Matching Estimators with Few Treated and Many Control Observations

Ferman, Bruno

Sao Paulo School of Economics - FGV

4 May 2017

Matching Estimators with Few Treated and Many Control Observations*

Bruno Ferman†

Sao Paulo School of Economics - FGV

First Draft: May, 2017

This Draft: June, 2017

Please click here for the most recent version

Abstract

We analyze the properties of matching estimators when the number of treated observations is fixed and the number of control observations is large. We show that, under standard assumptions, the nearest neighbor matching estimator for the average treatment effect on the treated is asymptotically unbiased, even though this estimator is not consistent. Since large sample inferential techniques are not adequate in our setting, we provide inferential procedures based on the theory of randomization tests under approximate symmetry. We show that these tests are asymptotically valid when the number treated observations is fixed and the number of control observations goes to infinity. Our simulation results suggest that our inference methods provide better size and power when compared to existing alternatives even when the number of control observations is not large.

Keywords: matching estimator, treatment effect, hypothesis testing, randomization inference, bootstrap
JEL Codes: C12; C13; C21

*I would like to thank Sergio Firpo, Ricardo Masini, Cristine Pinto, Vitor Possebom, and Pedro Sant'Anna for comments and suggestions. Deivis Angeli provided outstanding research assistance.

†bruno.ferman@fgv.br

1 Introduction

Matching estimators have been widely used for the estimation of treatment effects under a conditional independence assumption (CIA).[1] In many cases, matching estimators have been applied in settings where (1) the interest is in the average treatment effect for the treated (ATET), and (2) there is a large reservoir of potential controls (Imbens and Wooldridge (2009)). Abadie and Imbens (2006) study the theoretical properties of matching estimators when the number of control observations grows at a higher rate than the number of treated observations. However, their asymptotic results still depend on both the number of treated and control observations going to infinity.

In this paper, we analyze the properties of matching estimators when the number of treated observations is fixed while the number of control observations is large. We show that the nearest neighbor matching estimator is asymptotically unbiased for the ATET under standard assumptions used in the literature on estimation of treatment effects under selection on unobservables. This result is consistent with Abadie and Imbens (2006), who show that the conditional bias of the matching estimator can be ignored, provided that the number of control observations increases faster enough relative to the number of treated observations. In their setting, the matching estimator would be consistent and asymptotically normal. Differently from Abadie and Imbens (2006), since we consider the case in which the number of treated observations is fixed, the variance of the matching estimator does not converge to zero and the estimator will not generally be asymptotically normal in our setting. Our theoretical results should provide a better approximation to the behavior of the matching estimator relative to Abadie and Imbens (2006) in settings where not only there is a larger number of control observations relative to treated observations, but also the number of treated observations are not large enough, so that we cannot rely on asymptotic results.[2] When the dimensionality of the covariates is low and we consider matching estimators with few nearest neighbors, our Monte Carlo (MC) simulation results suggest that the bias of the matching estimator is close to zero even when the number of control observations is not particularly large, regardless of the number of treated observations. Increasing the dimensionality of the covariates and/or increasing the number of nearest neighbors implies that we need an increasing number of controls so that our approximation remains reliable.

The fact that the matching estimator is not asymptotically normal in our setting poses important challenges when it comes to inference. Inference based on the asymptotic distribution of the matching estimator

[1] See Imbens (2004), Imbens and Wooldridge (2009), and Imbens (2014) for reviews.

[2] The finite sample properties of matching estimators have been evaluated in detail in simulations in Frolich (2004) and Busso et al. (2014). Differently, we provide theoretical and simulation results holding the number of treated observations fixed, but relying on the number of control observations going to infinity.

derived in Abadie and Imbens (2006) would not be valid if the number of treated observations is small, even if there are many control observations. In finite samples, Rosenbaum (1984) and Rosenbaum (2002) consider permutation tests for observational studies under strong ignorability. However, these tests rely on strong assumptions.[3] We consider alternative inference methods in our setting. We first provide two inference procedures based on the theory of randomization tests under an approximate symmetry assumption developed in Canay et al. (2017). One test relies on permutations while the other one relies on group transformations given by sign changes. We show that these tests provide asymptotically valid hypothesis testing when the number of control observations goes to infinity, even when the number of treated observations is fixed. We also consider the approach suggested in Rothe (2017), which provides valid inference in finite samples when the outcome is normally distributed, and a wild bootstrap procedure proposed in Otsu and Rai (2015).[4,5]

Our simulation results show that, with few treated observations, the test based on the asymptotic distribution derived in Abadie and Imbens (2006) and the test based on wild bootstrap over-reject under the null. We find over-rejection even when the number of treated observations is not particularly small, for example, with 50 treated observations. In the absence of finite-sample bias, the two randomization inference methods we propose and the method suggested in Rothe (2017) control well for size with few treated observations in all scenarios, even when the number of control observations is not large. The randomization inference test based on permutations is the most powerful among these three tests in most scenarios. However, it relies on a sharper null hypothesis that, conditional on observables, the distribution of potential outcomes when treated and untreated is the same. The randomization inference test based on sign changes and the test based on Rothe (2017) rely on less stringent null hypotheses, but they have poor power in some scenarios.[6] These tests remain with correct size even when we consider the bias of the matching estimator, as long as the number of nearest neighbors used in the estimation and the dimension of the matching covariates are relatively low. With matching estimators using many nearest neighbors and/or multidimensional covariates we may need a large number of control observations so that we do not have over-rejection under the null. Taken together, our MC results suggest that the alternatives we propose may be more reliable than tests

[3] Rosenbaum (1984) assumes that the propensity score follows a logit model, while Rosenbaum (2002) assumes that observations are matched in pairs such that the probability of treatment assignment is the same conditional on the pair.

[4] The approach suggested in Rothe (2017) is valid in finite samples if the bias of the matching estimator is negligible. If the number of treated observations is small but the number of control observations is large, then we show that this will be the case.

[5] Otsu and Rai (2015) suggest a weighted bootstrap procedure in which the wild bootstrap is a particular case. We focus on the wild bootstrap because, with few treated and many control observations, the non-parametric version of their weighted bootstrap would have a potential problem that some bootstrap samples would not have any treated observation.

[6] The test based on sign changes has poor power when the number of nearest neighbors used for estimation is large relative to the number of control observations, while the test based on Rothe (2017) has poor power when we use few nearest neighbors in the estimation. Also, note that while these tests rely on less stringent null hypotheses, the test based on sign changes require that errors are symmetric around zero and the test based on Rothe (2017) rely on normality (although, as explained in Rothe (2017), this assumption is an "asymptotically irrelevant").

that rely on large number of treated and control observations even when the number of treated observations is not particularly small and when the number of control observations is not particularly large. For example, our permutation test provided more reliable hypothesis testing relative to existing alternatives even when we have 100 observations equally divided in two groups.

The remainder of this paper proceeds as follows. We present our theoretical setup in Section 2. The intuition behind our main assumptions are exactly the same as in standard models under CIA, although they are stated differently in order to consider our setting with fixed number of treated observations and many control observations. In Section 3, we derive the asymptotic distribution of the matching estimator and derive conditions under which it is asymptotically unbiased. In Section 4, we consider alternative inference methods for our setting. In Section 5, we evaluate in MC simulations the properties of the matching estimator and we contrast alternative inferential methods. Concluding remarks, including a discussion on potential implications of our results for Synthetic Control applications, are presented in Section 6.

2 Setting and Notation

We observe a sample $\{Y_i, X_i\}_{i=1}^{N_1}$ that receives treatment ($W_i = 1$) and a sample $\{Y_i, X_i\}_{i=N_1+1}^{N}$ that does not receive treatment ($W_i = 0$), where Y_i is the observed outcome of observation i, and X_i is a set of covariates. We assume that X_i is a continuous random vector of dimension k in $\mathbb{R}^k$.[7] Following Rubin (1973), let $Y_i(1)$ denote the potential outcome if observation i received treatment and $Y_i(0)$ denote the potential outcome if observation i did not receive treatment. Therefore, $Y_i = W_iY_i(1) + (1 - W_i)Y_i(0)$. We consider the case in which the number of treated observations, N_1, is finite, while the number of control observations, $N_0 = N - N_1$, is large. Let $\mathcal{I}_w$ denote the set of indexes for observations with $W_i = w$. We aim to estimate the treatment effect on the treated, which we denote by:

$$\tau = \frac{1}{N_1} \sum_{i \in \mathcal{I}_1} \mathbb{E}\left[Y_i(1) - Y_i(0) | X_i, W_i = 1\right] \tag{1}$$

Note that we focus on the estimation of the treatment effect on the treated because, given our setting with N_1 finite and N_0 large, there is no hope in constructing a counterfactual for the control observations using only a finite set of treated observations. Also, for most of our results we will consider the properties

[7] We abstract from the case in which components of X_i is discrete because, as argued in Abadie and Imbens (2006), discrete covariates with a finite number of support points can be easily dealt with by analyzing estimation of average treatment effects within subsamples defined by their values.

of the matching estimator conditional on the realization of $\{X_i\}_{i\in\mathcal{I}_1}$.[8] We consider the unconditional case in remark 2.

We present our main assumptions in a slightly different way relative to Abadie and Imbens (2006) in order to consider the case in which the number of control observations goes to infinity while the number of treated observations is fixed. The main intuition behind our assumptions, however, remain the same.

We assume that the sample we observe for the treated (control) observations consists of i.i.d. observations of individuals with $W_i = 1$ ($W_i = 0$), and that treated and control observations are independent.

Assumption 1 (Sample) *$\{Y_i(0), Y_i(1), X_i\}_{i\in\mathcal{I}_w}$ consists of N_w i.i.d. observations with $W_i = w$. Furthermore, we assume that individuals in the treated and control samples are independent.*

The following assumption restricts the way in which the distributions of the treatment and control observations may differ.

Assumption 2 (Conditional Independence Assumption) *Conditional on X_i, the distribution of $Y_i(0)$ is the same for i in the treated and in the control groups.*

Assumption 2 is equivalent to the conditional independence assumption (CIA). While in assumption 1 we allow for different distributions of $(Y_i(0), Y_i(1), X_i)$ whether i is treated or control, assumption 2 restricts that the conditional distribution of $Y_i(0)$ given X_i is the same for both treatment and control observations. However, the density $f_1(X_i)$ for $i \in \mathcal{I}_1$ can potentially be different from the density $f_0(X_i)$ for $i \in \mathcal{I}_0$. This is what generates potential bias in a simple comparison of means between treated and control groups, without taking into account that these groups might have different distributions of covariates X_i.

The next assumption states that possible values of X_i for the treated observations are in the support of X_i for the control observations.

Assumption 3 (Overlap) *$\mathbb{X}_1 \subset \mathbb{X}_0$, where $\mathbb{X}_w$ is the support of $f_w(X_i)$, for $w \in \{0, 1\}$*

Assumption 3 replaces the standard assumption that $Pr(W = 1|X = x) < 1 - \eta$ for some $\eta > 0$. This assumption will guarantee that, for each i in the treated group, we will be able to find an observation j in the control group with covariates X_j arbitrarily close to X_i when $N_0 \to \infty$.

The main identification problem arises from the fact that we observe either $Y_i(1)$ or $Y_i(0)$ for each observation i. Note that, if we had two observations, $i \in \mathcal{I}_1$ and $j \in \mathcal{I}_0$, with $X_i = X_j = x$, then, under

[8] Note that our analysis is a mixture of finite sample (N_1 is finite) and large sample ($N_0 \to \infty$), which is similar to the setting considered in Ferman and Pinto (2015) for the differences-in-differences estimator. We consider our results conditional on the realization of the treated covariates in an analogy to what is usually done in the study of finite sample properties of estimators.

assumption 2, $\mathbb{E}[Y_i|W_i = 1, X_i = x] - \mathbb{E}[Y_j|W_j = 0, X_j = x] = \mathbb{E}[Y_i(1) - Y_i(0)|X_i = x, W_i = 1]$. That is, we would be able estimate the average treatment effect conditional on each value of the covariates $X_i = x$. Then we would be able to aggregate these effects to construct the ATET. The main challenge is that, with a continuous random variable X_i, the probability of finding observations with exactly the same X_i is zero. The idea of the nearest neighbor matching estimator is to input the missing potential outcomes of a treated observation $i \in \mathcal{I}_1$ with observations in the control group $j \in \mathcal{I}_0$ that are as close as possible in terms of covariates X_i. More specifically, for a given metric $d(a, b)$ in $\mathbb{R}^k$, let $\mathcal{J}_M(i)$ be the set of M nearest neighbors in the control group of observation $i \in \mathcal{I}_1$. Then the matching estimator is given by:

$$\hat{\tau} = \frac{1}{N_1} \sum_{i \in \mathcal{I}_1} \left[Y_i - \frac{1}{M} \sum_{j \in \mathcal{J}_M(i)} Y_j \right] \tag{2}$$

3 Asymptotic Unbiasedness

For $w \in \{0, 1\}$, we define $\mu(x, w) = \mathbb{E}[Y|X = x, W = w]$ and $\epsilon_i = Y_i - \mu(X_i, W_i)$. Since we are focusing on the ATET, we also define $\mu_w(x) = \mathbb{E}[Y(w)|X = x, W_i = 1]$.[9] Under assumption 2, we have that $\mu(x, 0) = \mu_0(x)$. Using this notation, note that the parameter of interest (ATET) is given by:

$$\tau = \frac{1}{N_1} \sum_{i \in \mathcal{I}_1} [\mu_1(X_i) - \mu_0(X_i)] \tag{3}$$

and:

$$\hat{\tau} = \frac{1}{N_1} \sum_{i \in \mathcal{I}_1} \left[\left(\mu_1(X_i) - \frac{1}{M} \sum_{j \in \mathcal{J}_M(i)} \mu_0(X_j) \right) + \left(\epsilon_i - \frac{1}{M} \sum_{j \in \mathcal{J}_M(i)} \epsilon_j \right) \right] \tag{4}$$

We show that $\hat{\tau}$ is an asymptotically unbiased estimator for the ATET when the number of treated observations is fixed and the number of control observations grows, and we derive its asymptotic distribution in this setting.

Proposition 1 *Under assumptions 1, 2, and 3:*

1. *If $\mu_0(x)$ is continuous and bounded, then $\mathbb{E}[\hat{\tau}|\{X_i\}_{i \in \mathcal{I}_1}] \to \frac{1}{N_1} \sum_{i \in \mathcal{I}_1} [\mu_1(X_i) - \mu_0(X_i)]$*

2. *If $\tilde{f}(x) = \mathbb{E}[f(Y(0))|X = x]$ is continuous and bounded for any $f(y)$ continuous and bounded, then,*

[9] Note that Abadie and Imbens (2006) define $\mu_w(x) = \mathbb{E}[Y(w)|X = x]$. We use a slightly different definition because we are focusing on the ATET.

conditional on $\{X_i\}_{i\in\mathcal{I}_1}$*:*

$$\hat{\tau} \xrightarrow{d} \frac{1}{N_1}\sum_{i\in\mathcal{I}_1}(\mu_1(X_i)-\mu_0(X_i)) + \frac{1}{N_1}\sum_{i\in\mathcal{I}_1}\left(\epsilon_i - \frac{1}{M}\sum_{m=1}^{M}\epsilon_m(X_i)\right) \tag{5}$$

where $\epsilon_m(X_i) \overset{d}{=} Y_i(0)|X_i - \mu_0(X_i)$ *for* $i \in \mathcal{I}_1$*, and* $\epsilon_m(X_i)$ *is independent across* m *and* i*.*

Proof. Let $X^i_{(m)}$ be the covariate value of the m-closest match to observation i. The main intuition for the results in Proposition 1 is that, for a fixed $X_i = \bar{x}$, $X^i_{(m)} \xrightarrow{p} \bar{x}$ when $N_0 \to \infty$, because we will always be able to find M observations in the control group that are arbitrarily close to $\bar{x}$. Independence of $\epsilon_m(X_i)$ across m and i follows from the fact that the probability of two treated observations sharing the same nearest neighbor converges to zero. See details in Appendix A.1. ■

Proposition 1 shows that, conditional on the realization of $\{X_i\}_{i\in\mathcal{I}_1}$, the expected value of the matching estimator converges to $\tau = \frac{1}{N_1}\sum_{i\in\mathcal{I}_1}(\mu_1(X_i)-\mu_0(X_i))$, which is the ATET. We also derive the asymptotic distribution of the matching estimator, which is centered on τ. This result is important for the construction of the inference methods we propose in Section 4.

Remark 1 The condition that $\mu_0(x)$ is continuous and bounded would be satisfied if we assume that $\mu_0(x)$ is continuous and $\mathbb{X}_0$ is compact, as is assumed in Abadie and Imbens (2006). The assumption used in part 2 of Proposition 1 implies that the conditional distribution of $Y(0)$ given $X = x$ changes smoothly with x. This guarantees that the outcome of the m-closest match to treated observation i, $Y^i_{(m)}$, converges in distribution to $Y_i(0)|X_i = \bar{x}$ when $X^i_{(m)} \xrightarrow{p} \bar{x}$.

Remark 2 We focus on the properties of the matching estimator conditional on $\{X_i\}_{i\in\mathcal{I}_1}$. We might be interested, however, on the unconditional properties of the matching estimator. For example, we may think that $\{Y_i, X_i\}_{i\in\mathcal{I}_1}$ is a finite sample from a super population.[10] Under the assumptions from part 1 of Proposition 1, we also have that $\mathbb{E}[\hat{\tau}] = \mathbb{E}\left\{\mathbb{E}[\hat{\tau}|\{X_i\}_{i\in\mathcal{I}_1}]\right\}$ converges to $\mathbb{E}[\mu_1(X_i)-\mu_0(X_i)|i\in\mathcal{I}_1]$, which is the average treatment effect on the treated *population*. Alternatively, we may think that there is indeed a finite N_1 population of treated individuals, but these individuals were selected to receive treatment from a larger population. See details in Appendix A.1.

Remark 3 With N_1 fixed, the estimator is not consistent. This happens because, with a fixed number of treated observations, we cannot apply a law of large numbers to the average of the error of the treated

[10] See Imbens and Wooldridge (2009) and Abadie et al. (2014) for a discussion on defining the estimand of interest as the treatment effect on the finite population at hand versus on a super population.

observations. Also, the matching estimator will not be asymptotically normal, unless we assume that the error ϵ_i is normal.

Remark 4 The nearest-neighbor matching estimator is not, in general, unbiased for a fixed N_0. This happens because, for a fixed N_0, it is not possible to guarantee a perfect match in terms of covariates. As shown in Abadie and Imbens (2006) and Abadie and Imbens (2011), in a setting in which the number of treated and control observations grow (even if the number of control observations grows at a faster rate), nearest-neighbor matching estimators include a conditional bias term that converges to zero at a rate that may be slower than $N^{1/2}$. In our setting, however, since the variance of the estimator does not go to zero when $N_0 \to \infty$, this bias will always be of a lower order relative to the variance of the estimator. For this reason, we are also able to consider slightly less restrictive assumptions when we derive the asymptotic properties of the estimator in our setting.

Remark 5 With additional assumptions, we can also guarantee that the bias-corrected matching estimator has the same asymptotic distribution as the matching estimator without bias correction. The intuition again is that $\hat{\mu}_0(X_i) - \hat{\mu}_0(X^i_{(m)})$ converge to zero when $N_0 \to \infty$ because $X^i_{(m)} \xrightarrow{p} X_i$. See details in Appendix A.1.

4 Inference

The fact that the matching estimator is not asymptotically normal in our setting poses an important challenge when it comes to inference. In particular, the analytic asymptotic variance estimator derived in Abadie and Imbens (2006) should not provide a good approximation in our setting. We therefore consider alternative inference methods in this setting. We propose two tests based on the theory of randomization tests under an approximate symmetry assumption developed in Canay et al. (2017), and we show that they are asymptotically valid when $N_0 \to \infty$, even with fixed N_1. The first test is based on group transformations given by permutations, while the second test is based on group transformations given by sign changes.[11] Then we consider a test based on Rothe (2017) confidence intervals for treatment effects under limited overlap and a test based on wild bootstrap derived in Otsu and Rai (2015). These tests differ in their underlying assumptions and null hypotheses. Moreover, the size and power of these tests depends crucially on the number of observations in the treatment and control groups, and also on the number of nearest neighbors

[11] A test based on permutations has been studied in the context of an approximate symmetry assumption in Canay and Kamat (2016) for regression discontinuity designs, while a test based on sign changes has been studied in the context of an approximate symmetry assumption in Canay et al. (2017) for a series of applications.

used in the estimation. In Section 5 we consider the finite sample properties of these tests, and we analyze in detail the conditions under which these tests provide valid size and non-trivial power.

4.1 Randomization Inference Test Based on Permutations

Consider a function of the data given by:

$$\tilde{S}_{N_0} = \left(\tilde{S}^0_{N_0,1}, \tilde{S}^1_{N_0,1}, ..., \tilde{S}^M_{N_0,1}, ..., \tilde{S}^0_{N_0,N_1}, \tilde{S}^1_{N_0,N_1}, ..., \tilde{S}^M_{N_0,N_1}\right)' \tag{6}$$

where $\tilde{S}^0_{N_0,i} = Y_i$ and $\tilde{S}^m_{N_0,i} = Y^i_{(m)}$ for $m = 1, ..., M$. That is, $\tilde{S}_{N_0}$ is a vector containing the outcomes of the treated observations and of their M-nearest neighbors. Note that the distribution of $\tilde{S}_{N_0}$ depends on N_0 because the quality of the matches will depend on the number of control observations. Note that the matching estimator is given by:

$$\hat{\tau} = \frac{1}{N_1}\sum_{i=1}^{N_1}\left(\tilde{S}^0_{N_0,i} - \frac{1}{M}\sum_{j=1}^{M}\tilde{S}^j_{N_0,i}\right) \tag{7}$$

Let $\tilde{G}_i$ be the set of all permutations $\pi_i = (\pi_i(0), ..., \pi_i(M))$ of $\{0, 1, ..., M\}$ and let $\pi = \otimes_{i=1}^{N_1}\pi_i$ and $\tilde{\mathbf{G}} = \otimes_{i=1}^{N_1}\tilde{G}_i$. Therefore, $\tilde{S}^{\pi}_{N_0} = \left(\tilde{S}^{\pi_1(0)}_{N_0,1}, \tilde{S}^{\pi_1(1)}_{N_0,1}, ..., \tilde{S}^{\pi_1(M)}_{N_0,1}, ..., \tilde{S}^{\pi_{N_1}(0)}_{N_0,N_1}, \tilde{S}^{\pi_{N_1}(1)}_{N_0,N_1}, ..., \tilde{S}^{\pi_{N_1}(M)}_{N_0,N_1}\right)'$. Note that $\tilde{\mathbf{G}}$ is the set of all permutations that reassign the treatment status conditional on having exactly one treated observation for each group of treated observation i and its M nearest neighbors.

Let $\tilde{K} = |\tilde{\mathbf{G}}|$ and denote by:

$$\tilde{T}^{(1)}(\tilde{S}_{N_0}) \leq \tilde{T}^{(2)}(\tilde{S}_{N_0}) \leq ... \leq \tilde{T}^{(\tilde{K})}(\tilde{S}_{N_0}) \tag{8}$$

the ordered values of $\{\tilde{T}(\tilde{S}^{\pi}_{N_0}) : \pi \in \tilde{\mathbf{G}}\}$, where:

$$\tilde{T}(\tilde{S}^{\pi}_{N_0}) = \left[\frac{1}{N_1}\sum_{i=1}^{N_1}\left(\tilde{S}^{\pi_i(0)}_{N_0,i} - \frac{1}{M}\sum_{j=1}^{M}\tilde{S}^{\pi_i(j)}_{N_0,i}\right)\right]^2 \tag{9}$$

We set $\tilde{k} = \lceil \tilde{K}(1-\alpha) \rceil$, where α is the significance level of the test, and define:

$$\begin{aligned} \tilde{K}^+(\tilde{S}_{N_0}) &= |\{1 \leq j \leq \tilde{K} : \tilde{T}^{(j)}(\tilde{S}_{N_0}) > \tilde{T}^{(k)}(\tilde{S}_{N_0})\}| \\ \tilde{K}^0(\tilde{S}_{N_0}) &= |\{1 \leq j \leq \tilde{K} : \tilde{T}^{(j)}(\tilde{S}_{N_0}) = \tilde{T}^{(k)}(\tilde{S}_{N_0})\}| \end{aligned} \tag{10}$$

The randomization test is given by:

$$\tilde{\phi}(S_{N_0}) = \begin{cases} 1 & \text{if } \tilde{T}(\tilde{S}_{N_1}) > \tilde{T}^{(k)}(\tilde{S}_{N_1}) \\ a(\tilde{S}_{N_0}) & \text{if } \tilde{T}(\tilde{S}_{N_1}) = \tilde{T}^{(k)}(\tilde{S}_{N_1}) \\ 0 & \text{if } \tilde{T}(\tilde{S}_{N_1}) < \tilde{T}^{(k)}(\tilde{S}_{N_1}) \end{cases} \tag{11}$$

where:

$$\tilde{a}(\tilde{S}_{N_0}) = \frac{\tilde{K}\alpha - \tilde{K}^{+}(\tilde{S}_{N_1})}{\tilde{K}^{0}(\tilde{S}_{N_1})}$$

Proposition 2 *Under the same assumptions used in part 2 of Proposition 1, testing a null hypothesis that* $Y_i(0)|X_i \overset{d}{=} Y_i(1)|X_i$ *for all* $i \in \mathcal{I}_1$ *using the decision rule defined in 11 satisfies, under the null,* $\mathbb{E}[\tilde{\phi}(\tilde{S}_{N_1})] \to \alpha$ *for any* $\alpha \in (0, 1)$.

Proof.

We apply Theorem 3.1 from Canay et al. (2017). We only need to show that, when $N_0 \to \infty$, the limiting distribution of $\tilde{S}_{N_0}$ under the null is invariant to the transformations in $\tilde{\mathbf{G}}$. From the proof of Proposition 1, note that $Y^i_{(m)} \xrightarrow{d} Y_i(0)|X_i$. Therefore, under the null that $Y_i(0)|X_i \overset{d}{=} Y_i(1)|X_i$, we have that $\tilde{S}^j_{N_0,i} \xrightarrow{d} Y_i(0)|X_i$ for all $j = 0, ..., M$. Moreover, asymptotically, $\tilde{S}^j_{N_0,i}$ is independent across i and j. Therefore, the asymptotic distribution of $\tilde{S}_{N_0}$ is invariant to the transformations in $\tilde{\mathbf{G}}$. ■

Remark 6 Rosenbaum (2002) considers Fisher exact tests in observational studies with matched pairs. They show that, if the probability of treatment assignment is the same for both observations in each pair, then a permutation test conditional on the pair is valid even in finite samples. With a finite N_0 and continuous X, however, it is not possible to guarantee this condition even under assumption 2, since we will not have, in general, a perfect match in terms of covariates. We show that this condition can be approximately satisfied when $N_0 \to \infty$ using the theory of randomization inference under approximate symmetry developed in Canay et al. (2017).

Remark 7 The randomization induced by $\tilde{a}(\tilde{S}_{N_0})$ when $\tilde{T}(\tilde{S}_{N_1}) = T^{(k)}(\tilde{S}_{N_1})$ guarantees an asymptotic rejection rate of α despite the discreteness of the randomization distribution. As stated in Canay et al. (2017), a non-randomized test that rejects if $\tilde{T}(\tilde{S}_{N_1}) > \tilde{T}^{(k)}(\tilde{S}_{N_1})$ is level α and, unless N_1 is very small, this should not lead to severe under-rejection.

Remark 8 This test is also asymptotically valid for biased-corrected matching estimators. In this case, we

define $\tilde{S}^0_{N_0,i} = Y_i - \hat{\mu}_0(X_i)$ and $\tilde{S}^m_{N_0,i} = Y^i_{(m)} - \hat{\mu}_0(X^i_{(m)})$.

4.2 Randomization Inference Test Based on Sign Changes

We consider now an alternative function of the data given by:

$$S_{N_0} = (\hat{\tau}_1, ..., \hat{\tau}_{N_1})' \tag{12}$$

where $\hat{\tau}_i = Y_i - \frac{1}{M}\sum_{j \in \mathcal{J}_M(i)} Y_j$. Note that each $\hat{\tau}_i$ depends on the M nearest neighbors of observation i, so it implicitly depends on N_0.

Following Canay et al. (2017), we consider a test statistic given by:

$$T(S_{N_0}) = \frac{|\hat{\tau}|}{\sqrt{\frac{1}{N_1 - 1}\sum_{i=1}^{N_1}(\hat{\tau}_i - \hat{\tau})^2}} \tag{13}$$

where $\hat{\tau} = \frac{1}{N_1}\sum_{i \in \mathcal{I}_1} \hat{\tau}_i$ is the matching estimator for the treatment effects on the treated.

We consider group of transformations given by $\mathbf{G} = \{-1, 1\}^{N_1}$, where $gS_{N_0} = (g_1\hat{\tau}_1, ..., g_{N_1}\hat{\tau}_{N_1})'$. Let $K = |\mathbf{G}|$ and denote by:

$$T^{(1)}(S_{N_0}) \leq T^{(2)}(S_{N_0}) \leq ... \leq T^{(K)}(S_{N_0}) \tag{14}$$

the ordered values of $\{T(gS_{N_0}) : g \in \mathbf{G}\}$. Let $k = \lceil K(1-\alpha) \rceil$, where α is the significance level of the test, and define:

$$\begin{aligned} K^+(S_{N_0}) &= |\{1 \leq j \leq K : T^{(j)}(S_{N_0}) > T^{(k)}(S_{N_0})\}| \\ K^0(S_{N_0}) &= |\{1 \leq j \leq K : T^{(j)}(S_{N_0}) = T^{(k)}(S_{N_0})\}| \end{aligned} \tag{15}$$

The test is given by:

$$\phi(S_{N_0}) = \begin{cases} 1 & \text{if } T(S_{N_1}) > T^{(k)}(S_{N_1}) \\ a(S_{N_0}) & \text{if } T(S_{N_1}) = T^{(k)}(S_{N_1}) \\ 0 & \text{if } T(S_{N_1}) < T^{(k)}(S_{N_1}) \end{cases} \tag{16}$$

where:

$$a(S_{N_0}) = \frac{K\alpha - K^+(S_{N_1})}{K^0(S_{N_1})}$$

In words, we calculate the test statistic $T(gS_{N_0})$ for all possible $gS_{N_0} = (g_1\hat{\tau}_1, ..., g_{N_1}\hat{\tau}_{N_1})'$, and then we compare the actual test statistic $T(S_{N_0})$ with the distribution $\{T(gS_{N_0}) : g \in \mathbf{G}\}$.

Proposition 3 *Under the same assumptions used in part 2 of Proposition 1, if we further assume that ϵ_i is symmetric around zero for all $i = 1, ..., N_1$, then testing a null hypothesis that $\mu_1(X_i) = \mu_0(X_i)$ for all $i \in \mathcal{I}_1$ using the decision rule defined in 16 satisfies, under the null, $\mathbb{E}[\phi(S_{N_1})] \to \alpha$ for any $\alpha \in (0, 1)$.*

Proof.

Again, we apply Theorem 3.1 from Canay et al. (2017). We only need to show that, when $N_0 \to \infty$, the limiting distribution of S_{N_0} under the null is invariant to sign changes. This will be true if, asymptotically, $\hat{\tau}_i$ and $\hat{\tau}_j$ are independent for $i \neq j$, and the distribution of $\hat{\tau}_i$ is symmetric around zero. Note that it is not required that $\hat{\tau}_i$ has the same distribution across i.

From the results in Proposition 1, we know that, under the null, the asymptotic distribution of $\hat{\tau}_i$ conditional on $\{X\}_{i\in\mathcal{I}_1}$ is given by $\epsilon_i - \frac{1}{M}\sum_{m=1}^{M}\epsilon_m(X_i)$, which is symmetric around zero given the assumption that ϵ_i is symmetric around zero for all $i = 1, ..., N_1$. Moreover, we also know from Proposition 1 that $\hat{\tau}_i$ are independent across i. Therefore, the assumptions for Theorem 3.1 from Canay et al. (2017) are satisfied. ■

Remark 9 In the case $M = 1$ the randomization test based on permutation tests are equivalent to the test based on sign changes. In this case, $\hat{\tau}_i = Y_i - Y^i_{(1)}$ so a sign transformation $-\hat{\tau}_i = Y^i_{(1)} - Y_i$ is equivalent to permute the treatment assignment within each pair.

Remark 10 Note that we can test the null hypothesis that the average treatment effect is equal to zero conditional on each covariate value in $\{X_i\}_{i\in\mathcal{I}_1}$. This null hypothesis is implied by more narrowly defined null hypotheses that are usually considered in Fisher-type tests, such as $Y_i(0)|X_i \overset{d}{=} Y_i(1)|X_i$ or $Y_i(0) = Y_i(1)$ with probability one.

Remark 11 Remark 7 also applies to this test.

4.3 Test based on Rothe (2017)

Rothe (2017) constructs robust confidence intervals for treatment effects estimators under limited overlap. The main idea of his approach is that, under limited overlap, "local sample sizes" can be effectively very

small in applications, so that approximations based on asymptotic theory would not be reliable. Instead, he constructs confidence intervals based on classical approaches to small sample inference. He shows that inference for the matching estimator can be considered as a generalized version of the Behrens-Fisher problem, where the test statistic is a studentized version of a linear combination of independent means. In the case in which X is discrete and can take J different values, the matching estimator for the ATET would be a linear combination of $J+1$ sample means.[12] Under the assumption that outcomes are normally distributed, he constructs a confidence interval that guarantees coverage greater or equal than $1-\alpha$ (Proposition 2 in Rothe (2017)). With continuous covariates, Rothe (2017) considers a partition of the data based on an estimated propensity score. He shows that, if the bias is negligible, then the conclusion based on discrete covariates is still valid.

We consider a slightly different way to partition the data, based on the nearest neighbors of the treated observations. More specifically, we consider a partition in which a treated observation i is joint with its M nearest neighbors. Therefore, if treated observations i and i' share at least one nearest neighbor, then they belong to the same partition. Suppose we end up with J partitions, and let $S_j(i)=1$ if observation i belongs to partition j. Then the estimator for the ATET would be given by:

$$\hat{\tau}' = \hat{\mu}_1 - \sum_{j=1}^{J} \hat{\mu}_0(j)\hat{f}_1(j) \tag{17}$$

where $\hat{\mu}_1 = \frac{1}{N_1}\sum_{i\in\mathcal{I}_1} Y_i$ is the average of the treated observations, $\hat{\mu}_0(j) = \frac{1}{\sum_{i\in\mathcal{I}_0} S_j(i)}\sum_{i\in\mathcal{I}_0} S_j(i)Y_i$ is the average of the control observations in partition j, and $\hat{f}_1(j) = \frac{\sum_{i\in\mathcal{I}_1} S_j(i)}{N_1}$ is the proportion of the treated observations that belong to partition j. Since the probability that two treated observations share the same nearest neighbor goes to zero when N_1 is fixed and $N_0 \to \infty$, note that, for a fixed M, the estimators $\hat{\tau}$ and $\hat{\tau}'$ are asymptotically equivalent. Importantly, this estimator is a linear combination of independent sample means, so the results from Rothe (2017) apply to this case, and, if we assume that the finite sample bias of the matching estimator is negligible, then we can construct a test statistic and calculate a critical value that guarantees a rejection rate of at most α for an α-level test if $Y_i|X$ is normally distributed, even in finite samples.

Remark 12 The calculation of the critical values requires at lest two control observations for each partition of the data.

[12] One for the treated observations, and J for the control observations with each $X=x$.

Remark 13 Differently from the tests presented in Sections 4.2 and 4.1, the null hypothesis in this case is that the average treated effect on the treated is equal to zero.

4.4 Test based on wild bootstrap

We also consider a bootstrap procedure based on Otsu and Rai (2015). As explained in Abadie and Imbens (2008), naive bootstrap procedures are not valid for matching estimators because they fail to reproduce the distribution of the number of times each observation is used as a match. Otsu and Rai (2015) overcome this problem by considering bootstrap procedures that treat the number of times an observation is used for a match as one of the characteristics of the sample. More specifically, let $\tilde{\tau}$ be a bias corrected estimator for the ATET using $\hat{\mu}_0(x)$ as an estimator for $\mu_0(x)$. Otsu and Rai (2015) note that:[13]

$$\begin{aligned} \tilde{\tau} &= \frac{1}{N_1}\sum_{i=1}^{N}\left[W_i(Y_i-\hat{\mu}_0(X_i)) - (1-W_i)\frac{K_M(i)}{M}(Y_i-\hat{\mu}_0(X_i))\right] \\ &= \frac{1}{N_1}\sum_{i=1}^{N}\tilde{\tau}_i \end{aligned} \tag{18}$$

where $K_M(i)$ is the number of times a control observation i is used as a match and $\tilde{\tau}_i = W_i(Y_i-\hat{\mu}_0(X_i)) - (1-W_i)\frac{K_M(i)}{M}(Y_i-\hat{\mu}_0(X_i))$. The weighted bootstrap counterpart for $\sqrt{N_1}(\tilde{\tau}-\tau)$ is obtained as:

$$\sqrt{N_1}T^* = \sum_{i=1}^{N} e_i^*(\tilde{\tau}_i - W_i\tilde{\tau}) \tag{19}$$

where e_i^* are random variables satisfying specific conditions explained in Otsu and Rai (2015). Two particular cases that are encompassed in this model are nonparametric bootstrap (Efron (1979)) and wild bootstrap (Mammen (1993)). Since we are focusing on the case with few treated observations, a nonparametric bootstrap would likely generate bootstrap samples with no treated observations. Therefore, we focus on the wild bootstrap case.

An important disadvantage of this method relative to the randomization inference methods we propose is that this weighted bootstrap relies on the estimation of a conditional expectation function. Since, $\hat{\mu}_0(.)$ is chosen to fit Y_i for the treated observations, we expect that the observed $(Y_i-\hat{\mu}_0(X_i))$ should have a lower variance when compared to $(Y_i-\mu_0(X_i))$. Therefore, we expect that the bootstrap distribution will underestimate the variance of the estimator, leading to over-rejection in finite samples. Note that this is less

[13] We use a different notation compared to Otsu and Rai (2015).

of a problem in our inference method based on sign changes, even if we consider a bias-corrected estimator, because we would apply the sign change transformation on $\hat{\tau}_i = (Y_i - \hat{\mu}_0(X_i)) - \frac{1}{M}\sum_{j \in \mathcal{J}_M(i)}(Y_j - \hat{\mu}_0(X_j))$.

5 Monte Carlo Simulations

We use a data generating process (DGP) similar to the one used in Frolich (2004) and Busso et al. (2014) in our Monte Carlo (MC) simulations. Following Busso et al. (2014), these DGPs can be expressed as:

$$\begin{aligned} Y_i(0) &= m(Z_i) + \sigma\epsilon_i \\ W_i^* &= \alpha + \beta Z_i - U_i \end{aligned} \tag{20}$$

where $Z_i = \Lambda(\sqrt{2}X_i)$, and X_i is a normal covariate; the error term U_i is i.i.d. standard uniform and is independent of ϵ_i and X_i; W_i^* is the latent variable corresponding to treatment ($W_i = 1$ if $W_i^* > 0$). Since we want to consider the case in which N_1 is finite while N_0 is large, we generate a large population based on this DGP, and then we sampled a small number N_1 of treated observations and a large number N_0 of control observations.[14] Frolich (2004) considers five combinations of (α, β). For ease of exposition, we focus on the combination of (α, β) used in design 1 of Frolich (2004), which sets $\alpha = 0$ and $\beta = 1$. This is the design that induces the highest correlation between treatment assignment and covariate X among the parameters they consider.

We start presenting in Section 5.1 a simpler case in which $m(.) = 0$ and ϵ_i is normally distributed and independent of X, so that there is no selection on observables. This way we are able to focus on the size and power performance of the different inferential procedures in the absence of the finite sample bias of matching estimators. Note that, in this case, all assumptions in Rothe (2017) are satisfied. In Section 5.2, we consider a functional form $m(.)$ from Frolich (2004), so that the matching estimator is biased in finite samples.[15] This way, we can analyze how different specifications affect the finite sample bias of the matching estimator and the rejection rates for the different test procedures. For each scenario, we drew 10,000 samples for our MC simulations.

[14] We use the program avaliable at the supplemental appendix of Busso et al. (2014).

[15] For ease of exposition, we focus on specification 1 from 5.2. Results using alternative specifications are qualitatively the same. Results available upon request.

5.1 Simulations with no selection on observables

Test size

We start presenting results in a simpler case in which $Y_i(0)|X_i \sim N(0,1)$ and $Y_i(0) = Y_i(1)$. Note that, in this case, the matching estimator is unbiased even in finite samples. We present in Table 1 rejection rates for 5% tests using different inference methods for combinations of (N_1, N_0) where $N_1 \in \{5, 10, 25, 50\}$ and $N_0 \in \{20, 50, 500\}$. For ease of exposition, we include a superscript "+" when rejection rate is greater than 6% and a superscript "−" when rejection rate is lower than 4%.

We present in Panel A of Table 1 rejection rates using the test based on Abadie and Imbens (2006) for different matching estimators, varying the number of nearest neighbors, $M \in \{1, 2, 4, 10\}$. Note that rejection rates for a 5% test are higher than 12.4% when $N_1 = 5$ for all values of N_0 and M. This happens because the asymptotic distribution derived in Abadie and Imbens (2006) relies on $N_1 \to \infty$, even though they allow that N_0 grows at a faster rate than N_1. When N_1 increases, rejection rates go down. However, except for the case in which $N_1 = 500$, rejection rates do not become close to 5% when we increase N_1. For example, even with $N_0 = N_1 = 50$ we still find a rejection rate of greater than 7.3% for most specifications.

The results above suggest that rejection rates computed using the asymptotic variance derived in Abadie and Imbens (2006) may not be reliable when the number of treated observations is not large. We consider instead in Panel B of Table 1 rejection rates using the randomization inference test based on permutations. From Section 4.1, we know that this test is asymptotically valid when $N_0 \to \infty$ in part because the probability that different treated observations share the same nearest neighbor goes to zero. In finite samples, however, this may not be the case. In order to take that into account, we consider permutations of treatment status in partitions of the sample as discussed in Section 4.3.[16] Note that the probability that this finite sample adjustment is relevant goes to zero when $N_0 \to \infty$.[17] Rejection rates are remarkably close to 5% in all cases. The only exception is when $N_1 = 5$ and $M = 1$, because in this case there are relatively few possible permutations.[18]

In Panel C of Table 1 we present rejection rates using the randomization inference test based on sign changes, presented in Section 4.2. A key feature of the test based on sign changes is that $\hat{\tau}_i$ become

[16] We also consider the estimator $\hat{\tau}'$ described in 4.3. Note that $\hat{\tau}$ and $\hat{\tau}'$ are asymptotically equivalent. In our simulations, the correlation between these two estimators is around 0.95.

[17] Another alternative would be to consider a matching estimator without replacement. However, this would generate lower quality matches, which implies in more bias (Abadie and Imbens (2006)). Moreover, matching without replacement has the disadvantage that the estimator is not invariant to different sorting of the data.

[18] We use the non-randomized version of the test in which we do not reject in case of equality. Note that we could guarantee the correct size if we used a randomized version of the test. See details in Canay et al. (2017).

asymptotically independent, because the probability that two treated observations share the same nearest neighbor converges to zero. Note, however, that for finite N_0 there is a positive probability that $\hat{\tau}_i$ is correlated across i, as different treated observations may share the same nearest neighbor. For this reasons, we consider a slight modification on our test where we restrict to sign changes such that $g_i = g_j$ if i and j share the same nearest neighbor. Similar to the finite sample adjustment used in the test based on permutations, note that the probability that this modification is relevant converges to zero when $N_0 \to \infty$. When we consider the nearest-neighbor matching estimator with $M = 1$, rejection rates using this test are close to 5%, except when $N_1 = 5$. In this case, there are not many different group transformations, which explains why the test is conservative.[19] When we consider matching estimators with $M > 1$, then the test under-rejects even for larger N_1. This happens because increasing M increases the probability that different treated observations share the same nearest neighbors, which in turn reduces the number of group transformations. When $N_0 = 500$, this problem becomes less relevant, and rejection rates become close to 5%.[20]

We present in Panel D of Table 1 rejection rates for the test based on Rothe (2017), described in in Section 4.3. As explained in remark 12, this test is not well defined for the case with $M = 1$. While the test is well defined for $M = 2$, note that rejection rates are virtually equal to zero in this case. Therefore, while it is possible to guarantee that this test is level α even in finite samples, it is over-conservative for the case in which we use very few nearest neighbors. When we use more nearest neighbors, then rejection rates become closer to 5%. Finally, we present rejection rate using the wild bootstrap test in Panel E of Table 1. Following Otsu and Rai (2015), we estimate $\mu_0(x)$ using a linear regression with all control observations and we use the two point distribution suggested in Mammen (1993).[21] Note that this test over-rejects, except when we have $N_0 = 500$, in which case the estimation of $\hat{\mu}_0(x)$ does not underestimate the variance of $Y_i - \mu_0(x)$ for the control observations.[22]

Test power

Given that the two randomization inference tests and the test based on Rothe (2017) have correct test sizes (although, in some cases, they may be conservative), we consider the power of these tests. We present in Table

[19] Similar to the case of permutations, this happens because we use the non-randomized version of the test in which we do not reject in case of equality. Note that we could guarantee the correct size if we used a randomized version of the test.

[20] If we do not use the restriction of considering only sign changes such that $g_i = g_j$ if i and j share the same nearest neighbor, then rejection rates explode when we N_1 increases for a fixed N_0. This happens because the distribution generated by the sign change transformations will have a lower variance as it will not take into account that $\hat{\tau}_i$ are correlated across i. Results available upon request.

[21] That is, we assign $e_i^* = (\sqrt{5}-1)/2$ with probability $(\sqrt{5}+1)/2\sqrt{5}$ and $e_i^* = (\sqrt{5}+1)/2$ with probability $(\sqrt{5}-1)/2\sqrt{5}$.

[22] Consistent with this explanation, we find less over-rejection when we estimate $\mu_0(x)$ using only a constant. Results available upon request.

2 rejection rates when $Y_i(1) = Y_i(0)+0.4$ for these three tests. In most scenarios, the randomization inference test based on permutations have the highest power among these three alternatives. The randomization inference test based on sign changes has good power when we use few nearest neighbors relative to the number of control observations, but it has poor power otherwise. This is not surprising, given that this test is over-conservative when there are not many control observations relative to the number of nearest neighbors used in the estimation. Finally, the test based on Rothe (2017) has good power when we use many nearest neighbors, but it has poor power otherwise. Again, this is consistent with our previous finding that the test based on Rothe (2017) is over-conservative when we use a matching estimator with few nearest neighbors.

5.2 Simulations with selection on observables

In Section 5.1 we considered a simplified DGP such that potential outcomes are unrelated with covariates that determine treatment assignment. This allowed us to analyze the size and power properties of different inference methods in the absence of finite N_0 bias of the matching estimator. Now we consider a case in which potential outcomes are correlated with X, so the matching estimator is biased when N_0 is finite. We consider the first conditional expectation function $m(.)$ used in Frolich (2004), and we set $\sigma = \sqrt{0.1}$.[23]

We present in Panel A of Table 3 the average bias of the nearest-neighbor matching estimator. In columns 1 to 3 we consider the case with $M = 1$. For $N_0 = 20$, the matching estimator for the treatment effect on the treated has a bias of around 0.02 regardless of the number of observations in the treated group, which reflects the fact that, with a finite N_0, it is not possible to guarantee a perfect match in X for the treated observations and their nearest neighbors. This bias is equivalent to around 10% of the bias of a naive comparison between treated and control observations. Consistent with Proposition 1, the average bias goes down to zero when we increase the number of control observations, regardless of the number of treated observations. In particular, when we consider $N_0 = 500$, we cannot reject that the average bias is equal to zero with 10,000 simulations even when we consider as few as five treated observations. When we consider the matching estimator with more nearest neighbors, the bias gets significantly higher for any combination of (N_1, N_0), except for the case with $N_0 = 500$. This happens because we end up with poorer matches when we consider an estimator with more nearest neighbors. This loss in match quality is less relevant when we have many control observations, which explains why there is little increase in bias when we consider the case with $N_0 = 500$.

In Panel B of Table 3 we present the mean square error (MSE) of the matching estimators. While the

[23] Results using the other specifications are qualitatively the same. Results available upon request.

MSE is always decreasing in N_1 and N_0, there are two competing forces when we consider increases in M. On the one hand, using more nearest neighbors reduces the variance of the matching estimator. On the other hand, this increases the bias of the estimator. With $N_0 = 500$, since increasing M from one to ten has little impact on the bias, using more nearest neighbors - in this range - always reduces the MSE of the matching estimator. However, with smaller N_0 there are some cases in which increasing M actually increases the MSE, exposing the trade-off between bias and variance for the matching estimator.

We consider the implications of the finite N_0 bias of the matching estimator for our inference methods in Panels C-E of Table 3. With $M \in \{1, 2\}$, the test based on permutations still controls well for size. This happens because the finite N_0 bias of the matching estimator is negligible relative to the variance of the matching estimator, so it does not generate strong size distortions. When we consider $M \in \{4, 10\}$, however, we observe strong size distortions when N_0 is not large. This happens because both the bias increases and the variance of the estimator decreases, so the finite N_0 bias of the matching estimator becomes more relevant and generates larger size distortions.[24] The test based on sign changes never over-rejects. However, it is over-conservative (and has poor power) in the settings in which the bias would be largest. On the other extreme, the test based on Rothe (2017) only has good size and non-trivial power in specific scenarios when we use many nearest neighbors and there are many control observations.

5.3 Multidimensional covariates

The MC results in Section 5.2 are based on simulations in which the covariate X_i is unidimensional. As stressed in Abadie and Imbens (2006), the bias of the matching estimator converges to zero at a lower rate when X_i is multidimensional. While this does not affect our theoretical result in Proposition 1, this can have important effects on the finite N_0 behavioral of the matching estimator. In order to evaluate the implications of having a multidimensional X_i on our results while keeping our simulations comparable to the ones in Section 5.2, we include a marginal modification in the DGP by generating $k-1$ new random variables $\widetilde{X}_{2i}, ..., \widetilde{X}_{ki}$ with the same distribution as X_i that are independent of all other random variables in the model. Then we estimate the matching estimator using $\widetilde{X}_i = (X_i, \widetilde{X}_{2i}, ..., \widetilde{X}_{ki})'$ as covariates. Note that a mismatch in $\widetilde{X}_{k'i}$ for $k' = 2, ..., k$ would not directly generate bias in the matching estimator. However, the addition of these variables makes it more difficult to find a good match in terms of X_i, which might lead to higher bias.

[24] The over-rejection is more relevant if we set $\sigma = \sqrt{0.01}$ instead of $\sigma = \sqrt{0.1}$. This was expected, because decreasing the variance of ϵ_i reduces the variance of the matching estimator, but it does not affect the average finite N_0 bias. Still, rejection rates for the permutation test are close to 5% when we use $M = 1$ or when $N_0 = 500$. Results available upon request.

We present the MC results for the case with $k = 2$ in Table 4. Note that, for a given (N_1, N_0), the average bias of the matching estimator is higher when compared to the case with $k = 1$. Still, we find that the average bias goes to zero with N_0 for any given N_1, which is consistent with our Proposition 1. Rejection rates using our randomization inference test based on permutations remain close to 5% when we consider the matching estimator with $M = 1$. When we use more nearest neighbors, then we can still have rejection rates close to 5%, provided that we have a large number of control observations. The results in Table 5 show that we would require even larger N_0 to maintain the bias under control and rejection rates close to 5% when $k = 4$.

Overall, these results suggest that our permutation test can still be reliable with multidimensional covariates. However, one should be careful that our asymptotic approximations would require an increasing number of control observations to be reliable when one increases the number of covariates.

5.4 Bias-corrected matching estimator

The over-rejection we find in the DGP with selection on observables comes from the finite N_0 bias of the matching estimator. Following Abadie and Imbens (2011), we consider a bias-corrected matching estimator in which we estimate $\hat{\mu}_0(x)$ using a linear regression on X using only the control observations used as a match, with weights given by the number of times each observation was used. Then the bias corrected matching estimator is given by:

$$\tilde{\tau} = \frac{1}{N_1} \sum_{i \in \mathcal{I}_1} \left[(Y_i - \hat{\mu}_0(X_i)) - \frac{1}{M} \sum_{j \in \mathcal{J}_M(i)} (Y_j - \hat{\mu}_0(X_j)) \right] \tag{21}$$

We present results using this bias-corrected matching estimator in Table 6. While the average bias is reduced using this procedure, it generally comes at a cost of a higher MSE. The MSE is significantly higher when N_1 is very small, because in this case $\hat{\mu}_0(x)$ is estimated using very few observations. Note that this is the bias-corrected matching estimator one would estimate using the teffects command in Stata, so one should be careful when using this bias correction with few treated observations. Another interesting result is that there are some cases in which rejection rates using the randomization tests are higher when we use the bias-adjusted estimator, despite the fact that the average bias is lower. This happens because the bias adjustment $\hat{\mu}_0(X_i)$ is chosen to fit Y_i for the control observations, so in finite samples we under-estimate the variation generated by the control observations. Note that this is less problematic in our tests when compared to the wild bootstrap, but this still leads to some over-rejection. We also consider another bias-

adjustment in which we estimate $\hat{\mu}_0(X_i)$ using all control observations. In this case, we do not have the severe increase in MSE when N_1 is small. However, we still have that the bias adjustment induces some over-rejection when N_0 is not large. Moreover, misspecification of the conditional expectation function can severely reduce the bias-reduction gains in this case.[25]

Overall, it might be possible to construct a bias-corrected matching estimator if we have a large number of control observations. In this case, we would be able to use, for example, non-parametric estimation and have a good approximation to $\mu_0(0)$. However, with a fixed number of treated observations, in this case the matching estimator without correction would also work well in terms of bias and the randomization inference tests would provide good size and power, so it is unclear whether such bias correction would be warranted. When N_0 is not large, then one should be careful, as the bias correction can potentially do more harm than good.

6 Conclusion

We consider the asymptotic properties of matching estimators when the number of control observations is large, but the number of treated observations is fixed. We show that, in this setting, the nearest neighbor matching estimator is asymptotically unbiased for the ATET under standard assumptions used in the literature on estimation of treatment effects under selection on unobservables. Moreover, we provide tests based on the theory of randomization tests under approximate symmetry that are asymptotically valid when the number of control observations goes to infinity. Our theoretical results should provide a better approximation to the behavior of the matching estimator and more reliable hypothesis testing relative to Abadie and Imbens (2006) in settings in which not only there is a much larger number of control observations relative to treated observations, but also that the number of treated observations are not large enough, so that we cannot rely on asymptotic results. The results from our Monte Carlo (MC) suggest that our inference methods may be more reliable and more powerful than existing inference methods even when the number of control observations is not particularly large.

Finally, note that our setting is also related to the Synthetic Control (SC) method, which is an alternative to estimate treatment effects in comparative case studies (Abadie and Gardeazabal (2003), Abadie et al. (2010), and Abadie et al. (2015)). Díaz et al. (2015) explore the idea of a matching estimator that considers convex combinations of the characteristics of the individuals in the corresponding counterfactual. In this

[25] Results available upon request.

sense, the SC estimator would be a matching estimator as in Díaz et al. (2015) using the pre-treatment outcomes as covariates. The only difference is that the procedure used in Díaz et al. (2015) minimizes the sum of the distances between the characteristics of the treated observation and those of the control observations used in the convex combination. Our results indicate that, if treatment assignment is "as good as random" conditional on the pre-treatment outcomes, then a SC estimator that assigns weights using Díaz et al. (2015) procedure should be asymptotically unbiased when the number of control units goes to infinity and the number of pre-treatment periods is fixed.[26] Moreover, we provide tests that are asymptotically valid when the number of control units goes to infinity.[27] Our test could be a valid alternative to the placebo tests proposed in Abadie et al. (2010) for the SC estimator when there are multiple treated units and a large number of control units.[28] The main challenge to apply our results to the SC setting, however, is that one would need a very large number of control observations when the number of pre-treatment periods is large, so that our approximations become reliable, which may be infeasible in many SC applications.

[26] If however, treatment assignment is only "as good as random" conditional on common factors (which allows for some correlation between treatment assignment and post-treatment potential outcomes), then this would not be necessarily true. Gobillon and Magnac (2016) show that the SC estimator can be asymptotically unbiased if the number of control units and the number of pre-treatment periods go to infinity, while Abadie et al. (2010) show that, conditional on a perfect pre-treatment match, the bias of the SC estimator is bounded by a function that goes to zero when the number of pre-treatment periods increases, even if the number of control units is fixed. See also Ferman and Pinto (2016) for a discussion on the conditions for asymptotic unbiasedness for the SC estimator when the number of control units is fixed.

[27] Note that this test should only be asymptotically valid if we use Díaz et al. (2015) procedure to calculate the SC weights. Their procedure will guarantee that the SC unit for each treated unit will assign positive weights to only few donors that are very close in terms of pre-treatment outcomes as the corresponding treated unit, which will imply that the treatment effect estimators for each treated unit will be independent. If we use Abadie et al. (2010) original procedure, it is not clear that this will be the case.

[28] Ferman and Pinto (2017) show that the placebo tests proposed in Abadie et al. (2010) will not, in general, satisfy the approximate symmetry property required in Canay et al. (2017). See also Firpo and Possebom (2016) and Hahn and Shi (2016) for considerations on the placebo tests proposed in Abadie et al. (2010).

References

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller, "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program," *Journal of the American Statiscal Association*, 2010, *105* (490), 493–505.

_ , _ , and _ , "Comparative Politics and the Synthetic Control Method," *American Journal of Political Science*, 2015, *59* (2), 495–510.

_ and Guido W. Imbens, "Large Sample Properties of Matching Estimators for Average Treatment Effects," *Econometrica*, 2006, *74* (1), 235–267.

_ and _ , "On the Failure of the Bootstrap for Matching Estimators," *Econometrica*, 2008, *76* (6), 1537–1557.

_ and _ , "Bias-Corrected Matching Estimators for Average Treatment Effects," *Journal of Business & Economic Statistics*, 2011, *29* (1), 1–11.

_ and Javier Gardeazabal, "The Economic Costs of Conflict: A Case Study of the Basque Country," *American Economic Review*, 2003, *93* (1), 113–132.

_ , Susan Athey, Guido W. Imbens, and Jeffrey M. Wooldridge, "Finite Population Causal Standard Errors," Working Paper 20325, National Bureau of Economic Research July 2014.

Busso, Matias, John DiNardo, and Justin McCrary, "New Evidence on the Finite Sample Properties of Propensity Score Reweighting and Matching Estimators," *The Review of Economics and Statistics*, December 2014, *96* (5), 885–897.

Canay, Ivan A. and Vishal Kamat, "Approximate permutation tests and induced order statistics in the regression discontinuity design," Aug 2016.

_ , Joseph P. Romano, and Azeem M. Shaikh, "Randomization Tests under an Approximate Symmetry Assumption," *Econometrica*, 2017, *85* (3), 1013–1030.

Díaz, Juan, Tomás Rau, and Jorge Rivera, "A Matching Estimator Based on a Bilevel Optimization Problem," *The Review of Economics and Statistics*, October 2015, *97* (4), 803–812.

Efron, B., "Bootstrap Methods: Another Look at the Jackknife," *Ann. Statist.*, 01 1979, *7* (1), 1–26.

Ferman, Bruno and Cristine Pinto, "Inference in Differences-in-Differences with Few Treated Groups and Heteroskedasticity," MPRA Paper 67665, University Library of Munich, Germany November 2015.

_ **and** _ , "Revisiting the Synthetic Control Estimator," MPRA Paper 73982, University Library of Munich, Germany September 2016.

_ **and** _ , "Placebo Tests for Synthetic Controls," MPRA Paper 78079, University Library of Munich, Germany April 2017.

Firpo, Sergio and Vitor Possebom, "Synthetic Control Estimator: A Generalized Inference Procedure and Confidence Sets," April 2016.

Frolich, Markus, "Finite-Sample Properties of Propensity-Score Matching and Weighting Estimators," *The Review of Economics and Statistics*, 2004, *86* (1), 77–90.

Gobillon, Laurent and Thierry Magnac, "Regional Policy Evaluation: Interative Fixed Effects and Synthetic Controls," *Review of Economics and Statistics*, 2016. Forthcoming.

Hahn, Jinyong and Ruoyao Shi, "Synthetic Control and Inference," 2016.

Imbens, Guido, "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review," *Review of Economics and Statistics*, 2004.

_ , "Matching Methods in Practice: Three Examples," NBER Working Papers 19959, National Bureau of Economic Research, Inc March 2014.

Imbens, Guido W. and Jeffrey M. Wooldridge, "Recent Developments in the Econometrics of Program Evaluation," Technical Report 1 2009.

Mammen, Enno, "Bootstrap and Wild Bootstrap for High Dimensional Linear Models," *Ann. Statist.*, 03 1993, *21* (1), 255–285.

Otsu, Taisuke and Yoshiyasu Rai, "Bootstrap inference of matching estimators for average treatment effects," *Journal of the American Statistical Association*, 2015.

Rosenbaum, Paul R., "Conditional Permutation Tests and the Propensity Score in Observational Studies," *Journal of the American Statistical Association*, 1984, *79* (387), 565–574.

_ , "Covariance Adjustment in Randomized Experiments and Observational Studies," *Statist. Sci.*, 08 2002, *17* (3), 286–327.

Rothe, Christoph, "Robust Confidence Intervals for Average Treatment Effects Under Limited Overlap," *Econometrica*, 2017, *85* (2), 645–660.

Rubin, Donald B., "Matching to Remove Bias in Observational Studies," *Biometrics*, 1973, *29* (1), 159–183.

A Supplemental Appendix for "Matching Estimators with Few Treated and Many Control Observations

A.1 Proof of main results

Proposition 1

For a given realization of $X_i = \bar{x}$ for an observation in the treated group and for a given $\epsilon > 0$, consider the probability that the M-closest realizations of $\{X_j\}_{j\in\mathcal{I}_0}$ are such that $d(X_j, \bar{x}) < \epsilon$. Let $X^i_{(M)}$ be the M-closest match of observation i. Then:

$$\begin{aligned}\Pr\left(d(X^i_{(M)}, \bar{x}) > \epsilon\right) &= \sum_{m=0}^{M-1} \Pr\left(d(X_j, \bar{x}) < \epsilon \text{ for exactly } m \text{ observations}\right) \\ &= \sum_{m=0}^{M-1} \begin{pmatrix} N_0 \\ m \end{pmatrix} [\Pr(d(X_j, \bar{x}) < \epsilon)]^m [\Pr(d(X_j, \bar{x}) > \epsilon)]^{N_0 - m} \end{aligned} \tag{22}$$

Since $\bar{x} \in \mathbb{X}_0$, we have that $Pr(d(X_j, \bar{x}) < \epsilon) > 0$, which implies that $\Pr(d(X_j, \bar{x}) > \epsilon) < 1$. Therefore, we have that $\Pr\left(d(X^i_{(M)}, \bar{x}) > \epsilon\right) \to 0$. By analogy, the m-nearest neighbor of i for $m < M$ will also converge in probability to $\bar{x}$.

Now consider:

$$\mathbb{E}[\hat{\tau}|\{X_i\}_{i\in\mathcal{I}_1}] = \frac{1}{N_1} \sum_{i\in\mathcal{I}_1} \left(\mu_1(X_i) - \mathbb{E}\left[\frac{1}{M}\sum_{m=1}^{M} \mu_0(X^i_{(m)})\right]\right) \tag{23}$$

Since $\mu_0(x)$ is continuous and bounded and $X^i_{(m)} \xrightarrow{p} X_i$, then we have that $\mathbb{E}[\mu_0(X^i_{(m)})|X_i] \to \mu_0(X_i)$, which proves of proposition 1.

For part 2, assume that $\tilde{f}(x) = \mathbb{E}[f(Y(0))|X = x]$ is continuous and bounded for any $f : \mathbb{R} \to \mathbb{R}$ continuous and bounded. Let $Y^i_{(m)}$ be the outcome of the m-nearest neighbor of treated observation i. Therefore, for any $f(y)$ continuous and bounded, and for a given $X_i = \bar{x}$, we have that:

$$\mathbb{E}[f(Y^i_{(m)})] \quad = \quad E\left\{\mathbb{E}[f(Y^i_{(m)})|X^i_{(m)}]\right\} = E\left\{\tilde{f}(X^i_{(m)})\right\} \to \tilde{f}(\bar{x}) = E[f(Y(0))|X = \bar{x}] \tag{24}$$

By the Portmanteau Lemma, we have that $Y^i_{(m)} \xrightarrow{d} Y(0)|\{X = \bar{x}\}$. Under assumption 2, $Y^i_{(m)} \xrightarrow{d} \mu_0(X_i) + e_m(X_i)$, where $e_m(X_i) \overset{d}{=} Y_i(0)|X_i - \mu_0(X_i)$. Therefore, conditional on $\{X_i\}_{i\in\mathcal{I}_1}$:

$$\hat{\tau} = \frac{1}{N_1}\sum_{i\in\mathcal{I}_1}\left[Y_i - \frac{1}{M}\sum_{m=1}^{M} Y^i_{(m)}\right] \xrightarrow{d} \frac{1}{N_1}\sum_{i\in\mathcal{I}_1}\left[(\mu_1(X_i) - \mu_0(X_i)) + \left(\epsilon_i - \frac{1}{M}\sum_{m=1}^{M}\epsilon_m(X_i)\right)\right] \tag{25}$$

Now we just have to show that $\epsilon_m(X_i)$ is independent across m and i. Since X_i is a continuous random variable, then $X_i \neq X_j$ with probability one for $i \neq j$ with $i, j \in \mathcal{I}_1$. Since there is a finite number of treated observations, then it must be that, conditional on $\{X_i\}_{i=1}^{N_1}$, there is an $\eta > 0$ such that $d(X_i, X_j) > \eta$ for all $i, j \in \mathcal{I}_1$ with $i \neq j$. However, we know that $Pr(d(X_i, X^i_{(m)}) > \epsilon) \to 0$ for all $\epsilon > 0$. Therefore, the probability that $k \in \mathcal{I}_0$ belongs to $\mathcal{J}_M(i)$ and $\mathcal{J}_M(j)$ converges to zero. Therefore, under the assumption that the errors ϵ_i are independent across i (which is guaranteed from assumption 1), we have that $\epsilon_m(X_i)$ is independent across m and i.

Unconditional Expectation

Now we consider the unconditional expectation of $\hat{\tau}$:

$$\mathbb{E}[\hat{\tau}] = \mathbb{E}\{\mathbb{E}[\hat{\tau}|\{X_i\}_{i\in\mathcal{I}_1}]\} = \frac{1}{N_1}\sum_{i\in\mathcal{I}_1}\mathbb{E}\left[\mu_1(X_i) - \frac{1}{M}\sum_{m=1}^{M}\mu_0(X^i_{(m)})\right] \tag{26}$$

We need that $\mathbb{E}[\mu_0(X^i_{(m)})] \to \mathbb{E}[\mu_0(X_i)]$. We know that $\mathbb{E}[\mu_0(X^i_{(m)})|X_i] \to \mu_0(X_i)$ for all X_i. Again using the fact that $\mu_0(x)$ is continuous and bounded, we have that $\mathbb{E}[\mu_0(X^i_{(m)})] = \mathbb{E}\{\mathbb{E}[\mu_0(X^i_{(m)})|X_i]\} \to \mathbb{E}[\mu_0(X_i)]$. Therefore:

$$\mathbb{E}[\hat{\tau}] \to \mathbb{E}\left[\mu_1(X_i) - \mu_0(X_i)\right] \tag{27}$$

where this expectation is taken according to $f_1(x)$, the density function of the treated units.

Bias-corrected Matching Estimator

We consider the bias-corrected matching estimator using linear least squares on the nearest neighbors to estimate $\mu_0(x)$. This is the model used in Abadie and Imbens (2011). Considering, for simplicity, the case with $k = 1$, note that:

$$\hat{\tau}_{biasadj} = \hat{\tau} + \hat{\beta}\left(X^i_{(m)} - X_i\right) \tag{28}$$

where $\hat{\beta} = \frac{\sum_{i\in\mathcal{I}_1}\left(X^i_{(m)}-\bar{X}_1\right)Y^i_{(m)}}{\sum_{i\in\mathcal{I}_1}\left(X^i_{(m)}-\bar{X}_1\right)^2}$ and $\bar{X} = \frac{1}{N_1}\sum_{i\in\mathcal{I}_1} X^i_{(m)}$. We assume that $Y_i(0)|X_i = x$ is uniformly bounded for almost all $x \in \mathbb{X}_0$ and that X_i is bounded.[29] Define $\mathcal{X} = \sum_{i\in\mathcal{I}_1}(X^i_{(m)} - \bar{X}_1)^2$. If we have at least two treated observations, then note that $\exists C_1 > 0$ such that $\Pr(\mathcal{X} < C_1) \to 0$. Therefore:

$$\begin{aligned}
\Pr\left(|\hat{\beta}| \geq c\right) &= \Pr\left(\left|\frac{\sum_{i\in\mathcal{I}_1}\left(X^i_{(m)} - \bar{X}_1\right)Y^i_{(m)}}{\mathcal{X}}\right| \geq c\right) \leq \Pr\left(\frac{\sum_{i\in\mathcal{I}_1}\left|X^i_{(m)} - \bar{X}_1\right|\left|Y^i_{(m)}\right|}{\mathcal{X}} \geq c\right) \\
&\leq \Pr\left(\frac{C_2\sum_{i\in\mathcal{I}_1}\left|Y^i_{(m)}\right|}{\mathcal{X}} \geq c \mid \mathcal{X} < C_1\right)\Pr(\mathcal{X} < C_1) + \Pr\left(\frac{C_2\sum_{i\in\mathcal{I}_1}\left|Y^i_{(m)}\right|}{C_1} \geq c \mid \mathcal{X} > C_1\right)\Pr(\mathcal{X} > C_1)
\end{aligned} \tag{29}$$

Since $\Pr(\mathcal{X} < C_1) \to 0$, the first term converges to zero. Since we assume that $Y_i(0)|X_i = x$ is uniformly bounded for almost all $x \in \mathbb{X}_0$, we can always find c such that the second term is lower than any $\eta > 0$, which implies that $\hat{\beta} = O_p(1)$. Therefore, $\hat{\beta}(X^i_{(m)} - X_i) = o_p(1)$, so $|\hat{\tau}_{biasadj} - \hat{\tau}| = o_p(1)$.[30]

[29] Note that these assumptions are weaker than the assumptions in Abadie and Imbens (2011).

[30] Note that the proof would be easier if we used all control observations to estimate $\mu_0(x)$ using linear least squares. In this case, $\hat{\beta}$ would converge to the population OLS coefficients.

Table 1: **Test Sizes - no selection on observable**

	$M=1$			$M=2$			$M=4$			$M=10$		
	$N_0=20$	$N_0=50$	$N_0=500$	$N_0=20$	$N_0=50$	$N_0=500$	$N_0=20$	$N_0=50$	$N_0=500$	$N_0=20$	$N_0=50$	$N_0=500$
	(1)	(2)	(3)	(4)	(5)	(6)						
Panel A: test based on AI (2006)												
$N_1=5$	0.139^+	0.144^+	0.156^+	0.129^+	0.136^+	0.154^+	0.124^+	0.127^+	0.148^+	-	-	-
$N_1=10$	0.097^+	0.097^+	0.099^+	0.089^+	0.090^+	0.091^+	0.088^+	0.087^+	0.092^+	0.082^+	0.080^+	0.088^+
$N_1=25$	0.086^+	0.073^+	0.064^+	0.090^+	0.076^+	0.065^+	0.084^+	0.068^+	0.064^+	0.066^+	0.069^+	0.065^+
$N_1=50$	0.093^+	0.072^+	0.056	0.100^+	0.075^+	0.057	0.093^+	0.079^+	0.058	0.072^+	0.066^+	0.060
Panel B: test based on RI, permutation												
$N_1=5$	0.020^-	0.020^-	0.017^-	0.045	0.046	0.048	0.049	0.047	0.046	-	-	-
$N_1=10$	0.048	0.051	0.052	0.048	0.048	0.050	0.048	0.050	0.048	0.047	0.048	0.049
$N_1=25$	0.050	0.049	0.047	0.049	0.049	0.048	0.048	0.045	0.049	0.050	0.051	0.047
$N_1=50$	0.048	0.049	0.049	0.052	0.048	0.049	0.050	0.052	0.050	0.052	0.048	0.052
Panel C: test based on RI, sign changes												
$N_1=5$	0.002^-	0.008^-	0.015^-	0.001^-	0.003^-	0.012^-	0.000^-	0.001^-	0.009^-	-	-	-
$N_1=10$	0.025^-	0.042	0.050	0.006^-	0.024^-	0.046	0.000^-	0.006^-	0.042	0.000^-	0.000^-	0.033^-
$N_1=25$	0.044	0.052	0.049	0.018^-	0.046	0.051	0.000^-	0.024^-	0.052	0.000^-	0.000^-	0.049
$N_1=50$	0.049	0.049	0.048	0.030^-	0.056	0.049	0.000^-	0.033^-	0.050	0.000^-	0.000^-	0.051
Panel D: test based on Rothe (2017)												
$N_1=5$	-	-	-	0.006^-	0.001^-	0.000^-	0.025^-	0.024^-	0.021^-	-	-	-
$N_1=10$	-	-	-	0.004^-	0.001^-	0.000^-	0.029^-	0.026^-	0.024^-	0.038^-	0.039^-	0.043
$N_1=25$	-	-	-	0.005^-	0.001^-	0.000^-	0.032^-	0.027^-	0.025^-	0.043	0.046	0.043
$N_1=50$	-	-	-	0.009^-	0.001^-	0.000^-	0.039^-	0.032^-	0.025^-	0.046	0.043	0.047
Panel E: test based on wild bootstrap												
$N_1=5$	0.095^+	0.074^+	0.061^+	0.110^+	0.085^+	0.079^+	0.116^+	0.101^+	0.098^+	-	-	-
$N_1=10$	0.098^+	0.062^+	0.051	0.105^+	0.068^+	0.058	0.108^+	0.075^+	0.072^+	0.121^+	0.086^+	0.078^+
$N_1=25$	0.125^+	0.072^+	0.050	0.124^+	0.077^+	0.056	0.122^+	0.076^+	0.058	0.146^+	0.084^+	0.062^+
$N_1=50$	0.145^+	0.084^+	0.051	0.141^+	0.082^+	0.053	0.138^+	0.088^+	0.054	0.155^+	0.094^+	0.057

Note: This Table presents simulation results using design 1 from Frolich (2004) and Busso et al. (2014). Potential outcomes are normally distributed with mean zero and variance one. Panel A presents rejection rates under the null based on the asymptotic distribution of the matching estimator derived in Abadie and Imbens (2006) (AI). Panel B presents rejection rates under the null for the randomization inference test based on permutations, proposed in Section 4.2 (RI, permutation). Panel C presents rejection rates under the null for the randomization inference test based on sign changes, proposed in Section 4.1 (RI, sign changes). Panel D presents rejection rates under the null for the test based on the robust confidence intervals derived in Rothe (2017). Finally, Panel E presents rejection rates under the null for the test based on wild bootstrap. We include a superscript "+" when rejection rate is greater than 6% and a superscript "−" when rejection rate is lower than 4%. For each combination (N_1, N_0), we run 10,000 simulations.

Table 2: **Test Power - no selection on observable**

	$M=1$			$M=2$			$M=4$			$M=10$		
	$N_0=20$	$N_0=50$	$N_0=500$	$N_0=20$	$N_0=50$	$N_0=500$	$N_0=20$	$N_0=50$	$N_0=500$	$N_0=20$	$N_0=50$	$N_0=500$
	(1)	(2)	(3)	(4)	(5)	(6)						
Panel A: test based on RI, permutation												
$N_1=5$	0.031^-	0.030^-	0.024^-	0.075^+	0.085^+	0.089^+	0.095^+	0.098^+	0.112^+	-	-	-
$N_1=10$	0.088^+	0.106^+	0.118^+	0.116^+	0.133^+	0.149^+	0.133^+	0.148^+	0.181^+	0.155^+	0.169^+	0.208^+
$N_1=25$	0.122^+	0.164^+	0.242^+	0.151^+	0.197^+	0.316^+	0.171^+	0.243^+	0.361^+	0.235^+	0.284^+	0.418^+
$N_1=50$	0.144^+	0.203^+	0.396^+	0.175^+	0.264^+	0.515^+	0.217^+	0.304^+	0.590^+	0.298^+	0.375^+	0.651^+
Panel B: test based on RI, sign changes												
$N_1=5$	0.004^-	0.011^-	0.020^-	0.001^-	0.005^-	0.020^-	0.000^-	0.001^-	0.021^-	-	-	-
$N_1=10$	0.045	0.085^+	0.116^+	0.010^-	0.053	0.125^+	0.000^-	0.014^-	0.141^+	0.000^-	0.000^-	0.113^+
$N_1=25$	0.117^+	0.164^+	0.245^+	0.047	0.169^+	0.300^+	0.000^-	0.076^+	0.327^+	0.000^-	0.000^-	0.333^+
$N_1=50$	0.159^+	0.226^+	0.407^+	0.097^+	0.252^+	0.503^+	0.000^-	0.153^+	0.547^+	0.000^-	0.000^-	0.547^+
Panel C: test based on Rothe (2017)												
$N_1=5$	-	-	-	0.008^-	0.004^-	0.000^-	0.053	0.052	0.055	-	-	-
$N_1=10$	-	-	-	0.010^-	0.004^-	0.000^-	0.084^+	0.095^+	0.113^+	0.133^+	0.144^+	0.178^+
$N_1=25$	-	-	-	0.016^-	0.008^-	0.000^-	0.128^+	0.173^+	0.265^+	0.220^+	0.268^+	0.401^+
$N_1=50$	-	-	-	0.026^-	0.011^-	0.001^-	0.156^+	0.211^+	0.484^+	0.285^+	0.361^+	0.641^+

Note: This Table presents simulation results using design 1 from Frolich (2004) and Busso et al. (2014). Potential outcomes are normally distributed with mean zero and variance one. For the treated observations, we add a treatment effect of 0.4. Panel A presents rejection rates for the randomization inference test based on permutations, proposed in Section 4.2 (RI, permutation). Panel B presents rejection rates for the randomization inference test based on sign changes, proposed in Section 4.1 (RI, sign changes). Panel D presents rejection rates for the test based on the robust confidence intervals derived in Rothe (2017), We include a superscript "+" when rejection rate is greater than 6% and a superscript "−" when rejection rate is lower than 4%. For each combination (N_1, N_0), we run 10,000 simulations.

Table 3: **MC results with selection on observable**

	$M=1$			$M=2$			$M=4$			$M=10$		
	$N_0=20$	$N_0=50$	$N_0=500$	$N_0=20$	$N_0=50$	$N_0=500$	$N_0=20$	$N_0=50$	$N_0=500$	$N_0=20$	$N_0=50$	$N_0=500$
	(1)	(2)	(3)	(4)	(5)	(6)						
Panel A: average $\lvert bias \times 1000\rvert$												
$N_1=5$	25.85	4.82	3.94	32.63	12.43	2.06	57.02	26.10	3.06	-	-	-
$N_1=10$	26.49	10.28	0.90	33.82	14.10	0.02	53.66	25.77	2.76	111.40	50.21	4.48
$N_1=25$	23.20	8.53	0.92	31.57	14.42	1.32	53.54	22.46	2.65	111.08	48.94	4.52
$N_1=50$	24.17	8.91	1.71	35.00	12.89	1.17	54.66	23.15	2.89	112.50	49.98	4.58
Panel B: mean squared error ($\times 1000$)												
$N_1=5$	57.85	48.31	41.84	43.83	35.51	30.70	38.62	30.98	24.91	-	-	-
$N_1=10$	39.38	28.73	20.90	29.49	22.14	15.76	26.97	18.50	13.18	32.76	18.47	11.71
$N_1=25$	28.69	17.29	8.99	22.08	13.69	6.93	19.97	11.29	5.86	25.73	11.44	5.13
$N_1=50$	24.60	13.35	5.23	19.02	10.49	4.04	17.32	8.90	3.31	23.64	9.56	2.96
Panel C: test based on RI, permutation												
$N_1=5$	0.018^-	0.021^-	0.014^-	0.041	0.048	0.047	0.050	0.049	0.046	-	-	-
$N_1=10$	0.047	0.047	0.045	0.049	0.049	0.050	0.063^+	0.050	0.050	0.157^+	0.075^+	0.049
$N_1=25$	0.051	0.048	0.044	0.057	0.053	0.048	0.087^+	0.059	0.051	0.311^+	0.122^+	0.050
$N_1=50$	0.050	0.052	0.049	0.056	0.052	0.048	0.104^+	0.061^+	0.049	0.446^+	0.173^+	0.051
Panel D: test based on RI, sign changes												
$N_1=5$	0.002^-	0.006^-	0.011^-	0.001^-	0.002^-	0.013^-	0.000^-	0.001^-	0.008^-	-	-	-
$N_1=10$	0.024^-	0.040^-	0.045	0.006^-	0.025^-	0.049	0.000^-	0.007^-	0.043	0.000^-	0.000^-	0.035^-
$N_1=25$	0.042	0.048	0.044	0.017^-	0.047	0.047	0.000^-	0.023^-	0.049	0.000^-	0.000^-	0.050
$N_1=50$	0.047	0.050	0.052	0.031^-	0.050	0.052	0.000^-	0.036^-	0.053	0.000^-	0.000^-	0.051
Panel E: test based on Rothe (2017)												
$N_1=5$	-	-	-	0.006^-	0.002^-	0.000^-	0.025^-	0.024^-	0.023^-	-	-	-
$N_1=10$	-	-	-	0.006^-	0.001^-	0.000^-	0.035^-	0.026^-	0.020^-	0.134^+	0.054	0.031^-
$N_1=25$	-	-	-	0.011^-	0.002^-	0.000^-	0.064^+	0.033^-	0.019^-	0.298^+	0.102^+	0.031^-
$N_1=50$	-	-	-	0.016^-	0.003^-	0.000^-	0.085^+	0.041	0.018^-	0.435^+	0.159^+	0.032^-

Note: This Table presents simulation results using design 1 the conditional expectation function 1 from Frolich (2004) and Busso et al. (2014). Panel A reports the average bias (multiplied by 1000), while Panel B reports the mean squared error (multiplied by 1000) of the matching estimator. Panel C presents rejection rates for the randomization inference test based on permutations, proposed in Section 4.2 (RI, permutation). Panel D presents rejection rates for the randomization inference test based on sign changes, proposed in Section 4.1 (RI, sign changes). Panel E presents rejection rates for the test based on the robust confidence intervals derived in Rothe (2017), We include a superscript "+" when rejection rate is greater than 6% and a superscript "−" when rejection rate is lower than 4%. For each combination (N_1, N_0), we run 10,000 simulations.

Table 4: **MC results with selection on observable,** $k = 2$

	$M = 1$			$M = 2$			$M = 4$			$M = 10$		
	$N_0 = 20$	$N_0 = 50$	$N_0 = 500$	$N_0 = 20$	$N_0 = 50$	$N_0 = 500$	$N_0 = 20$	$N_0 = 50$	$N_0 = 500$	$N_0 = 20$	$N_0 = 50$	$N_0 = 500$
	(1)	(2)	(3)	(4)	(5)	(6)						
Panel A: average $\lvert bias \times 1000 \rvert$												
$N_1 = 5$	46.94	24.22	5.02	58.06	36.19	6.07	82.35	51.28	8.69	-	-	-
$N_1 = 10$	45.88	24.42	4.32	63.39	33.73	5.39	87.35	48.20	10.16	133.41	80.30	17.11
$N_1 = 25$	46.75	27.57	3.49	61.31	35.78	7.20	85.74	49.10	8.74	132.30	81.45	17.17
$N_1 = 50$	47.23	27.02	5.62	61.20	35.83	6.54	85.54	50.19	9.21	133.16	81.74	17.83
Panel B: mean squared error ($\times 1000$)												
$N_1 = 5$	53.91	46.00	40.80	43.34	36.12	30.13	41.06	32.55	25.42	-	-	-
$N_1 = 10$	35.31	26.68	20.69	30.12	21.14	15.72	29.69	19.33	13.41	37.66	21.81	11.97
$N_1 = 25$	23.66	15.51	8.74	21.27	12.74	6.79	21.78	11.93	5.74	30.42	15.06	5.16
$N_1 = 50$	19.75	10.96	4.85	18.04	9.61	3.80	19.82	9.70	3.25	28.16	12.87	3.05
Panel C: test based on RI, permutation												
$N_1 = 5$	0.023^-	0.023^-	0.016^-	0.051	0.050	0.045	0.069^+	0.058	0.048	-	-	-
$N_1 = 10$	0.051	0.053	0.045	0.069^+	0.055	0.050	0.112^+	0.068^+	0.050	0.204^+	0.144^+	0.054
$N_1 = 25$	0.063^+	0.057	0.047	0.102^+	0.074^+	0.048	0.212^+	0.126^+	0.048	0.363^+	0.350^+	0.061^+
$N_1 = 50$	0.061^+	0.054	0.049	0.135^+	0.089^+	0.054	0.323^+	0.219^+	0.057	0.482^+	0.570^+	0.097^+
Panel D: test based on RI, sign changes												
$N_1 = 5$	0.003^-	0.009^-	0.017^-	0.000^-	0.003^-	0.012^-	0.000^-	0.001^-	0.012^-	-	-	-
$N_1 = 10$	0.030^-	0.045	0.045	0.005^-	0.024^-	0.045	0.000^-	0.004^-	0.047	0.000^-	0.000^-	0.028^-
$N_1 = 25$	0.055	0.054	0.046	0.016^-	0.053	0.050	0.000^-	0.009^-	0.051	0.000^-	0.000^-	0.046
$N_1 = 50$	0.058	0.051	0.049	0.022^-	0.066^+	0.054	0.000^-	0.015^-	0.054	0.000^-	0.000^-	0.049
Panel E: test based on Rothe (2017)												
$N_1 = 5$	-	-	-	0.007^-	0.003^-	0.001^-	0.039^-	0.031^-	0.024^-	-	-	-
$N_1 = 10$	-	-	-	0.010^-	0.002^-	0.000^-	0.082^+	0.041	0.022^-	0.176^+	0.117^+	0.037^-
$N_1 = 25$	-	-	-	0.023^-	0.003^-	0.000^-	0.184^+	0.094^+	0.022^-	0.355^+	0.337^+	0.044
$N_1 = 50$	-	-	-	0.039^-	0.008^-	0.000^-	0.281^+	0.181^+	0.029^-	0.469^+	0.571^+	0.073^+

Note: This Table replicates the simulations presented in Table 3 with the difference that it considers a matching estimator on X and $\tilde{X}_2$, where $\tilde{X}_2$ is a random variable independent of all other random variables in the model.

Table 5: **MC results with selection on observable,** $k = 4$

	$M = 1$			$M = 2$			$M = 4$			$M = 10$		
	$N_0 = 20$	$N_0 = 50$	$N_0 = 500$	$N_0 = 20$	$N_0 = 50$	$N_0 = 500$	$N_0 = 20$	$N_0 = 50$	$N_0 = 500$	$N_0 = 20$	$N_0 = 50$	$N_0 = 500$
	(1)	(2)	(3)	(4)	(5)	(6)						
Panel A: average $\lvert bias \times 1000 \rvert$												
$N_1 = 5$	82.61	55.64	24.32	97.64	67.89	27.27	115.26	80.56	30.66	-	-	-
$N_1 = 10$	84.43	58.95	21.21	100.01	72.87	26.57	119.87	83.75	30.97	152.37	112.08	46.03
$N_1 = 25$	86.12	60.29	23.10	99.67	70.75	26.91	117.22	87.20	33.27	151.47	112.19	44.97
$N_1 = 50$	81.55	61.42	22.95	97.00	71.69	27.08	114.23	87.78	34.24	149.99	113.39	46.73
Panel B: mean squared error ($\times 1000$)												
$N_1 = 5$	59.36	49.46	41.05	50.67	40.66	31.96	48.91	36.49	26.68	-	-	-
$N_1 = 10$	39.22	28.39	21.24	35.44	25.03	16.31	36.94	23.99	14.21	43.23	27.83	13.70
$N_1 = 25$	26.54	16.71	9.21	26.16	15.53	7.35	28.12	17.04	6.64	35.82	21.06	6.91
$N_1 = 50$	21.87	12.47	5.07	22.62	12.55	4.31	24.96	14.40	4.27	32.56	18.90	4.91
Panel C: test based on RI, permutation												
$N_1 = 5$	0.025^-	0.022^-	0.016^-	0.069^+	0.057	0.048	0.093^+	0.069^+	0.052	-	-	-
$N_1 = 10$	0.069^+	0.055	0.050	0.098^+	0.080^+	0.051	0.160^+	0.109^+	0.058	0.233^+	0.199^+	0.067^+
$N_1 = 25$	0.087^+	0.078^+	0.054	0.184^+	0.133^+	0.061^+	0.301^+	0.261^+	0.076^+	0.387^+	0.448^+	0.125^+
$N_1 = 50$	0.101^+	0.095^+	0.060^+	0.262^+	0.227^+	0.077^+	0.413^+	0.473^+	0.124^+	0.493^+	0.682^+	0.319^+
Panel D: test based on RI, sign changes												
$N_1 = 5$	0.009^-	0.011^-	0.014^-	0.001^-	0.004^-	0.013^-	0.000^-	0.000^-	0.010^-	-	-	-
$N_1 = 10$	0.049	0.055	0.050	0.005^-	0.039^-	0.050	0.000^-	0.001^-	0.049	0.000^-	0.000^-	0.023^-
$N_1 = 25$	0.089^+	0.080^+	0.057	0.014^-	0.085^+	0.060^+	0.000^-	0.003^-	0.065^+	0.000^-	0.000^-	0.043
$N_1 = 50$	0.100^+	0.094^+	0.061^+	0.023^-	0.124^+	0.072^+	0.000^-	0.006^-	0.088^+	0.000^-	0.000^-	0.067^+
Panel E: test based on Rothe (2017)												
$N_1 = 5$	-	-	-	0.009^-	0.002^-	0.000^-	0.062^+	0.043	0.027^-	-	-	-
$N_1 = 10$	-	-	-	0.017^-	0.003^-	0.000^-	0.131^+	0.077^+	0.026^-	0.206^+	0.166^+	0.051
$N_1 = 25$	-	-	-	0.053	0.008^-	0.000^-	0.275^+	0.230^+	0.039^-	0.381^+	0.437^+	0.098^+
$N_1 = 50$	-	-	-	0.093^+	0.031^-	0.000^-	0.381^+	0.441^+	0.079^+	0.484^+	0.690^+	0.289^+

Note: This Table replicates the simulations presented in Table 3 with the difference that it considers a matching estimator on X, $\tilde{X}_2$, $\tilde{X}_3$ and $\tilde{X}_4$.

Table 6: **MC results with selection on observable, bias-corrected estimator**

	$M=1$			$M=2$			$M=4$			$M=10$		
	$N_0=20$	$N_0=50$	$N_0=500$	$N_0=20$	$N_0=50$	$N_0=500$	$N_0=20$	$N_0=50$	$N_0=500$	$N_0=20$	$N_0=50$	$N_0=500$
	(1)	(2)	(3)	(4)	(5)	(6)						
Panel A: average $\|bias \times 1000\|$												
$N_1=5$	16.42	72.55	6.74	58.15	12.61	4.20	13.63	15.07	6.34	-	-	-
$N_1=10$	13.07	9.64	3.38	16.12	14.39	4.34	16.15	15.45	3.02	13.87	15.54	6.12
$N_1=25$	12.94	8.22	3.80	13.56	12.04	2.34	15.70	15.78	4.49	12.35	15.48	6.58
$N_1=50$	15.19	10.61	2.35	14.12	12.93	3.32	18.46	14.64	4.82	11.83	15.45	6.79
Panel B: mean squared error ($\times 1000$)												
$N_1=5$	>100	>100	42.02	>100	>100	30.84	98.57	44.42	26.43	-	-	-
$N_1=10$	76.28	33.95	21.31	>100	29.83	16.26	45.08	22.02	13.32	33.72	19.61	12.00
$N_1=25$	42.55	20.35	9.37	32.71	15.98	7.12	27.39	13.47	5.86	23.35	12.17	5.39
$N_1=50$	38.32	16.73	5.17	27.80	12.77	4.06	23.29	10.90	3.36	20.65	9.52	3.05
Panel C: test based on RI, permutation												
$N_1=5$	0.018^-	0.020^-	0.015^-	0.052	0.052	0.046	0.085^+	0.059	0.053	-	-	-
$N_1=10$	0.052	0.048	0.048	0.066^+	0.056	0.051	0.094^+	0.061^+	0.050	0.142^+	0.080^+	0.053
$N_1=25$	0.062^+	0.056	0.047	0.082^+	0.064^+	0.050	0.111^+	0.068^+	0.051	0.205^+	0.097^+	0.053
$N_1=50$	0.074^+	0.059	0.049	0.082^+	0.064^+	0.054	0.116^+	0.076^+	0.047	0.256^+	0.110^+	0.054
Panel D: test based on RI, sign changes												
$N_1=5$	0.002^-	0.007^-	0.015^-	0.001^-	0.003^-	0.013^-	0.000^-	0.000^-	0.009^-	-	-	-
$N_1=10$	0.025^-	0.032^-	0.047	0.005^-	0.027^-	0.047	0.000^-	0.006^-	0.046	0.000^-	0.000^-	0.037^-
$N_1=25$	0.043	0.049	0.051	0.018^-	0.047	0.050	0.000^-	0.019^-	0.046	0.000^-	0.000^-	0.048
$N_1=50$	0.044	0.048	0.046	0.027^-	0.052	0.054	0.000^-	0.029^-	0.048	0.000^-	0.000^-	0.049

Note: This Table replicates the simulations presented in Table 3 with the difference that it considers a bias-corrected matching estimator suggested in Abadie and Imbens (2011).