $i'j'$  are matched to the same individual in the same cluster, and  $Var[\eta^2]$  if  $ij$  and  $i'j'$  are matched to different individuals in the same cluster. Individual  $l(ij)k(ij)$  can be matched

at most  $\bar{I} \cdot \bar{L}(k)$  times. Therefore, for each  $ij$  we have

$$\begin{aligned} & (\gamma_{ij})^2 (\gamma_{i'j})^2 \sum_{i',j' \neq i,j} E[\varepsilon_{l(ij)k(ij)}^2 \varepsilon_{l(i'j')k(i'j')}^2 - \sigma_\varepsilon^2(X_{l(ij)k(ij)}, W_{ij}) \sigma_\varepsilon^2(X_{l(i'j')k(i'j')}, W_{i'j'})] | \mathbf{X}, \mathbf{W} \\ & \leq (1 + \bar{L}(k))^4 \cdot \bar{I} \cdot \bar{L}(k) \cdot \max\{Var[\varepsilon^2], Var[\eta^2]\} = (1 + \bar{L}(k))^4 \cdot \bar{I} \cdot \bar{L}(k) \cdot Var[\varepsilon^2] \end{aligned}$$

and the result follows.

Consider the second term in  $\hat{V}^E - E[\hat{V}^E | \mathbf{X}, \mathbf{W}]$ . To keep the argument simpler, we assume that it is always the case that  $k(ij) \neq k(i'j)$  for  $i' \neq i$ , so that

$$\begin{aligned} & N^{-1} \sum_{i,j} [\gamma_{ij} \gamma_{i'j} (\hat{\sigma}^2(X_{ij}, X_{i'j}) - E[\hat{\sigma}^2(X_{ij}, X_{i'j}) | \mathbf{X}, \mathbf{W}])] \\ = & N^{-1} \sum_{i,j} \left[ \sum_{i' \neq i} \gamma_{ij} \gamma_{i'j} (\eta_j^2 - \sigma_\eta^2) \right] \quad \text{(A)} \\ & + N^{-1} \sum_{i,j} \left[ \sum_{i' \neq i} \gamma_{ij} \gamma_{i'j} \eta_j \left( (\omega_{ij} - \omega_{l(ij)k(ij)} - \eta_{k(ij)}) + (\omega_{i'j} - \omega_{l(i'j)k(i'j)} - \eta_{k(i'j)}) \right) \right] \quad \text{(B)} \\ & + N^{-1} \sum_{i,j} \left[ \sum_{i' \neq i} \gamma_{ij} \gamma_{i'j} (\omega_{ij} - \omega_{l(ij)k(ij)} - \eta_{k(ij)}) (\omega_{i'j} - \omega_{l(i'j)k(i'j)} - \eta_{k(i'j)}) \right] \quad \text{(C)} \\ & + N^{-1} \sum_{i,j} \left[ \sum_{i' \neq i} \gamma_{ij} \gamma_{i'j} (\varepsilon_{ij} - \varepsilon_{l(ij)k(ij)}) \left( \mu_{W_{i'j}}(X_{i'j}) - \mu_{W_{i'j}}(X_{l(i'j)k(i'j)}) \right) \right] \quad \text{(D)} \\ & + N^{-1} \sum_{i,j} \left[ \sum_{i' \neq i} \gamma_{ij} \gamma_{i'j} (\varepsilon_{i'j} - \varepsilon_{l(i'j)k(i'j)}) \left( \mu_{W_{ij}}(X_{ij}) - \mu_{W_{ij}}(X_{l(ij)k(ij)}) \right) \right] \quad \text{(E)} \end{aligned}$$

Each of these five terms has mean zero due to the construction of  $\hat{\sigma}^2(X_{ij}, X_{i'j})$  and this remains true even if  $k(ij) = k(i'j)$ . The result follows since the conditional variances of all five terms vanish as  $J \rightarrow \infty$ . As above, the calculations rely on the fact that the number of times a unit (or a cluster) can be used in a match is bounded. For instance, consider the conditional variance of term (A). We have

$$\begin{aligned} & E \left[ \left( N^{-1} \sum_j \sum_i \sum_{i' \neq i} \gamma_{ij} \gamma_{i'j} (\eta_j^2 - \sigma_\eta^2) \right)^2 | \mathbf{X}, \mathbf{W} \right] = N^{-2} \sum_j \left( \sum_i \sum_{i' \neq i} \gamma_{ij} \gamma_{i'j} \right)^2 Var[\eta_j^2] \\ & \leq J^{-1} \left( \bar{I} (\bar{I} - 1) (1 + \bar{L}(k))^2 \right)^2 Var[\eta_j^2] \rightarrow 0 \end{aligned}$$

as  $J \rightarrow \infty$ . Therefore, we have  $Var[\hat{V}^E - E[\hat{V}^E | \mathbf{X}, \mathbf{W}] | \mathbf{X}, \mathbf{W}] \rightarrow 0$  which shows that  $\hat{V}^E \xrightarrow{p} V^E$ . ■

## B.2 Proof of Proposition 2

**Proposition 3** *Suppose assumptions 1 through 5 hold and let*

$$\begin{aligned} \hat{V}^{\tau(X)} = & N^{-1} \sum_j [\sum_i (\hat{Y}_{ij}(1) - \hat{Y}_{ij}(0) - \hat{\tau}_M)]^2 \quad (8) \\ & - N^{-1} \sum_j [\sum_i \sum_{i'} (2W_{ij}-1)(2W_{i'j}-1) \hat{\sigma}^2(X_{ij}, X_{i'j})] \\ & + 2N^{-1} \sum_j [M^{-1} \sum_i \sum_{i'} \sum_{s't' \in \mathcal{J}_M(i'j)} (2W_{ij}-1)(2W_{i'j}-1) 1\{t'=j\} \hat{\sigma}^2(X_{ij}, X_{s't'})] \\ & - N^{-1} \sum_j [M^{-2} \sum_i \sum_{i'} \sum_{st \in \mathcal{J}_M(ij)} \sum_{s't' \in \mathcal{J}_M(i'j)} (2W_{ij}-1)(2W_{i'j}-1) 1\{t=t'\} \hat{\sigma}^2(X_{st}, X_{s't'})]. \end{aligned}$$

Then, holding fixed cluster sizes, as  $J \rightarrow \infty$ ,  $\hat{V}^{\tau(X)} \xrightarrow{p} V^{\tau(X)}$ . An estimate of the marginal variance can then be computed as  $\hat{V}^E + \hat{V}^{\tau(X)}$ .

**Proof.** Note that

$$\begin{aligned} \sum_i \left( \hat{Y}_{ij}(1) - \hat{Y}_{ij}(0) - \tau \right) &= \sum_i \left( \mu_1(X_{ij}) - \mu_0(X_{ij}) - \tau \right) \\ &\quad + \sum_i (2W_{ij} - 1) \left( \varepsilon_{ij} - M^{-1} \sum_{st \in \mathcal{J}_M(ij)} \varepsilon_{st} \right) \\ &\quad + \sum_i (2W_{ij} - 1) \left[ \mu_{1-W_{ij}}(X_{ij}) - M^{-1} \sum_{st \in \mathcal{J}_M(ij)} \mu_{1-W_{ij}}(X_{st}) \right]. \end{aligned}$$

As  $J \rightarrow \infty$ , the matching discrepancy vanishes so we can ignore the last term and we have

$$\begin{aligned} E \left[ \left( \sum_i \left( \hat{Y}_{ij}(1) - \hat{Y}_{ij}(0) - \tau \right) \right)^2 \right] &\approx E \left[ \left( \sum_i (\tau(X_{ij}) - \tau) \right)^2 \right] + E \left[ \left( \sum_i (2W_{ij} - 1) \varepsilon_{ij} \right)^2 \right] \\ &\quad - 2E \left[ \begin{aligned} & \left( \sum_i (2W_{ij} - 1) \varepsilon_{ij} \right) \\ & \times \left( M^{-1} \sum_{i'} \sum_{st \in \mathcal{J}_M(i'j)} (2W_{i'j} - 1) \varepsilon_{st} \right) \end{aligned} \right] \\ &\quad + E \left[ \left( M^{-1} \sum_i \sum_{st \in \mathcal{J}_M(ij)} (2W_{ij} - 1) \varepsilon_{st} \right)^2 \right] \end{aligned}$$

Since we are interested in  $E \left[ \left( \sum_i (\tau(X_{ij}) - \tau) \right)^2 \right]$ , we need to compute the expected value of the final three terms under our assumed error components model. First, we have

$$E \left[ \left( \sum_i (2W_{ij} - 1) \varepsilon_{ij} \right)^2 \mid \mathbf{X}, \mathbf{W} \right] = \sum_i \left[ \sigma_\varepsilon^2(X_{ij}, W_{ij}) + \sum_{i' \neq i} (2W_{ij} - 1) (2W_{i'j} - 1) \sigma_\eta^2 \right].$$

Next, we have

$$\begin{aligned} &E \left[ \left( \sum_i (2W_{ij} - 1) \varepsilon_{ij} \right) \left( M^{-1} \sum_{i'} \sum_{st \in \mathcal{J}_M(i'j)} (2W_{i'j} - 1) \varepsilon_{st} \right) \mid \mathbf{X}, \mathbf{W} \right] \\ &= M^{-1} \sum_i \sum_{i'} \sum_{s't' \in \mathcal{J}_M(i'j)} (2W_{ij} - 1) (2W_{i'j} - 1) \begin{pmatrix} 1 \{t' = j\} \sigma_\eta^2 \\ +1 \{s't' = ij\} \sigma_\omega^2(X_{ij}, W_{ij}) \end{pmatrix} \end{aligned}$$

Finally, we have

$$\begin{aligned} &E \left[ \left( M^{-1} \sum_i \sum_{st \in \mathcal{J}_M(ij)} (2W_{ij} - 1) \varepsilon_{st} \right)^2 \mid \mathbf{X}, \mathbf{W} \right] \\ &= M^{-2} \sum_i \sum_{i'} \sum_{st \in \mathcal{J}_M(ij)} \sum_{s't' \in \mathcal{J}_M(i'j)} (2W_{ij} - 1) (2W_{i'j} - 1) \begin{pmatrix} 1 \{t = t'\} \sigma_\eta^2 \\ +1 \{st = s't'\} \sigma_\omega^2(X_{st}, W_{st}) \end{pmatrix}. \end{aligned}$$

Thus, an estimator of  $V^{\tau(X)}$  can be based on the the clustered variance of the imputed treatment effects (i.e.  $N^{-1} \sum_j \left[ \sum_i \left( \hat{Y}_{ij}(1) - \hat{Y}_{ij}(0) - \hat{\tau}_M \right) \right]^2$ ) and three correction terms that account for residual clustering. The estimator is obtain by replacing  $\sigma_\eta^2$  with the relevant  $\hat{\sigma}^2(X_{ij}, X_{st})$  term and  $\sigma_\varepsilon^2(X_{ij}, W_{ij}) = \sigma_\eta^2 + \sigma_\omega^2(X_{ij}, W_{ij})$  by  $\hat{\sigma}^2(X_{ij}, X_{ij}) = \hat{\sigma}_\varepsilon^2(X_{ij}) = \hat{\varepsilon}_{ij}^2/2$  in each of the above terms. ■

The formula for  $\hat{V}^{\tau(X)}$  simplifies when all units in each cluster have the same treatment

status. In this case,  $(2W_{ij} - 1)(2W_{i'j} - 1) = 1$  and the third term in (8) vanishes because other units in cluster  $j$  are never be used to impute missing outcomes for any  $ij$ , so that

$$\begin{aligned}\hat{V}^{\tau(X)} &= N^{-1} \sum_j \left[ \sum_i \left( \hat{Y}_{ij}(1) - \hat{Y}_{ij}(0) - \hat{\tau}_M \right) \right]^2 - N^{-1} \sum_j \left[ \sum_i \sum_{i'} \hat{\sigma}^2(X_{ij}, X_{i'j}) \right] \\ &\quad - N^{-1} \sum_j \left[ M^{-2} \sum_i \sum_{i'} \sum_{st \in \mathcal{J}_M(ij)} \sum_{s't' \in \mathcal{J}_M(i'j)} 1 \{t = t'\} \hat{\sigma}^2(X_{st}, X_{s't'}) \right]\end{aligned}$$

We now consider the role of duplicative matchings. Consider the third term in (8). This term reflects that fact that the  $\varepsilon_{ij}$  may be correlated with some  $\varepsilon_{st}$  for some  $st \in \mathcal{J}_M(i'j)$  if units in cluster  $j$  are used to impute missing outcomes for  $i'j$ . Assuming  $P = 0$ , this term should vanish as  $J \rightarrow \infty$ . Consider the fourth term in (8). This term itself can be written as the sum of three pieces

$$\begin{aligned}& M^{-2} \sum_i \sum_{i'} \sum_{st \in \mathcal{J}_M(ij)} \sum_{s't' \in \mathcal{J}_M(i'j)} (2W_{ij} - 1)(2W_{i'j} - 1) \left( \begin{array}{c} 1 \{t = t'\} \sigma_\eta^2 \\ +1 \{st = s't'\} \sigma_\omega^2(X_{st}, W_{st}) \end{array} \right) \\ = & M^{-2} \sum_i \sum_{st \in \mathcal{J}_M(ij)} (\sigma_\eta^2 + \sigma_\omega^2(X_{st}, W_{st})) \\ & + M^{-2} \sum_i \sum_{st \in \mathcal{J}_M(ij)} \sum_{(s't') \neq st \in \mathcal{J}_M(i'j)} 1 \{t = t'\} \sigma_\eta^2 \\ & + M^{-2} \sum_i \sum_{i' \neq i} \sum_{st \in \mathcal{J}_M(ij)} \sum_{s't' \in \mathcal{J}_M(i'j)} (2W_{ij} - 1)(2W_{i'j} - 1) \left( \begin{array}{c} 1 \{t = t'\} \sigma_\eta^2 \\ +1 \{st = s't'\} \sigma_\omega^2(X_{st}, W_{st}) \end{array} \right)\end{aligned}$$

The first piece reflects the sum of the variances of the residuals for  $ij$ 's matches and corresponds to a correction term in Abadie and Imbens (2006) that arises in the absence of clustering. The second component reflects the fact that multiple units in some cluster  $s$  might be used to impute outcomes for  $ij$ . If all units  $ij$  are only matched to units in distinct clusters (that is, we never use multiple units in some cluster  $s$  to impute outcomes for  $ij$ ), this term is zero. Thus, if  $P = 0$  this term should vanish as  $J \rightarrow \infty$ . Finally, the third piece represents overlap arising from the fact that multiple units in cluster  $j$  may be matched to some cluster  $s$ . This term should converge to 0 as  $J \rightarrow \infty$  if  $P=0$ . Thus, in the simplest case where  $P = 0$  so all problems with duplication vanish as  $J \rightarrow \infty$  we can consistently estimate  $V^{\tau(X)}$  using

$$\begin{aligned}\hat{V}^{\tau(X)} &= N^{-1} \sum_j \left[ \sum_i \left( \hat{Y}_{ij}(1) - \hat{Y}_{ij}(0) - \hat{\tau}_M \right) \right]^2 - N^{-1} \sum_j \sum_i \frac{K_M(ij)}{M^2} \hat{\sigma}_\varepsilon^2(X_{ij}) \\ &\quad - N^{-1} \sum_j \sum_i \sum_{i'} (2W_{ij} - 1)(2W_{i'j} - 1) \hat{\sigma}^2(X_{ij}, X_{i'j}).\end{aligned}$$