Telescope Matching: A Flexible Approach to Estimating Direct Effects*

Matthew Blackwell†
Anton Strezhnev‡

August 4, 2018

Abstract

Estimating the direct effect of a treatment fixing the value of a consequence of that treatment is becoming a common part of social science research. In many cases, however, these effects are difficult to estimate using standard methods since they can induce post-treatment bias. More complicated methods like marginal structural models or structural nested mean models can recover direct effects in these situations but require parametric models for the outcome or the post-treatment covariates. In this paper, we propose a novel two-step matching approach to estimating direct effects that we call *telescope matching*. This method uses matching to impute missing counterfactual outcomes in a flexible manner. We derive the asymptotic properties of this estimator, show that there is a bias that dominates its asymptotic distribution, and develop a bias-corrected estimator that augments matching with regression. Using simulation and empirical studies, we show how this approach weakens model dependence for researchers estimating direct treatment effects.

*Thanks to Alberto Abadie, Paul Kellstedt, Jamie Robins, Jann Spiess, and Yiqing Xu for valuable feedback and discussions. Any remaining errors are our own.

†Department of Government and Institute for Quantitative Social Science, Harvard University, 1737 Cambridge St, MA 02138. web: http://www.mattblackwell.org email: mblackwell@gov.harvard.edu

‡University of Pennsylvania Law School, 3501 Sansom Street, Philadelphia, PA 19104, web: https://www.antonstrezhnev.com/ email: astrezhn@law.upenn.edu

1 Introduction

Researchers in the social sciences often want to estimate the direct effect of a treatment for fixed values of a downstream or mediating variable. These direct effects form the bases of mediation analyses and can help to adjudicate theories for why a total causal effect of treatment exists at all (Imai, Keele and Yamamoto, 2010; VanderWeele, 2015; Acharya, Blackwell and Sen, 2016). These quantities of interest may also detect heterogeneous effects that can help to design more effective treatments in the future. Direct effects are also of interest in time-series and time-series cross-sectional settings, where the lagged effect of past treatment fixing contemporaneous treatment is commonly estimated (Blackwell and Glynn, 2018). Often, these effects are estimated under the assumption that (a) treatment assignment is unconfounded conditional on baseline covariates, and (b) the mediator is unconfounded conditional on baseline covariates, treatment status, and intermediate covariates potentially affected by treatment. Several parametric and semi-parametric methods have been developed for estimating these quantities, including the parametric g-formula, structural nested models, and marginal structural models (Robins, 1986, 1997; Richardson and Rotnitzky, 2014).

Unfortunately, these extant approaches to estimating direct effects require the (correct) specification of several models. In particular, analysts must be able to estimate both effect of the treatment and the effect of the mediator. Thus, this requires at least two models to estimate effects and the decision of how to control for both the baseline and intermediate covariates. Even if treatment is randomly assigned so that there are no baseline confounders, the mediator is often simply observed without any intervention and requires intermediate covariates to block confounding in the estimated effects. Furthermore, the performance of the resulting estimators depend on choosing the correct model specification across all of these models. Researchers interested in these effects thus can face a daunting number of modeling choices to estimate these effects.

In this paper, we present a new method for estimating direct effects based on a matching approach. Matching is a popular strategy for reducing model dependence and for estimating average treatment effects, but it has seen limited application to dynamic settings like this (Ho et al., 2006; Abadie and Imbens, 2006). We develop a two-stage matching procedure which first creates create

imputations of counterfactual outcomes for fixed values of the mediating variable using nonparametric matching. We then use these imputations to carry out a second stage of matching in order to estimate the direct effect of treatment holding constant the mediating variable. This two-stage approach, which we call "telescope matching," adjusts for both pre- and post-treatment confounders in the first stage, "telescoping out" to only the pre-treatment confounders in the second stage. Our method allows us to sidestep modeling assumptions on either the relationship between the outcome and covariates or between the treatment and covariates by weakening dependence on strong parametric models.

We derive the large-sample properties of this matching approach and show that it is consistent for the controlled direct effect. However, as in the case with single-shot matching (Abadie and Imbens, 2006), we derive bias terms that prevent the estimator from converging to a stable asymptotic distribution. To help avoid this issue, we develop a bias-correction method that uses regression estimators in a similar manner to Abadie and Imbens (2011). We find that this combination of matching and regression strikes an appropriate balance between robustness and efficiency. While our method is not as efficient when the true regression models are known, we show through a simulation study that it can be more robust when the regression models are misspecified.

This paper proceeds as follows. First, we describe the relevant quantities of interest, including the controlled direct effect, and the assumptions necessary to identify these effects. We then define our telescope matching approach to estimating these direct effects, discuss its large-sample properties, and describe the bias-correction approach. Then, we discuss variance estimation using the weighted bootstrap of Otsu and Rai (2017) and compare our approach to others. Next, we conduct a simulation study that shows how these various estimators perform when a researcher has a correct and incorrect specification of the outcome regression model. Finally, we demonstrate the method in an empirical setting of an experimental study of the effect of media frames on attitudes towards immigration as mediated by emotional response (Brader, Valentino and Suhay, 2008).

2 Proposed method

2.1 Notation and assumptions

Let $A_i \in \{0,1\}$ denote values of treatment for unit i. Let $M_i \in \{0,1\}$ denote the value of a post-treatment variable that we seek to hold constant. This may be the same treatment administered a future date or it might be some consequence of the treatment that the analyst believes is part of a causal mechanism. For brevity, we call this variable the mediator since it is post-treatment and may mediate some or all of the effect of treatment. The goal of the analysis is to estimate the effect of treatment on some outcome, Y_i. We define potential outcomes for this under the various combinations of treatment and mediator, $Y_i(a,m)$ (Rubin, 1974; Robins, 1986). For instance, $Y_i(1,1)$ represents the outcome we would see if unit i had been assigned $A_i = 1$ and $M_i = 1$. We make the usual consistency assumption, $Y_i = Y_i(a,m)$ if $A_i = a$ and $M_i = m$, which states that the observed outcome for unit i is the potential outcome for that unit at its observed level of A_i and M_i. Note that, because M_i can be affected by A_i, it too has potential outcomes, $M_i(a)$, that also follow a consistency assumption, $M_i = A_i M_i(1) + (1 - A_i)M_i(0)$.

We define two sets of relevant covariates: baseline and intermediate. The baseline covariates, X_i, are causally prior to both A_i and M_i. Thus, researchers can adjust for these covariates using typical causal inference techniques such as regression, weighting, or matching. The intermediate covariates, Z_i, can be affected by A_i, but are causally prior to M_i and confound the outcome-mediator relationship. These covariates pose problems for standard models when trying to estimate the effect of the treatment and the mediator at the same time due to the potential for post-treatment bias induced by conditioning on them (Rosenbaum, 1984).

Our goal in this paper is to estimate the controlled direct effect for unit i fixing the mediator M_i to a particular level (Robins and Greenland, 1992). For a given individual, this can be defined in terms of the counterfactual outcomes:

$$\tau_i = Y_i(1,0) - Y_i(0,0).$$

This is the effect of moving from control to treatment when we force i to have $M_i = 0$. Because it

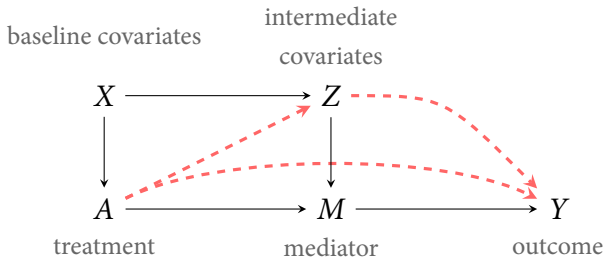

Figure 1: Directed acyclic graph showing the causal relationships in the present setting. Dashed red lines represent the controlled direct effect of the treatment not through the mediator. Unobserved errors are omitted.

is difficult to estimate the individual-level effects without very strong assumptions, we focus on the average controlled direct effect (ACDE), defined as:

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

This quantity represents the average effect of treatment when a potentially post-treatment mediator is fixed at a particular value. For this paper, we focus on the ACDE when setting $M_i = 0$, but it is straightforward to extend the discussion and the method to investigate the controlled direct effect at other levels of M_i. We can also define the conditional ACDE:

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

This is the direct effect of treatment within levels of the covariates. We can recover the ACDE from the conditional effects by averaging over the distribution of the data: $\tau = \mathbb{E}[\tau(X_i)]$.

The direct effect stands in contrast to the overall average treatment effect (ATE), which is the difference in average potential outcomes when we just manipulate treatment:

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

This quantity is the "total" effect of treatment, including both its direct effect and any effects through the mediator. Previous work has focused on the difference between the ATE and the ACDE as evidence for M_i playing a role in the causal mechanism of A_i (Acharya, Blackwell and Sen, 2016, 2018).

Throughout, we assume that $\{(Y_i, M_i, Z_i, A_i, X_i)\}_{i=1}^N$ are independent and identically distributed. We let N_{am} be the number of units with $A_i = a$ and $M_i = m$, with N_a being the marginal number

of units with $A_i = a$. We make the following sequential ignorability assumption about the treatment and mediator:

Assumption 1 (Sequential Ignorability). *For every value, a, m, x, z:*

$$\{Y_i(a,m), M_i(a), Z_i(a)\} \perp\!\!\!\perp A_i | X_i = x \tag{1}$$

$$Y_i(a,m) \perp\!\!\!\perp M_i | X_i = x, Z_i = z, A_i = a \tag{2}$$

The first part of this assumption states that the treatment is independent of the potential outcome and the potential values of the mediator, conditional on baseline covariates. The second part states that the mediator is independent of the potential outcomes, conditional on the treatment and the baseline and intermediate covariates. This assumption essentially requires two "selection-on-observables" conditions, one for the treatment and one for the mediator. Thus, there must be no unmeasured confounders for the treatment-outcome relationship after conditioning on X_i and no unmeasured confounders for the mediator-outcome relationship after conditioning on $\{X_i, A_i, Z_i\}$.

We further assume that the distributions of the treatment and mediator are not degenerate at any values of the covariates.

Assumption 2 (Positivity). *For every value, a, x, z, and for some values $\eta > 0$ and $\nu > 0$:*

$$\eta < P(A_i = 1 | X_i = x) < 1 - \eta \tag{3}$$

$$\nu < P(M_i = 1 | X_i = x, Z_i = z, A_i = a) < 1 - \nu \tag{4}$$

The first part of this assumption requires that the treated and control distributions of the baseline covariates have the same support. The second part extends this assumption to the $M_i = 1$ and $M_i = 0$ covariate distributions.

A few other pieces of notation will be useful. First, we define a series of conditional expectation functions (CEF) of the potential outcomes, conditional on different sets of covariates. In particular, we define $\mu_{am}(x,z,a) = E[Y(a,m)|X_i = x, Z_i = z, A_i = a]$ and $\mu_{am}(x,a) = E[Y_i(a,m)|X_i = x, A_i = a]$. Let $\mu(x,z,a,m) = E[Y_i|X_i = x, Z_i = z, A_i = a, M_i = m]$ be the CEF of the observed outcome, noting that under Assumption 1, $\mu_{am}(x,z,a) = \mu(x,z,a,m)$ We also define two types of residuals,

$\varepsilon_i = Y_i - \mu(X_i, Z_i, A_i, M_i)$ and $\eta_i = \mu_{A_i 0}(X_i, Z_i, A_i) - \mu_{A_i 0}(X_i, A_i)$. The first is the CEF error for Y_i and the second captures the variation in the CEF of the potential outcomes that is due to Z_i. Given these definitions, we have $E[\eta_i|\mathbf{X}, \mathbf{A}] = 0$ and $E[\varepsilon_i|\mathbf{X}, \mathbf{Z}, \mathbf{A}, \mathbf{M}] = 0$, where $\mathbf{X}$ and $\mathbf{Z}$ are the entire $N \times k_x$ and $N \times k_z$ matrices of pretreatment and intermediate covariates, and $\mathbf{A}$ and $\mathbf{M}$ are the N vectors of the treatment and mediator.

2.2 Simple two-stage matching procedure

How can we estimate the ACDE? If we simply were to compare the average outcome across levels of the treatment, we will be combining the direct effect and any effect due to changes in the mediator. Instead, we would like to compare the potential outcomes for fixed values of the mediator. To do this, several approaches have been put forward. If both A_i and M_i are randomized, then standard tools for multileveled treatments can be used to estimate the direct effect of treatment since there are no covariates for which to adjust. When there are only baseline confounders, then standard selection-on-observable methods for multi-leveled treatments can be applied (Imbens, 2004). However, when there are post-treatment confounders for the relationship between M_i and Y_i, we must turn to other methods to adjust for this form of confounding.

Our proposed approach, which we call *telescope matching*, imputes values of the missing potential outcomes in a flexible manner. For any particular unit, we only observe one of four possible potential outcomes, an issue sometimes called the fundamental problem of causal inference. To estimate the ACDE when $M_i = 0$, we would like to observe values for $Y_i(1, 0)$ and $Y_i(0, 0)$ for all units. The goal of telescope matching is to use matching methods in order to obtain reasonable imputations of these values for all units.

Let $V_i = (X_i, Z_i)$ be the vector of covariates, both baseline and intermediate. The first step of telescope matching is to match each unit with $M_i = 1$ to some number of units with $M_i = 0$ that have similar values of covariates V_i and identical treatment status A_i. We follow Abadie and Imbens (2006) in much of our discussion of matching estimators. Given a particular distance metric on the support of V_i (such as the Euclidean norm or the Mahalanobis distance) and given a particular unit

with $M_i = 1$, we choose L units, here indexed by j, that are the closest to i in terms of covariate distance that have $M_j = 0$. Let $\mathcal{J}_L^m(i)$ denote this set of units that are matched to some unit i with $M_i = 1$. Matching is done with replacement so a control unit might be matched to multiple treated units and we let $K_L^m(i) = \sum_{j=1}^{N} \mathbb{I}\{i \in \mathcal{J}_L^m(j)\}$ be the number of times that unit i is used as a match in stage, where $\mathbb{I}\{\cdot\}$ is an indicator function. As in Abadie and Imbens (2006), this quantity is important to the asymptotic distribution of the matching estimator.

Typically, matching would be used to estimate the effect of M_i on Y_i, but here we are actually more interested in obtaining a good estimate of the potential outcome under $M_i = 0$ for all units, including those observed to have $M_i = 1$. To do this, we define the following imputation:

$$\widehat{Y}_{i0} = \begin{cases} Y_i & \text{if } M_i = 0 \\ \frac{1}{L} \sum_{j \in \mathcal{J}_L^m(i)} Y_j & \text{if } M_i = 1 \end{cases}$$

For units observed with $M_i = 0$, consistency grants that the observed outcome equals that unit's potential outcome under assignment to $M_i = 0$. However, for units with $M_i = 1$, we need to impute the missing counterfactual outcome under $M_i = 0$. We do so by averaging the outcome among those units with $M_i = 0$ which were matched to unit i. These units have identical treatment levels A_i and are the closest to i (for a given distance metric) on both the baseline and intermediate covariates.

Once we impute the potential outcomes for all observations in the sample, we can complete the method by taking these imputations as the outcome in a second matching stage in which we attempt to estimate the effect of A_i adjusting for confounding by the baseline covariates, X_i. That is, we use the same matching technique as in the first stage to generate matches for each unit i with respect to treatment status. In particular, we match each unit to L units of the opposite treatment status with similar values of the baseline covariates.[1] Let $\mathcal{J}_L^a(i)$ be the indices of the units matching to treated unit i such that $A_j = 1 - A_i$ for all $j \in \mathcal{J}_L^a(i)$. We define the number of times i is used as a match in the A_i stage as $K_L^a(i) = \sum_{j=1}^{N} \mathbb{I}\{i \in \mathcal{J}_L^a(j)\}$. Note that here, as opposed to in the first stage, we are matching treated to control and control to treated to ensure that we can estimate the overall ACDE

[1] We could allow for different matching ratios in the two stages, but for simplicity, we focus on the case where the ratios are the same across the two stages.

(rather than the ACDE conditional on the treated). We can then define the following imputed values of the potential outcomes under treatment:

$$\widehat{Y}_{i10} = \begin{cases} \widehat{Y}_{i0} & \text{if } A_i = 1 \\ \frac{1}{L} \sum_{j \in \mathcal{J}_L^a(i)} \widehat{Y}_{j0} & \text{if } A_i = 0 \end{cases}$$

We can also define similar values for the potential outcomes under control:

$$\widehat{Y}_{i00} = \begin{cases} \frac{1}{L} \sum_{j \in \mathcal{J}_L^a(i)} \widehat{Y}_{j0} & \text{if } A_i = 1 \\ \widehat{Y}_{i0} & \text{if } A_i = 0 \end{cases}$$

With these definitions in hand, we can then apply a standard difference in means matching estimator. In particular, the simple two-stage matching estimate of the ACDE then becomes,

$$\widehat{\tau} \equiv \frac{1}{N} \sum_{i=1}^{N} \left(\widehat{Y}_{i10} - \widehat{Y}_{i00} \right).$$

Note that this second-stage matching could also be used to estimate the overall ATE as well.

Both $K_L^m(i)$ and $K_L^a(i)$ tell us how much unit i is contributing to the overall estimate through being matched in the first and second stages, respectively. Of course, units with $M_i = 0$ might also contribute *indirectly* if they are matched to a $M_i = 1$ unit in the first stage and that $M_i = 1$ unit is used as a match in the second stage. To account for such indirect contributions of a unit, let $K_L^{am}(i) = \sum_{j=1}^{N} \mathbb{I}\{i \in \mathcal{J}_L^m(j)\} K_L^a(j)$ be the number of times a first-stage match with $M_i = 0$ is implicitly used as a match in the second stage because the unit to which it was matched is selected as a match in the second stage.

2.3 Bias and consistency

To understand the benefits and potential drawbacks of such a matching approach, we investigate the large-sample properties of this estimator. Abadie and Imbens (2006) showed that in the context of estimating the overall ATE, the equivalent simple matching procedure was biased due to imperfect matches. Further, they showed that with a fixed size for the matched set, L, this bias converges to 0 as

the sample size increased, but at a rate slow enough to affect the asymptotic normality of the matching estimator. In this section, we show that a similar account holds in the present setting.

We can decompose the estimation error of $\widehat{\tau}$ as follows:

$$\widehat{\tau} - \tau = \left(\frac{1}{N} \sum_{i=1}^{N} \tau(X_i) - \tau \right) + E_L^m + E_L^a + B_L^m + B_L^a \tag{5}$$

The first term in the decomposition, $(1/N) \sum_{i=1}^{N} \tau(X_i) - \tau$, is the difference between the sample average of the conditional ACDEs and the true ACDE, which converges to 0 under a standard law of large numbers. Next in the decomposition are two weighted sums of the residuals:

$$E_L^m = \frac{1}{N} \sum_{i=1}^{N} (2A_i - 1) \left(1 + \frac{K_L^a(i)}{L} + \frac{K_L^m(i)}{L} + \frac{K_L^{am}(i)}{L^2} \right) (1 - M_i) \varepsilon_i$$

$$E_L^a = \frac{1}{N} \sum_{i=1}^{N} (2A_i - 1) \left(1 + \frac{K_L^a(i)}{L} \right) \eta_i$$

The error due to the first-stage is mean-zero conditional on all variables, $\mathbb{E}[E_L^m | \mathbf{X}, \mathbf{A}, \mathbf{Z}, \mathbf{M}] = 0$, and the error due to the second-stage is mean-zero conditional on the baseline covariates and the treatment, $\mathbb{E}[E_L^a | \mathbf{X}, \mathbf{A}] = 0$. Thus, the first three terms impose no bias on the matching estimator.

Finally, the last two terms capture the bias of the matching procedure due to the first and second-stages of matching:

$$B_L^m = \frac{1}{N} \sum_{i=1}^{N} (2A_i - 1) \left(1 + \frac{K_L^a(i)}{L} \right) M_i \left(\frac{1}{L} \sum_{\ell \in \mathcal{J}_L^m(i)} \mu_{A_i 0}(X_\ell, Z_\ell, A_i) - \mu_{A_i 0}(X_i, Z_i, A_i) \right)$$

$$B_L^a = \frac{1}{N} \sum_{i=1}^{N} (2A_i - 1) \left[\frac{1}{L} \sum_{j \in \mathcal{J}_L^a(i)} \mu_{1-A_i,0}(X_i, 1 - A_i) - \mu_{1-A_i,0}(X_j, 1 - A_i) \right]$$

These bias terms reflect the matching discrepancy at each stage of matching. For instance, the last term in the definition of B_L^m is the difference in the expectation of the outcome for the covariates for unit i and for the units matched to i. If the matches were perfect, then we would have $X_i = X_\ell$ and $Z_i = Z_\ell$ for all $\ell \in \mathcal{J}_L^m$ and $X_i = X_j$ for all $j \in \mathcal{J}_L^a(i)$, and both of these bias terms would be equal to 0. In general, however, matches are imperfect when we have any continuous covariates and so these bias terms will not be mean-zero (Abadie and Imbens, 2006). Importantly for the results below, though, these values do converge to 0 as N increases.

With this in hand, we can state a few asymptotic features of the simple matching estimator. In these results, we invoke a series of regularity conditions that we describe in Assumption 3 in the Appendix. Briefly, these conditions impose smoothness on conditional expectations and variances as functions of the covariates and ensure that sufficient moments of the outcome exist to allow for convergence in distribution. First, we note that even though the simple matching estimator is biased, it is consistent for the true ACDE.

Theorem 1. *Suppose that Assumptions 1, 2, and 3 hold. Then, (i)* $\widehat{\tau} - \tau \xrightarrow{p} 0$ *and (ii)*

$$\sqrt{N}\left(\widehat{\tau} - B_L^m - B_L^a\right) \xrightarrow{d} N(0, \sigma^2),$$

where

$$\begin{aligned}\sigma^2 = {} & \mathbb{E}[(\tau(X_i) - \tau)^2] \\ & + \mathbb{E}\left[\left(1 + \frac{K_L^a(i)}{L}\right)^2 \sigma_\eta^2(X_i, A_i)\right] \\ & + \mathbb{E}\left[(1 - M_i)\left(1 + \frac{K_L^a(i)}{L} + \frac{K_L^m(i)}{L} + \frac{K_L^{am}(i)}{L^2}\right)^2 \sigma^2(X_i, Z_i, A_i, 0)\right].\end{aligned}$$

The crux of part (i) of this result comes from the fact that the terms in the decomposition in equation 5 all converge to 0 in probability. Unfortunately, in the present setting, as in Abadie and Imbens (2006), the bias terms dominate the distribution of the estimator as $N \to \infty$, so that the simple matching estimator will not converge in distribution at the root-N rate. The second part of this theorem shows that when the bias terms are removed, the matching estimator is asymptotically normal with a variance that depends on the distribution of the number of times a unit is used as a match. Even though these results ignore the bias terms, they are still useful because the bias correction that we describe next will converge at a fast enough rate so it can be ignored asymptotically (Abadie and Imbens, 2011).

2.4 Bias correction

As described above, the bias of the simple matching procedure tends to 0 as N increases, but the rate is slow enough to dominate the asymptotic distribution of the estimator. Due to this, Abadie

and Imbens (2011) proposed a bias-corrected estimator that estimates and removes the bias from a simple matching procedure. In this section, we extend this idea to the present two-stage setting. In particular, we propose estimating the two bias terms with regression estimators of the two relevant CEFs, $\widehat{\mu}(x,z,a,m)$ and $\widehat{\mu}_{a0}(x,a)$. We can then use these regressions to obtain estimates of the bias terms themselves:

$$\widehat{B}_L^m = \frac{1}{N}\sum_{i=1}^{N}(2A_i-1)\left(1+\frac{K_L^a(i)}{L}\right)M_i\left(\frac{1}{L}\sum_{j\in J_L^m(i)}\widehat{\mu}(X_j,Z_j,A_i,0)-\widehat{\mu}(X_i,A_i,Z_i,0)\right)$$

$$\widehat{B}_L^a = \frac{1}{N}\sum_{i=1}^{N}(2A_i-1)\left[\frac{1}{L}\sum_{j\in J_L^a(i)}\widehat{\mu}_{1-A_i,0}(X_i,1-A_i)-\widehat{\mu}_{1-A_i,0}(X_j,1-A_i)\right]$$

If the regression estimators are consistent for their respective CEFs, then $\widehat{B}_L^m$ and $\widehat{B}_L^a$ converge in probability to the bias terms B_L^m and B_L^a, respectively.

To implement this procedure, we must regress Y_i on X_i and Z_i within levels of A_i and M_i to estimate $\widehat{\mu}(X_i,Z_i,A_i,0)$. For $\widehat{\mu}_{A_i0}(X_i,A_i)$, we use this first-stage regression and the first-stage matching to create a bias-corrected imputation of the missing potential outcome, $Y_i(A_i,0)$:

$$\widetilde{Y}_{i0} = \begin{cases} Y_i & \text{if } M_i = 0 \\ \frac{1}{L}\sum_{j\in\mathcal{J}_L^m(i)}\left(Y_j+\widehat{\mu}(X_i,Z_i,A_i,0)-\widehat{\mu}(X_j,Z_j,A_j,0)\right) & \text{if } M_i = 1 \end{cases}$$

We then treat that imputation as the dependent variable and regress it onto the baseline covariates. Abadie and Imbens (2011) proposed series estimators to consistently estimate these CEFs without knowledge of the exact functional form for how the covariates relate to the outcomes. In our simulations and empirical examples below, we use a simple linear, additive regression model, which is a type of series estimator. Other, more flexible estimation methods like generalized additive models may give better performance.

With these estimates in hand, we define the following bias-corrected estimator:

$$\widetilde{\tau} = \widehat{\tau} - \widehat{B}_L^m - \widehat{B}_L^a. \tag{6}$$

In the Appendix, we describe the conditions under which the bias-corrected estimator and the simple-matching estimator both converge to the same distribution. As described by Abadie and Imbens

(2011) and Otsu and Rai (2017), this will occur when the true CEFs are sufficiently smooth and the regression estimators sufficiently flexible to estimate the bias terms at a rate faster than root-N.

2.5 Weighted bootstrap

Conducting inference using telescope matching requires a valid method for estimating standard errors. While the bootstrap is a popular approach for many methods, it is well known that conventional non-parametric bootstrapping, resampling observations $\{Y_i, X_i, Z_i, A_i, M_i\}$, is invalid for matching estimators (Abadie and Imbens, 2008). This is due to the inability of the naive boostrap to preserve the distributions of $K_m^a(i)$, the counts of the number of times unit i is used as a match, across resamples. In the case of our proposed estimator, the same issue persists for the other match counts: $K_L^m(i)$ and $K_L^{am}(i)$.

Recently, Otsu and Rai (2017) proposed a method for using a variety of bootstrap techniques in the matching setting. They show that when the bias-corrected matching estimator is written in a linearized form such that $\widetilde{\tau} = \sum_{i=1}^N \widetilde{\tau}_i$ where $\widetilde{\tau}_i$ consists only of functions of observation i, one could use a weighted bootstrap of the residuals, $\widetilde{\tau}_i - \widetilde{\tau}$, to obtain valid confidence intervals for matching estimators. This "weighted" boostrap resamples the ith contribution to the overall estimate rather than resampling units and matching again in the resampled units. We show in the Appendix that it is possible to write the contribution of the ith observation to our bias-corrected estimator as:

$$\begin{aligned}\widetilde{\tau}_i = (2A_i - 1)\Bigg[&(1 - M_i)\left(1 + \frac{K_L^a(i)}{L} + \frac{K_L^m(i)}{L} + \frac{K_L^{am}(i)}{L^2}\right) Y_i \\ &- \left((1 - M_i)\left(\frac{K_L^m(i)}{L} + \frac{K_L^{am}(i)}{L^2}\right) - M_i\left(1 + \frac{K_L^a(i)}{L}\right)\right)\widehat{\mu}_{A_i,0}(X_i, A_i, Z_i) \\ &- \left(\widehat{\mu}_{1-A_i,0}(X_i, 1 - A_i) + \frac{K_L^a(i)}{L}\widehat{\mu}_{A_i,0}(X_i, A_i)\right)\Bigg] \quad (7)\end{aligned}$$

Otsu and Rai (2017) shows that one can bootstrap the sampling distribution, $(\widetilde{\tau} - \tau)$ with

$$T^* = \frac{1}{N}\sum_{i=1}^N W_i^*(\widetilde{\tau}_i - \widetilde{\tau}),$$

where W_i^* are random variables that satisfy a few basic regularity conditions.[2] For our purposes, we use the nonparametric bootstrap (Efron, 1979), which draws the $\{W_i^*\}$ from a multinomial distribution with N draws of N categories of equal probability. As discussed by Otsu and Rai (2017), this approach avoids the issues with the naive row-resampling bootstrap method that is commonly used and analyzed by Abadie and Imbens (2008) in the matching context. In particular, the weighted bootstrap ensures that the distribution of the $K_L^a(i)$, $K_L^m(i)$, and $K_L^{am}(i)$ are preserved across resamples. In Section 3 we show the performance of the weighted bootstrap in constructing confidence intervals.

2.6 Relationship to other approaches

Controlled direct effects have been the focus of a great deal of statistical and empirical studies over the last few decades. As pointed out by Robins (1986) and Rosenbaum (1984), these direct effects are not identified from standard approaches that condition on M_i and Z_i due to the potential post-treatment bias, which is sometimes called collider bias. This has been interpreted as preventing the application of standard regression or matching estimators to this setting. Instead, estimation of these effects has focused on three general approaches: (1) parametric g-estimation, (2) structural nested models, and (3) inverse probability of treatment weighting (IPTW). Parametric g-estimation exploits the fact that the mean of the potential outcomes can be identified by integrating over the distribution of the post-treatment covariates, Z_i. This integration, though, requires parametric models for the outcome and for joint distribution of the covariates, which can be very demanding when there are more than a handful of covariates.

Structural nested mean models (SNMMs) focus on modeling the effects of both the treatment and the mediator, and their implementation requires models for the outcome conditional on the covariates or the treatment and mediator conditional on the covariates. Sequential ignorability implies a particular estimating equation approach that can be used to estimate the controlled direct effect, but these models can also be seen in terms of an imputation approach. In particular, a SNMM captures

[2] See Appendix A of Otsu and Rai (2017) for more details on the regularity conditions for the weighted bootstrap.

the effect of the mediator in "blip-down" or "demediation" function:

$$\gamma_m(x, z, a) = E[Y_i(a, 1) - Y_i(a, 0)|X_i = x, A_i = a, Z_i = z, M_i = 1]$$

In the binary context, these functions modeling the (conditional) average treatment effect on the treated for both A_i and M_i. These functions are parameterized in terms of the covariates and levels of treatment, a, which can permit effect modification for these effects. For example, we might have:

$$\gamma_m(x, z, a; \beta) = \beta_0 + \beta_1 a.$$

Here, the effect of the mediator would be β_0 when $A_i = 0$ and $\beta_0 + \beta_1$ when $A_i = 1$. Most approaches to SNMMs can be seen as using the following imputation for the missing potential outcomes under $M_i = 0$:

$$\overline{Y}^*_{i0} = \begin{cases} Y_i & \text{if } M_i = 0 \\ Y_i - \widehat{\gamma}_m(X_i, Z_i, A_i) & \text{if } M_i = 1 \end{cases}$$

Thus, SNMMs rely on a consistent estimate of the (possibly heterogeneous) causal effects of M_i. Of course, this estimation will typically rely on a regression model such as the one used in the bias correction above. When linear regression models are used to estimate the parameters of the γ function, this approach has been called *sequential g-estimation*. Below, we compare telescope matching to this approach in our simulations.

Inverse probability of treatment weighting represents the third popular approach to estimating direct effects. In the current setting, it would require consistent estimates of the probability of the mediator given the treatment and intermediate and baseline covariates, $e_m(x, z, a) = \Pr(M_i = 1|X_i = x, Z_i = z, A_i = a)$, and the probability of treatment given the baseline covariates, $e_a(x) = \Pr(A_i = 1|X_i = x)$. Weights within levels of A_i for each unit can then be constructed as $W_i(a) = \frac{M_i}{e_m(X_i,Z_i,a)} + \frac{1-M_i}{1-e_m(X_i,Z_i,a)}$, so that the weight for unit i is the inverse of the probability of receiving the level of M_i it actually had. If we do have consistent estimates of these functions, weighting the data by the above weights will remove the confounding (due to X_i and Z_i) of the relationship between M_i and Y_i. This allows M_i to simply be included as an additional control in any matching, weighting, or regression

approach to estimating the (direct) effect of A_i on Y_i. Unfortunately, in practice, IPTW can have poor performance due to unstable weights when the probability of $M_i = 1$ is close to 0 or 1.[3]

Finally, a host of *doubly robust* methods have been developed that combine features of the SNMM and weighting approaches. These methods require models for both (a) the outcome-covariate relationship and (b) the mediator- and treatment-covariate relationships. These methods are doubly robust in the sense that they are consistent for direct effects when either (a) or (b) are correctly specified. With telescope matching on the other hand, we model the outcome then use matching to help guard against misspecification.

3 Simulation results

We evaluate the performance of telescope matching against existing SNMM methods—specifically, sequential g-estimation—using a simulation in which we artificially introduce model misspecification. We show that while the performance of telescope matching remains comparable to sequential g-estimation when the true model is known, it is much more robust when the outcome models are incorrectly specified. This simulation follows an approach similar to that of Kang and Schafer (2007) which considered the robustness of "doubly robust" estimators in situations where the functional form of the outcome-confounder relationship was not known.

Our assumed data generating process, which reflects a common situation encountered by researchers, is as follows. We assume a binary treatment, denoted A_i and a binary mediator M_i. We have two observed pre-treatment confounders, X_{i1} and X_{i2}, and one post-treatment confounder Z_{i1}. The true X_{i1} and X_{i2} are assumed to both be random normal variates with mean 0 and variance 1.

The probability unit i receives treatment has a logistic functional form:

$$Pr(A_i = 1|X_{i1}, X_{i2}) = \frac{1}{1 + \exp(-(-X_{i1} + .5X_{i2}))}$$

The post-treatment confounder, Z_{i1}, is a function of treatment and of another, unobserved, confounding factor affecting both Z_{i1} and outcome Y_i. Therefore, while Z_{i1} is causally affected by treatment A_i,

[3] For example, in results not shown here, the IPTW approach does quite poorly in the simulation study below when the variables are misspecified. This result is generally consistent with Kang and Schafer (2007).

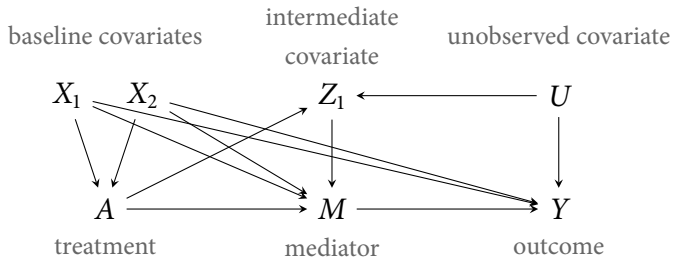

Figure 2: Directed acyclic graph showing the structure of the simulation.

it itself does not directly affect Y_i. Rather, it is a control variable that can block confounding due to the unobserved common cause, denoted U_i. Our specific form for Z_{i1} is chosen as:

$$Z_{i1} = .5A_i + \gamma_i + \delta U_i$$

where $\gamma_i \sim \text{Normal}(0, 1)$, $U_i \sim \text{Normal}(0, 0.2)$. The parameter δ captures the amount of confounding between the intermediate covariate and the outcome and is a parameter that we vary in our simulations. The stronger this confounding, the larger the post-treatment bias for the ACDE when conditioning on Z_{i1} in a naive manner.

The probability that the mediator equals 1 follows a logistic form and is a function of treatment and all confounders:

$$Pr(M_i = 1|A_i, X_{i1}, X_{i2}, Z_{i1}) = \frac{1}{1 + \exp(-(X_{i1} - .75X_{i2} + .5A_{i1} + .75Z_{i1}))}$$

Finally, the observed outcome Y_i is simulated as

$$Y_i = 210 + 27.4 * M_i + 13.7X_{i1} + 13.7X_{i2} + \delta U_i + \varepsilon_i$$

where $\varepsilon_i \sim \text{Normal}(0, 1)$ and δ is the same parameter that appears in the functional form of Z_{i1}. In this case, the effect of A_i flows entirely through its effect on the mediator. Holding the mediator constant, the true controlled direct effect is 0. Figure 2 illustrates the simulation structure in a directed acyclic graph.

As in Kang and Schafer (2007), we simulate model misspecification by considering a scenario where the confounders are not measured directly but rather as the non-linear transformations X^*_{i1},

X^*_{i2}, and Z^*_{i1}:

$$X^*_{i1} = \exp(X_{i1}/2)$$
$$X^*_{i2} = 1/(1 + \exp(X_{i2})) + 10$$
$$Z^*_{i1} = (Z_{i1}/25 + .6)^3$$

Were these non-linear transformations known to the researcher, it would be possible to specify the true linear regression model in terms of a correct transformation of the confounders. However, in practice, researchers do not know the exact non-linear transformation that would yield a correctly specified model. Instead, they will typically use models that simply assume linearity and additivity.

Our simulation varies two parameters: sample size and the magnitude of post-treatment confounding (δ). For each simulated dataset, we estimate the controlled direct effect using (1) a naive additive regression estimate that conditions on the mediator and all pre- and post-treatment confounders, (2) a sequential g-estimation approach that assumes the outcome model is linear and additive in all variables, and (3) our telescope matching approach that makes the same assumptions about the regression model as in sequential g-estimation, but uses Mahalanobis distance matching and bias correction using the same regression modeling as with the sequential g-estimation approach. In our simulations, we set L, the number of units matched to each treated unit ($M_i = 1$), to 3.[4]

Figure 3 plots the absolute value of the bias and the root mean squared error (RMSE) for all three approaches under correct and incorrect model specifications. On the x-axis, we vary the size of post-treatment confounding, which is measured as the partial correlation between U_i and Y_i. For each combination of parameter values, we carried out 10000 iterations of our simulation. We find that both telescope matching and sequential g-estimation are unbiased when the model is correctly specified, with sequential g-estimation having a slight advantage over matching in terms of variance—a gap that decreases significantly as sample size increases. At the larger sample sizes, the increase in variance resulting from including a more flexible imputation model is rather minimal. As ex-

[4] We also considered a fourth method which we discussed earlier in the paper, inverse propensity of treatment weighting (IPTW). While it performs reasonably well in large samples given correct model specifications for both the mediator and outcome, we omit it from the graphs for expository reasons as the bias under misspecification is far larger than for any of the other methods, consistent with Kang and Schafer (2007).

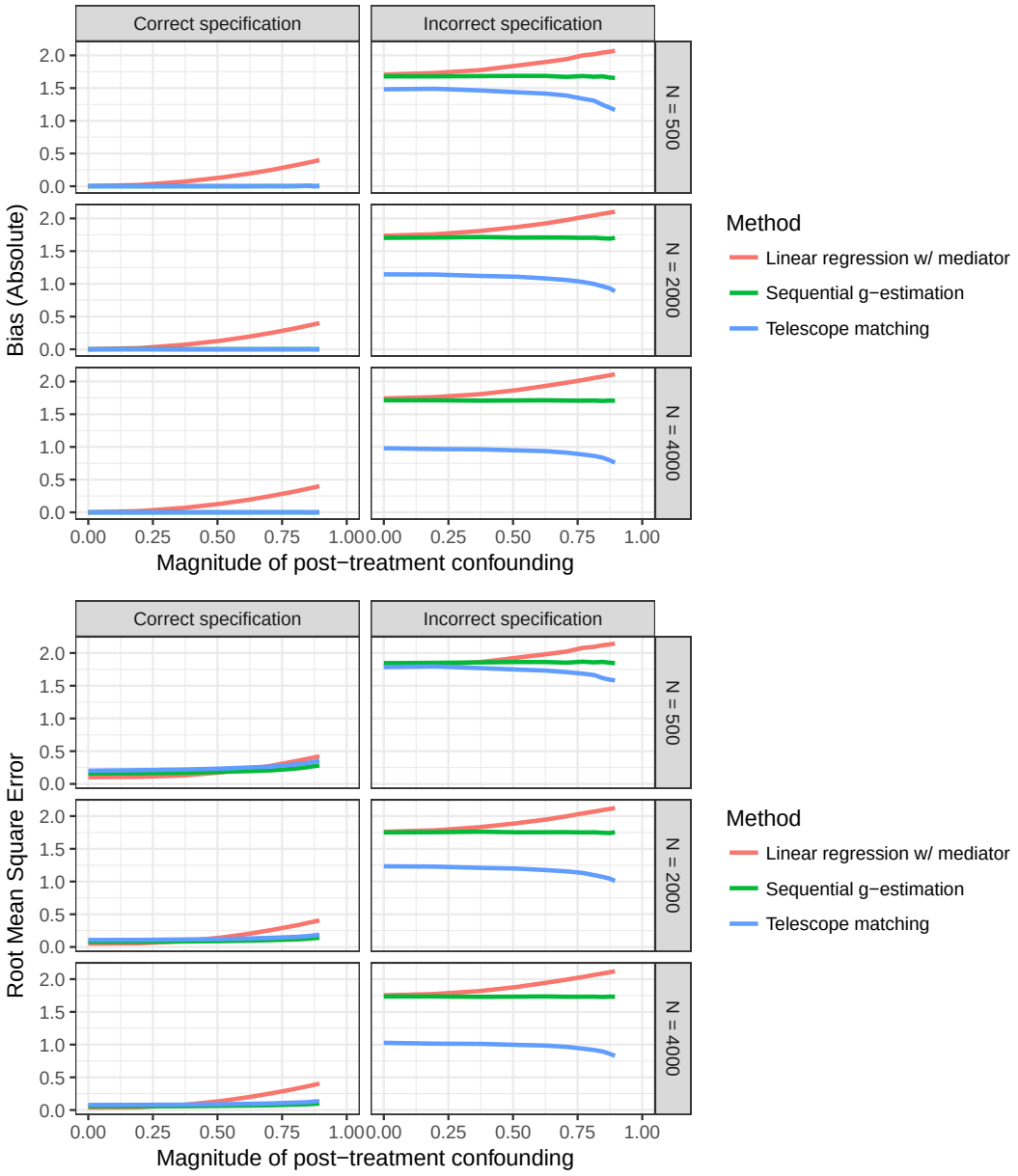

Figure 3: Performance of regression controlling for intermediate covariates, sequential g-estimation, and telescope matching under simulated data with correct and misspecified models.

pected, simply including all of the covariates in a single regression model suffers from the problem of post-treatment bias, the magnitude of which grows as we increase the correlation between the post-treatment confounder and Y_i. When we introduce our misspecification into the imputation model, the performance of sequential g-estimation is notably worse than that of matching, particularly for the larger sample sizes where the efficiency gains of using sequential g-estimation are smaller. In

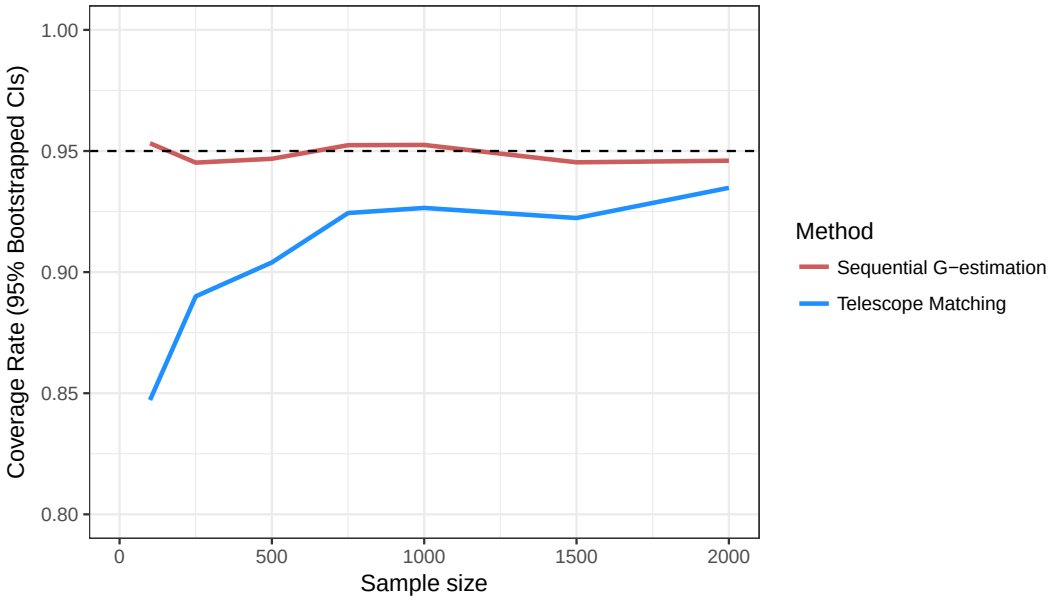

Figure 4: Coverage rate of confidence intervals using bootstrapped standard error estimates under correct model specification. Standard errors estimated using 1000 bootstrap iterations. Each coverage rate estimated using 2500 simulation iterations.

our simulations, we find that the RMSE of sequential g-estimation is about 75% worse than telescope matching for the largest sample size (4000) when the outcome model is incorrectly specified and inconsistent for the true outcome.

We also evaluate the performance of the weighted bootstrap as a way of obtaining standard errors and confidence intervals for our matching imputation method. For a fixed level of post-treatment confounding ($\delta = 3.5$), we simulated the coverage rate of 95% confidence intervals calculated using the conventional nonparametric bootstrap for sequential g-estimation and the nonparametric weighted bootstrap for telescope matching. Figure 4 plots the coverage rates across different sample sizes.

We find that while the boostrapped confidence intervals for telescope matching slightly undercover the true parameter values for smaller sample sizes like $N = 100$, once the sample size is greater than about 750, the coverage rate of the bootstrap is generally close to the desired nominal rate. While we still find undercoverage of about 2 to 3 percentage points, this result is comparable to the performance of the weighted bootstrap in Otsu and Rai (2017). We find that for reasonable sample sizes,

weighted bootstraping provides a reliable method for constructing confidence intervals and conducting hypothesis tests when using telescope matching.

Overall, the simulation results are promising for our proposed method. The findings are are consistent with the argument made in Ho et al. (2006) that matching allows researchers to avoid some of the pitfalls of having to choose the "correct" imputation model. Moreover, at least under the data generating process of this simulation, the loss of power when the true model is somehow known is minimal and far outweighed by the reduction in bias under the more likely case where the researcher happens to select a specification that does not quite match the truth. However, we do caution that larger sample sizes are typically necessary in order to make reliable inferences about controlled direct effects compared to simple average treatment effects.

4 Empirical application: Brader, Valentino, Suhay (2008)

Brader, Valentino and Suhay (2008) analyzes an experiment to assess the effect of different types of media messages on individuals' support for greater levels of immigration. They consider two dimensions along which the media might affect attitudes: tone and ethnic and racial cues. They hypothesize that the combination of a negative tone and an out-group racial prime—stories emphasizing the costs of immigration and a non-white immigrant—will be most likely to provoke opposition to immigration among white respondents.

To evaluate these hypotheses they carried out an experiment on a sample of 354 white, non-Latino adults in the United States. They presented respondents with a mock news article which varied along two dimensions. The article could either emphasize the benefits of immigration or the costs of immigration, and it featured either a white European immigrant or a Latino immigrant. They found that only one combination of the two treatments—a story emphasizing the costs of immigration featuring a Latino immigrant—had a statistically significant effect on increasing respondents' opposition to immigration. Article tone was found to not have a statistically significant effect among respondents exposed to the white European immigrant cue.

The researchers hypothesize that this joint treatment effect of race and tone is mediated by respondents' levels of anxiety about increased levels of immigration to the United States. The original paper uses a "product-of-coefficients" approach to assessing mediation. In a paper on methods for causal mediation analysis, Imai et al. (2011) re-evaluate this hypothesis by estimating what they term the "average causal mediation effect"—the average effect on the outcome of setting the mediator to the value it would take under treatment relative to the value it would take under control, holding treatment constant.

We re-analyze the version of the experimental dataset used in Imai et al. (2011) to instead estimate the controlled direct effect of the combined negative framing/Hispanic immigrant cue when the anxiety mediator is held fixed. While the controlled direct effect considered in this paper is distinct from the ACME defined in Imai et al. (2011), it still can provide researchers with a way of assessing different hypotheses about the causal setting. Specifically, the controlled direct effect gives the effect that remains from treatment if we could somehow control the mediator and set to a common level for the entire population. The existence of a non-zero controlled direct effect is evidence that there exists some amount of the treatment effect that is not entirely transmitted through the mediator being considered.

Figure 5 plots the estimated ACDE of the joint negative tone/Latino treatment on respondents' level of opposition to immigration (measured on a scale from 0 to 4) holding fixed respondents' level of anxiety. Anxiety is a post-treatment quantity which we dichotomize into respondents indicating "low" anxiety (those answering that immigration made them "a little anxious" or "not anxious at all") and high. Our ACDE of interest is the effect of treatment on the four-level outcome fixing respondents' anxiety to "low." As confounders of the mediator, we include four observed pre-treatment covariates measured by the researchers: age, gender, income (measured on a 19-point scale), whether respondents have some college education. We also include one post-treatment confounder: respondents' measured levels of perceived harm due to immigration. Brader, Valentino and Suhay (2008) treat this as an alternative potential mechanism explaining the effect of treatment on respondents' attitudes. However, it is also a plausible confounder of respondents' anxiety levels as beliefs about

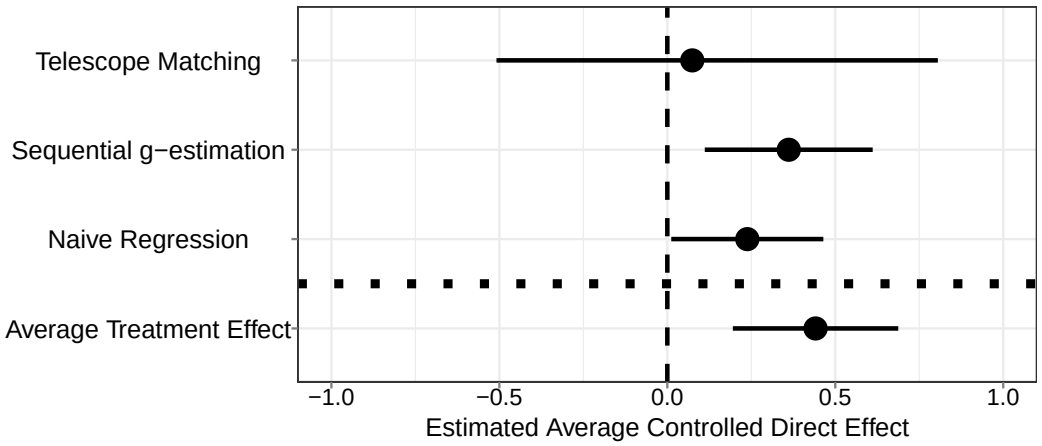

Figure 5: Estimated ACDE fixing anxiety of tone/racial cue treatment on opposition to immigration. N = 265. Lines denote 95% confidence intervals. Sequential g-estimation and telescope matching CIs estimated using 1000 bootstrap iterations.

perceived harm likely affect respondents' emotional response to immigration. While the treatment is randomized, we include covariates in both the first and second-stage models. Though the covariates are not confounders of treatment, including them in the regression helps reduce the variance of each of the estimators. To allow for more flexibility in the telescope matching estimator given the small sample size, we set L, the number of units matched in each stage, to 2.

Compared to the estimated average treatment effect, all estimated ACDEs are smaller, suggesting that intervening on the mediator would eliminate some of the effect of treatment, consistent with the anxiety mediation hypothesis in Brader, Valentino and Suhay (2008). However, both sequential g-estimation and the naive post-treatment regression estimators suggest that there exists a statistically significant effect of treatment that remains unmediated by anxiety. In other words, even holding anxiety at a low level, the tone/race treatment still raises respondents' opposition to immigration on average. Conversely, we find no statistically significant ACDE using telescope matching, obtaining a point estimate very close to 0. Notably, sequential g-estimation yields a much more exaggerated point estimate, while incorporating matching in the imputations attenuates the estimated ACDE closer to zero. While telescope matching does inflate the uncertainty in the estimator, we still have strong reasons to prefer the telescope matching results given robustness to model misspecification. Overall, we do not find strong evidence for a persistent effect of the treatment when anxiety levels are

held fixed, consistent with the argument in Brader, Valentino and Suhay (2008) that the emotional response to immigration plays a primary role in how individuals react to informational cues from the media. In other words, intervening and holding fixed emotional state appears to go a long way towards eliminating the effect of informational cues on immigration attitudes.

5 Conclusion

In this paper, we have introduced a novel method for estimating the direct effect of treatment for fixed values of a mediator. This matching-based approach flexibly imputes missing values of the potential outcomes and appears to be more robust to model misspecification that other semiparametric approaches like sequential g-estimation. This method could be of use to many applied researchers who want to estimate direct effects but have a large degree of uncertainty about the correct model specification for baseline and intermediate covariates. Furthermore, we derived several properties of the estimator, including its large-sample distribution, that allowed us to develop a bias-corrected version of this estimator that augments the matching with regression.

There are several avenues for future work on this frontier. First, it would be interesting to understand how these methods could be extended to estimate quantities of interest in mediation analyses like the natural direct and indirect effect, when the assumptions of that setting holds. Second, we have only compared telescope matching to sequential g-estimation, but it would be fruitful to compare this method to the broader suite of doubly robust methods for direct effects. Finally, we have explored bias correction through simple additive linear regression models but a range of more flexible regression techniques, from generalized additive models to cutting-edge machine learning methods, could plausibly be used as well. In general, this paper illustrates how estimation of controlled direct effects can be treated as a problem of imputing missing potential outcomes $Y_i(a, 0)$. We outline one particular imputation strategy, a two-stage matching estimator, but there are many other imputation methods, each with their own particular advantages and drawbacks, that could be investigated in

subsequent research.

Bibliography

Abadie, Alberto and Guido W. Imbens. 2006. "Large sample properties of matching estimators for average treatment effects." *Econometrica* 74(1):235–267.

Abadie, Alberto and Guido W. Imbens. 2008. "On the Failure of the Bootstrap for Matching Estimators." *Econometrica* 76(6):1537–1557.

Abadie, Alberto and Guido W. Imbens. 2011. "Bias-Corrected Matching Estimators for Average Treatment Effects." *Journal of Business & Economic Statistics* 29(1):1–11.

Abadie, Alberto and Guido W Imbens. 2012. "A martingale representation for matching estimators." *Journal of the American Statistical Association* 107(498):833–843.

Acharya, Avidit, Matthew Blackwell and Maya Sen. 2016. "Explaining Causal Findings Without Bias: Detecting and Assessing Direct Effects." *American Political Science Review* 110(3):512–529.

Acharya, Avidit, Matthew Blackwell and Maya Sen. 2018. "Analyzing Causal Mechanisms in Survey Experiments." *Political Analysis, Forthcoming* .

Billingsley, Patrick. 1995. *Probability and Measure*. 3 ed. New York: John Wiley and Sons.

Blackwell, Matthew and Adam Glynn. 2018. "How to Make Causal Inferences with Time-Series Cross-Sectional Data under Selection on Observables." Working Paper.

Brader, Ted, Nicholas A Valentino and Elizabeth Suhay. 2008. "What triggers public opposition to immigration? Anxiety, group cues, and immigration threat." *American Journal of Political Science* 52(4):959–978.

Efron, B. 1979. "Bootstrap methods: another look at the jackknife." *Annals of Statistics* 7:1–26.

Ho, Daniel E., Kosuke Imai, Gary King and Elizabeth A. Stuart. 2006. "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." *Political Analysis* 15(3):199.

Imai, Kosuke, Luke Keele, Dustin Tingley and Teppei Yamamoto. 2011. "Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies." *American Political Science Review* 105(4):765–789.

Imai, Kosuke, Luke Keele and Teppei Yamamoto. 2010. "Identification, Inference and Sensitivity Analysis for Causal Mediation Effects." *Statistical Science* 25(1):51–71.
URL: *http://projecteuclid.org/euclid.ss/1280841733*

Imbens, Guido W. 2004. "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review." *Review of Economics and Statistics* 86(1):4–29.
URL: *http://www.mitpressjournals.org/doi/abs/10.1162/003465304323023651*

Kang, Joseph D.Y. and Joseph L. Schafer. 2007. "Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data." *Statistical science* 22(4):523–539.

Otsu, Taisuke and Yoshiyasu Rai. 2017. "Bootstrap inference of matching estimators for average treatment effects." *Journal of the American Statistical Association* 112(520):1720–1732.

Richardson, Thomas S. and Andrea Rotnitzky. 2014. "Causal Etiology of the Research of James M. Robins." *Statistical Science* 29(4):459–484.

Robins, James M. 1986. "A new approach to causal inference in mortality studies with sustained exposure periods-Application to control of the healthy worker survivor effect." *Mathematical Modelling* 7(9-12):1393–1512.
URL: *http://biosun1.harvard.edu/ robins/new-approach.pdf*

Robins, James M. 1997. Causal Inference from Complex Longitudinal Data. In *Latent Variable Modeling and Applications to Causality*, ed. M. Berkane. Vol. 120 of *Lecture Notes in Statistics* New York:

Springer-Verlag pp. 69–117.

URL: *http://biosun1.harvard.edu/ robins/cicld-ucla.pdf*

Robins, James M. and Sander Greenland. 1992. "Identifiability and Exchangeability for Direct and Indirect Effects." *Epidemiology* 3(2):143–155.

Rosenbaum, Paul R. 1984. "The consquences of adjustment for a concomitant variable that has been affected by the treatment." *Journal of the Royal Statistical Society. Series A (General)* pp. 656–666.

URL: *http://www.jstor.org/stable/10.2307/2981697*

Rubin, Donald B. 1974. "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of educational Psychology* 66(5):688.

VanderWeele, Tyler. 2015. *Explanation in causal inference: methods for mediation and interaction*. Oxford: Oxford University Press.

A Appendix

A.1 Proofs

Let $\sigma^2(x,a,z,m) = \mathbb{V}[Y_i|X_i = x, A_i = a, Z_i = z, M_i = m]$, $\sigma^2_m(x,a,z) = \mathbb{V}[Y_i(a,m)|X_i = x, A_i = a, Z_i = z]$, and $\sigma^2_\eta(x,a) = \mathbb{V}[E[Y_i(a,0)|X_i = x, Z_i, A_i = a]|X_i = x, A_i = a] = E[\eta_i^2|X_i = x, A_i = a]$.

Assumption 3 (Regularity conditions)**.** *We assume the following:*

(i) Let $V_i = (Z_i, X_i)$ be a random vector of $k = k_z + k_x$ continuous covariates distributed on $\mathbb{R}^k$ with compact and convex support $\mathbb{V}$, with its density bounded and bounded away from zero.

(ii) $\{(Y_i, M_i, Z_i, A_i, X_i)\}_{i=1}^N$ are independent and identically distributed.

(iii) The functions $\mu(z,a,m)$, $\sigma^2(x,z,a,m)$, and $\sigma^2_\eta(x,a)$ are Lipschtiz on $\mathbb{V}$.

(iv) $E[Y_i^4|V_i = v, A_i = a, M_i = a]$ exists and is uniformly bounded in $\mathbb{V}$.

(v) $\sigma^2(x,z,a,m)$ *and* $\sigma^2_\eta(x,a)$ *are bounded away from o.*

Most of the above regularity conditions in Assumption 3 are used by Abadie and Imbens (2006) to derive the asymptotic properties of the simple matching estimator.

Lemma 1. *Suppose that Assumptions 1, 2, and 3 hold. Then, (i) the expectation* $\mathbb{E}[(K_L^{am}(i))^q]$ *is uniformly bounded in N.*

Proof of Lemma 1. Proof available on the author's website. □

Proof of Theorem 1. Let $D_N = \frac{1}{N}\sum_{i=1}^N \tau(X_i) - \tau + E_L^a + E_L^m$. We can write

$$\sqrt{N}D_N = \sum_{k=1}^{3N} \xi_{N,k},$$

where

$$\xi_{N,k} = \begin{cases} \frac{1}{\sqrt{N}}\left(\tau(X_i) - \tau\right), & \text{if } 1 \le k \le N \\ \frac{1}{\sqrt{N}}(2A_i - 1)\left(1 + \frac{K_L^a(i)}{L}\right)\eta_i, & \text{if } N+1 \le k \le 2N \\ \frac{1}{\sqrt{N}}(2A_i - 1)(1 - M_i)\left(1 + \frac{K_L^a(i)}{L} + \frac{K_L^m(i)}{L} + \frac{K_L^{am}(i)}{L^2}\right)\varepsilon_i & \text{if } 2N+1 \le k \le 3N \end{cases}$$

Let $\mathbf{X} = \{X_1, \ldots, X_N\}$, $\mathbf{A} = \{A_1, \ldots, A_N\}$, $\mathbf{Z} = \{Z_1, \ldots, Z_N\}$, and $\mathcal{M}_N = \{M_1, \ldots, M_N\}$. We then define the following σ-fields:

$$\mathcal{F}_{N,k} = \begin{cases} \sigma\{\mathbf{A}, X_1, \ldots, X_k\} & \text{for } 1 \le k \le N \\ \sigma\{\mathbf{A}, \mathbf{X}, Z_1, \ldots, Z_{k-N}\} & \text{for } N+1 \le k \le 2N \\ \sigma\{\mathbf{A}, \mathbf{X}, \mathbf{Z}, \mathbf{M}, Y_1, \ldots, Y_{k-2N}\} & \text{for } 2N+1 \le k \le 3N \end{cases}$$

Following the logic of Abadie and Imbens (2012), we note that

$$\left\{\sum_{j=1}^{i} \xi_{N,j}, \mathcal{F}_{N,i}, 1 \le i \le 3N\right\}$$

is a martingale for $N \ge 1$.

For $1 \leq k \leq N$, the conditional variances of the martingale differences are given by

$$\begin{aligned}\mathbb{E}[\xi_{N,k}^2|\mathcal{F}_{N,k-1}] &= \frac{1}{N}\mathbb{E}[(\tau(X_i) - \tau)^2 \,|A_i], \\ &= \frac{1}{N}\mathbb{E}[(\tau(X_i) - \tau)^2],\end{aligned}$$

where the second equality holds by ignorability. For $N+1 \leq k \leq 2N$, the conditional variances are

$$\begin{aligned}\mathbb{E}[\xi_{N,k}^2|\mathcal{F}_{N,k-1}] &= \frac{1}{N}\left(1 + \frac{K_L^a(i)}{L}\right)^2 \mathbb{E}[\eta_i^2|\mathbf{X}, \mathbf{A}], \\ &= \frac{1}{N}\left(1 + \frac{K_L^a(i)}{L}\right)^2 \sigma_\eta^2(X_i, A_i)\end{aligned}$$

Finally, for the $2N+1 \leq k \leq 3N$, the conditional variances of the martingale differences are

$$\begin{aligned}\mathbb{E}[\xi_{N,k}^2|\mathcal{F}_{N,k-1}] &= \frac{1}{N}(1 - M_i)\left(1 + \frac{K_L^a(i)}{L} + \frac{K_L^m(i)}{L} + \frac{K_L^{am}(i)}{L^2}\right)\mathbb{E}[\varepsilon_i^2|\mathbf{X}, \mathbf{A}, \mathbf{Z}, \mathbf{M}], \\ &= \frac{1}{N}(1 - M_i)\left(1 + \frac{K_L^a(i)}{L} + \frac{K_L^m(i)}{L} + \frac{K_L^{am}(i)}{L^2}\right)\sigma^2(X_i, Z_i, A_i, 0).\end{aligned}$$

Thus, we can invoke a weak law of large numbers argument to show that

$$\sum_{k=1}^{3N} \mathbb{E}[\xi_{N,k}^2|\mathcal{F}_{N,k-1}] \xrightarrow{p} \sigma^2,$$

where

$$\begin{aligned}\sigma^2 &= \mathbb{E}[(\tau(X_i) - \tau)^2] \\ &+ \mathbb{E}\left[\left(1 + \frac{K_L^a(i)}{L}\right)^2 \sigma_\eta^2(X_i, A_i)\right] \\ &+ \mathbb{E}\left[(1 - M_i)\left(1 + \frac{K_L^a(i)}{L} + \frac{K_L^m(i)}{L} + \frac{K_L^{am}(i)}{L^2}\right)^2 \sigma^2(X_i, Z_i, A_i, 0)\right]\end{aligned} \tag{8}$$

To establish the large sample distribution of D_N, we use a Lyapunov condition:

$$\sum_{k=1}^{3N} \mathbb{E}[|\xi_{N,k}|^{2+\delta}] \to 0, \qquad \text{for some } \delta > 0.$$

Note that this condition implies the more standard Lindeberg condition used in martingal central limit theorems (Billingsley, 1995; Abadie and Imbens, 2012).

From Lemma 3 of Abadie and Imbens (2006) and Lemma 1, we have that $E[(K_L^a(i)/L)^4]$, $E[(1 + K_L^m(i)/L)^4]$, and $\mathbb{E}[(K_L^{am}(i))^4]$ are uniformly bounded. Through iterated use of Minkowski inequality,

we have

$$\mathbb{E}\left[\left(1+\frac{K_L^a(i)}{L}+\frac{K_L^m(i)}{L}+\frac{K_L^{am}(i)}{L^2}\right)^4\right] \leq \mathbb{E}\left[\left(1+\frac{K_L^m(i)}{L}\right)^4\right]^{1/4}+\left(E\left[\left(\frac{K_L^a(i)}{L}\right)^4\right]^{1/4}+E\left[\left(\frac{K_L^{am}(i)}{L^2}\right)^4\right]^{1/4}\right)^{1/4}$$

which implies that $(1+K_L^a/L+K_L^m/L+K_L^{am}(i)/L^2)$ are uniformly bounded.

Letting $\delta = 2$, look at one term from $2N+1 \leq k \leq 3N$:

$$\mathbb{E}[\xi_{N,k}^4] = \frac{1}{N^2}\mathbb{E}\left[(1-M_i)\left(1+\frac{K_L^a(i)}{L}+\frac{K_L^m(i)}{L}+\frac{K_L^{am}(i)}{L^2}\right)^4 \mathbb{E}[\varepsilon_i^4|X_iA_i,Z_i,M_i]\right].$$

The expectation on the right-hand side is bounded given the bound on fourth moments for Y_i (part (iv) of Assumption 3) and the earlier mentioned uniform bound on $(1+K_L^a/L+K_L^m/L+K_L^{am}(i)/L^2)$. Similar analyses can be conducted for the other two parts of the martingale, which will ensure the Lyapunov condition holds. Thus, based on the martingale central limit theorem (Billingsley, 1995, Theorem 35.12), we have $\sqrt{N}D_N \xrightarrow{d} N(0,\sigma^2)$.

□

A.2 Bias correction

Let $\lambda = (\lambda_1,\ldots,\lambda_k)$ be a k-dimensional vector of nonnegative integers with order $|\lambda| = \sum_{i=1}^k \lambda_i$. This vector will define a polynomial in x, so that $x^\lambda = x_1^{\lambda_1}\cdots x_k^{\lambda_k}$. We collect these vectors into a series that is nondecreasing in its order. That is, we let $\{\lambda(G)\}_{G=1}^\infty$ be a series of all distinct λ vectors such that $|\lambda(G)|$ is nondecreasing. Let $p_G(x) = x^{\lambda(G)}$ be the polynomial induced by the Gth vector in this series. Let $p^G(x) = (p_1(x),\ldots,p_G(x))'$ be the vector of all such polynomials up to G. We now define two nonparametric series estimators for the two conditional expectations of interest:

$$\widehat{\mu}_{a0}(x,z) = p^{G(N)}(x,z)'\left(\sum_{i:A_i=a,M_i=0} p^{G(N)}(X_i,Z_i)p^{G(N)}(X_i)'\right)^-\left(\sum_{i:A_i=a,M_i=0} p^{G(N)}(X_i,Z_i)Y_i\right)$$

$$\widehat{\mu}_{a0}(x) = p^{G(N)}(x)'\left(\sum_{i:A_i=a} p^{G(N)}(X_i)p^{G(N)}(X_i)'\right)^-\left(\sum_{i:A_i=a} p^{G(N)}(X_i)\widetilde{Y}_{i0}\right)$$

Here, $(\cdot)^-$ represents a generalized inverse. Essentially, these are linear regression models regressing the outcome (or transformed outcome) on a vector of polynomials of the appropriate covariates for

that regression. Furthermore, the polynomials grow more complex as the sample size grows, which ensures that this approximation will converge to the true expectation under the appropriate smoothness conditions. Note that we use the first-stage regression $\widehat{\mu}_{A_i 0}(X_i, Z_i)$ to construct the bias-corrected outcome for the second stage, $\widetilde{Y}_{i0}$.

Assumption 4 (Smoothness for bias correction). *We assume the following:*

(i) $G(N) = O(N^v)$ *with* $0 < v < \min(2/(4k+3), 2/(4k^2-k))$;

(ii) there are constants $C_1 > 0$ *and* $C_2 > 0$ *such that for each* λ *with* $|\lambda| = k$, *the derivatives* $\partial^\lambda \mu(x, z, a, 0)$ *and* $\partial^\lambda \mu_{a0}(x, a)$ *exist and satisfy* $\sup_{(x,z)\in\mathbb{V}} |\partial^\lambda \mu(x, z, a, 0)| < C_1$ *and* $\sup_{x\in\mathbb{X}} |\partial^\lambda \mu_{a0}(x, a)| < C_2$.

It is possible write the bias-corrected estimator as follows: $\widetilde{\tau} = \widehat{\tau} - \widehat{B}_L^m - \widehat{B}_L^a$, where the first term is the simple matching estimator and the latter two terms are:

$$\widehat{B}_L^m = \frac{1}{N}\sum_{i=1}^{N}(2A_i - 1)\left(1 + \frac{K_L^a(i)}{L}\right) M_i \left(\frac{1}{L}\sum_{\ell\in J_L^m(i)} \widehat{\mu}_{A_i 0}(X_\ell, Z_\ell) - \widehat{\mu}_{A_i 0}(X_i, Z_i)\right)$$

$$\widehat{B}_L^a = \frac{1}{N}\sum_{i=1}^{N}(2A_i - 1)\left[\frac{1}{L}\sum_{j\in J_L^a(i)} \widehat{\mu}_{1-A_i,0}(X_i) - \widehat{\mu}_{1-A_i,0}(X_j)\right]$$

These converge in probability to the bias terms B_L^m and B_L^a, respectively. Define the following:

$$\widetilde{B}_L^a = \frac{1}{N}\sum_{i=1}^{N}(2A_i - 1)\left[\frac{1}{L}\sum_{j\in J_L^a(i)} \widetilde{\mu}_{1-A_i,0}(X_i, 1-A_i) - \widetilde{\mu}_{1-A_i,0}(X_j, 1-A_i)\right]$$

where, $\widetilde{\mu}_{a0}(x, a) = \mathbb{E}[\widetilde{Y}_{i0}|X_i = x, A_i = a]$.

Theorem 2 (Bias-corrected matching). *Suppose that Assumptions 1-4 hold. Then,*

$$\sqrt{N}\left(B_L^m + B_L^a - (\widehat{B}_L^m + \widehat{B}_L^a)\right) \xrightarrow{p} 0, \qquad \textit{and}$$

$$\sqrt{N}(\widetilde{\tau} - \tau) \xrightarrow{d} N(0, \sigma^2).$$

Proof of Theorem 2. Let Λ_ℓ be the set of vectors λ such that $|\lambda| = \ell$ and let $\partial^\lambda g(x) = \partial^{|\lambda|} g(x)/\partial x_1^{\lambda_1}\cdots\partial x_k^{\lambda_k}$. Finally, for $d \geq 0$, define $|g|_d = \max_{|\lambda|\leq d} \sup_x |\partial^\lambda g(x)|$. Lemma A.1 of Abadie and Imbens (2011)

uniformly bounded discrepancy between the partial derivatives of the series estimator and the true CEF, which here implies that:

$$|\widehat{\mu}_{a0} - \mu_{a0}|_d = \max_{|\lambda| \leq d} \sup_{(x,z) \in \mathbb{V}} |\partial^\lambda (\widehat{\mu}_{a0}(x,z) - \mu_{a0}(x,z))| = O_p(G^{1+2d}((G/N)^{1/2} + G^{-\zeta})). \tag{9}$$

This convergence ensures that $|\widehat{B}_L^m - B_L^m| = o_p(N^{-1/2})$ by the same logic as the proofs in Lemma A.2 and Theorem 2 of Abadie and Imbens (2011). Similar logic will ensure that $|\widehat{B}_L^a - \widetilde{B}_L^a| = o_p(N^{-1/2})$. By the triangle inequality, we have $|\widehat{B}_L^a - B_L^a| \leq |\widehat{B}_L^a - \widetilde{B}_L^a| + |\widetilde{B}_L^a - B_L^a|$.

We could apply the same logic as these other steps to $|\widetilde{B}_L^a - B_L^a|$ if we had a similar uniform convergence result for $|\widetilde{\mu}_a - \mu_a|_d$. To do so, we use the definition of $\widetilde{Y}_{i0}$ to derive the following:

$$\widetilde{\mu}_{a0}(x) = \mu_{a0}(x) + b(x,a)$$

where,

$$b_a(x) = \mathbb{E}\left[M_i \left(\frac{1}{L} \sum_{\ell \in J_L^m(i)} \widehat{\mu}(x,a,Z_i,0) - \widehat{\mu}(X_\ell,a,Z_\ell,0) - (\mu(x,a,Z_i,0) - \mu(X_\ell,a,Z_\ell,0)) \right) \middle| X_i = x, A_i = a \right]$$

Thus, we have: $|\widetilde{\mu}_{a0} - \mu_{a0}|_d = \max_{|\lambda| \leq d} \sup_{x \in \mathbb{X}} |\partial^\lambda b_a(x)|$. Note that in this second stage, we only consider partial derivatives with respect to the baseline covariates, where as in the first stage, we consider partial derivatives with respect to the baseline and intermediate covariates. Thus, for the functions, $\mu_{a0}(x,z)$ and its estimate, $\widehat{\mu}_{a0}(x,z)$, the above derivative discrepancy norm with respect to just x will be bounded by the one that covers x and z in (9). Thus, the interior of expectation of the $b_a(x)$ function can be bounded and is integrable, which means we can use Lebesgue's dominated convergence theorem to switch the order of differentiation and expectation (while also applying the triangle inequality):

$$|\partial^\lambda b_a(x)| \leq \mathbb{E}\left[\frac{1}{L} \sum_{\ell \in J_L^m(i)} \left| \partial^\lambda \left(\widehat{\mu}(x,a,Z_i,0) - \widehat{\mu}(X_\ell,a,Z_\ell,0) - (\mu(x,a,Z_i,0) - \mu(X_\ell,a,Z_\ell,0)) \right) \right| \middle| X_i = x, A_i = a \right]$$

Again, because the difference in the derivatives of these expectations are bounded, the entire function inside the expectation will be bounded by $2 \times |\widehat{\mu}_{a0} - \mu_{a0}|_d$. Thus, we have:

$$|\widetilde{\mu}_{a0} - \mu_{a0}|_d = O_p(G^{1+2d}((G/N)^{1/2} + G^{-\zeta})).$$

Using the same logic as Lemma A.2 in Abadie and Imbens (2011), this allows us to show that

$$\max_{i=1,\ldots,N} |\widetilde{\mu}_{a0}(X_i) - \widetilde{\mu}_{a0}(X_\ell) - (\mu_{a0}(X_i) - \mu_{a0}(X_\ell))| = o_p(N^{-1/2})$$

for $a = 0, 1$. We can then apply Theorem 2 of Abadie and Imbens (2011) to show $|\widehat{B}^a_L - B^a_L| = o_p(N^{-1/2})$, thus showing that the bias correction will not affect the asymptotic variance of the matching estimator.

□

A.3 Weighted bootstrap derivation

To derive the form of the individual $\widetilde{\tau}_i$ for the weighted bootstrap, we start by writing $\widetilde{\tau}$ in terms of the naive matching estimator $\widehat{\tau}$ and the two bias-corrections $\widehat{B}^m_L$ and $\widehat{B}^a_L$.

$$\begin{aligned}
\widetilde{\tau} &= \widehat{\tau} - \widehat{B}^m_L - \widehat{B}^a_L && (10)\\
&= \sum_{i=1} \widetilde{\tau}_i && (11)\\
&= \sum_{i=1}^{N} \widehat{\tau}_i - \widehat{B}^m_{Li} - \widehat{B}^a_{Li} && (12)\\
&= \sum_{i=1}^{N} \widehat{\tau}_i - \sum_{i=1}^{N} \widehat{B}^m_{Li} - \sum_{i=1}^{N} \widehat{B}^a_{Li} && (13)
\end{aligned}$$

First, to derive $\widehat{\tau}_i$, we write the naive matching estimator as:

$$\begin{aligned}
\hat{\tau} &= \sum_{i=1}^{N} (2A_i - 1) \left(1 + \frac{K^a_L(i)}{L}\right) \widehat{Y}_{i0} && (14)\\
&= \sum_{i=1}^{N} (2A_i - 1) \left[(1 - M_i) \left(1 + \frac{K^a_L(i)}{L}\right) Y_i + M_i \left(1 + \frac{K^a_L(i)}{L}\right) \frac{1}{L} \sum_{j \in \mathcal{J}^m_L(i)} Y_j\right] && (15)\\
&= \sum_{i=1}^{N} (2A_i - 1) \left[(1 - M_i) \left(1 + \frac{K^a_L(i)}{L} + \frac{K^m_L(i)}{L} + \frac{K^{am}_L(i)}{L^2}\right) Y_i\right] && (16)\\
\hat{\tau}_i &= (2A_i - 1) \left[(1 - M_i) \left(1 + \frac{K^a_L(i)}{L} + \frac{K^m_L(i)}{L} + \frac{K^{am}_L(i)}{L^2}\right) Y_i\right] && (17)
\end{aligned}$$

with the second to last line following from the fact that all units used for matching imputation in the first stage have $M_i = 0$.

The first-stage bias correction, $\widehat{B}_L^m$ can also be written as

$$\widehat{B}_L^m = \frac{1}{N}\sum_{i=1}^{N}(2A_i-1)\left(1+\frac{K_L^a(i)}{L}\right)M_i\left(\frac{1}{L}\sum_{\ell\in\mathcal{J}_l^m(i)}\widehat{\mu}_{A_i0}(X_\ell,Z_\ell,A_i)-\widehat{\mu}_{A_i0}(X_i,Z_i,A_i)\right) \tag{18}$$

$$= \frac{1}{N}\sum_{i=1}^{N}(2A_i-1)\left[\left(1+\frac{K_L^a(i)}{L}\right)M_i\left(\frac{1}{L}\sum_{\ell\in\mathcal{J}_l^m(i)}\widehat{\mu}_{A_i0}(X_\ell,Z_\ell,A_i)\right)-M_i\left(1+\frac{K_L^a(i)}{L}\right)\widehat{\mu}_{A_i0}(X_i,Z_i,A_i)\right] \tag{19}$$

$$= \frac{1}{N}\sum_{i=1}^{N}(2A_i-1)\left[(1-M_i)\left(\frac{K_L^m(i)}{L}+\frac{K_L^{am}(i)}{L^2}\right)\widehat{\mu}_{A_i0}(X_i,A_i,Z_i)-M_i\left(1+\frac{K_L^a(i)}{L}\right)\widehat{\mu}_{A_i0}(X_i,A_i,Z_i)\right] \tag{20}$$

$$= \frac{1}{N}\sum_{i=1}^{N}(2A_i-1)\widehat{\mu}_{A_i0}(X_i,A_i,Z_i)\left[(1-M_i)\left(\frac{K_L^m(i)}{L}+\frac{K_L^{am}(i)}{L^2}\right)-M_i\left(1+\frac{K_L^a(i)}{L}\right)\right] \tag{21}$$

$$\widehat{B}_{Li}^m = (2A_i-1)\widehat{\mu}_{A_i0}(X_i,A_i,Z_i)\left[(1-M_i)\left(\frac{K_L^m(i)}{L}+\frac{K_L^{am}(i)}{L^2}\right)-M_i\left(1+\frac{K_L^a(i)}{L}\right)\right] \tag{22}$$

Note that because of exact matching in the first stage on A_i, $A_i = A_\ell$ for all $\ell \in \mathcal{J}_L^m(i)$.

And finally the second-stage bias correction $\widehat{B}_L^a$ can be rewritten as

$$\widehat{B}_L^a = \frac{1}{N}\sum_{i=1}^{N}(2A_i-1)\left[\frac{1}{L}\sum_{j\in J_L^a(i)}\widehat{\mu}_{1-A_i,0}(X_i,1-A_i)-\widehat{\mu}_{1-A_i,0}(X_j,1-A_i)\right] \tag{23}$$

$$= \frac{1}{N}\sum_{i=1}^{N}(2A_i-1)\left(\widehat{\mu}_{1-A_i,0}(X_i,1-A_i)+\frac{K_L^a(i)}{L}\widehat{\mu}_{A_i,0}(X_i,A_i)\right) \tag{24}$$

$$\widehat{B}_{Li}^a = (2A_i-1)\left(\widehat{\mu}_{1-A_i,0}(X_i,1-A_i)+\frac{K_L^a(i)}{L}\widehat{\mu}_{A_i,0}(X_i,A_i)\right) \tag{25}$$

Combining the three linearized terms yields

$$\begin{aligned}\widetilde{\tau}_i = (2A_i-1)\Bigg[&(1-M_i)\left(1+\frac{K_L^a(i)}{L}+\frac{K_L^m(i)}{L}+\frac{K_L^{am}(i)}{L^2}\right)Y_i \\ &-\left((1-M_i)\left(\frac{K_L^m(i)}{L}+\frac{K_L^{am}(i)}{L^2}\right)-M_i\left(1+\frac{K_L^a(i)}{L}\right)\right)\widehat{\mu}_{A_i,0}(X_i,A_i,Z_i) \\ &-\left(\widehat{\mu}_{1-A_i,0}(X_i,1-A_i)+\frac{K_L^a(i)}{L}\widehat{\mu}_{A_i,0}(X_i,A_i)\right)\Bigg]\end{aligned} \tag{26}$$