
Interpretable Almost-Matching-Exactly With Instrumental Variables

M. Usaid Awan,* Yameng Liu,* Marco Morucci,* Sudeepa Roy, Cynthia Rudin, Alexander Volfovsky

Duke University
Durham, NC, USA

Abstract

Uncertainty in the estimation of the causal effect in observational studies is often due to unmeasured confounding, i.e., the presence of unobserved covariates linking treatments and outcomes. Instrumental Variables (IV) are commonly used to reduce the effects of unmeasured confounding. Existing methods for IV estimation either require strong parametric assumptions, use arbitrary distance metrics, or do not scale well to large datasets. We propose a matching framework for IV in the presence of observed categorical confounders that addresses these weaknesses. Our method first matches units exactly, and then consecutively drops variables to approximately match the remaining units on as many variables as possible. We show that our algorithm constructs better matches than other existing methods on simulated datasets, and we produce interesting results in an application to political canvassing.

1 INTRODUCTION

The gold standard for inferring the causal effect of a treatment (such as smoking, a tax policy, or a fertilizer) on an outcome (such as blood pressure, stock prices, or crop yield) is the randomized experiment: the analyst manually assigns the treatment to each of her units uniformly at random. Unfortunately, this manipulation is impossible or unethical for some treatments of practical interest, leading to the need for inferring causal relations from observational studies. In many observational studies, it is common for *instrumental variables* (IV) to be available. These variables are (a) allocated randomly across units,

(b) correlated with the treatment, and (c) affect the dependent variable only through their effect on the treatment. The fact that instrumental variables allow for consistent estimation of causal effect with non-randomized treatments is a hallmark of the causal inference literature, and has led to the use of IV methods across many different applied settings (e.g., Joskow, 1987; Gerber and Green, 2000; Acemoglu et al., 2001; Autor et al., 2013).

The most popular existing method that uses instrumental variables to conduct causal inference is Two-Stage Least Squares Regression (2SLS) (Angrist and Keueger, 1991; Card, 1993; Wooldridge, 2010). The 2SLS methodology makes strong parametric assumptions about the underlying outcome model (linearity), which do not generalize well to complex problems. Non-parametric approaches to IV-based causal estimates generalize 2SLS to more complex models (Newey and Powell, 2003; Frölich, 2007), but lack interpretability; it is difficult to troubleshoot or trust black box models. Matching methods that allow for nonparametric inference on average treatment effects without requiring functional estimation have recently been introduced for the IV problem in Kang et al. (2016): the full-matching algorithm presented in their work relaxes some of the strong assumptions of 2SLS, however, it does not scale well to massive datasets, and imposes a fixed metric on covariates. It also does not take into account that covariates have different levels of importance for matching.

The approach for instrumental variable analysis presented in this paper aims to handle the problems faced by existing methods: it is non-parametric, scalable, and preserves the interpretability of having high-quality matched groups. We create an Almost-Matching Exactly framework (Wang et al., 2019; Dieng et al., 2019) for the purpose of instrumental variable analysis. Our methodology estimates the causal effects in a non-parametric way and hence performs better than 2SLS or other parametric models. It improves over existing matching methods for instrumental variables when covariates are discrete, leveraging an

*Equal contribution.

adaptive distance metric. This adaptive distance metric is capable of systematically accounting for nuisance variables, discounting their importance for matching. The algorithm scales easily to large datasets (millions of observations) and can be implemented within most common database systems for optimal performance.

In what follows, first we introduce the problem of instrumental variable estimation for observational inference, and describe the role of matching within it. Second, we outline the Almost-Matching Exactly with Instrumental Variables (AME-IV) framework for creating matched groups. Third, we describe estimators with good statistical properties that can be used on the matched data. Finally, we present results from applying our methodology to both simulated and real-world data: we show that the method performs well in most settings and outperforms existing approaches in several scenarios.

2 RELATED WORK

Widely used results on definition and identification of IVs are given in Imbens and Rubin (1997); Angrist et al. (1996), and generalized in Brito and Pearl (2002); Chen et al. (2016). Methods for discovery of IVs are developed in Silva and Shimizu (2017).

The most popular method for IV estimation in the presence of observed confounders is two-stage least squares (2SLS) (Card, 1993). 2SLS estimators are consistent and efficient under linear single-variable structural equation models with a constant treatment effect (Wooldridge, 2010). One drawback of 2SLS is its sensitivity to misspecification of the model. Matching, on the other hand, allows for correct inference without the need to specify an outcome model.

Recent work on matching for IV estimation includes matching methods that match directly on covariates, rather than on summary statistics like propensity score (Ichimura and Taber, 2001). These matching methods can be very powerful nonparametric estimators; full matching (Kang et al., 2013) is one such approach, but has a limitation in that its distance metric between covariates is fixed, whereas ours is learned. Wang et al. (2019) provides an in-depth discussion of other matching methods including near-far and full-matching, in the context of AME.

Other IV methods in the presence of measured covariates include Bayesian methods (Imbens and Rubin, 1997), semiparametric methods (Abadie, 2003; Tan, 2006; Ogburn et al., 2015), nonparametric methods (Frölich, 2007) and deep learning methods (Hartford et al., 2017), but these methods do not enjoy the benefits of interpretability that matching provides.

3 METHODOLOGY

We consider the problem of instrumental variable estimation for a set of n units indexed by $i = 1, \dots, n$. Each unit is randomly assigned to a binary instrument level. Units respond to being assigned different levels of this instrument by either taking up the treatment or not: we denote with $t_i(1), t_i(0) \in \{0, 1\}$ the treatment level taken up by each unit after being exposed to value $z \in \{0, 1\}$ of the instrument. Subsequently, units respond to a treatment/instrument regime by exhibiting different values of the outcome variable of interest, which we denote by $y_i(t_i(1), 1), y_i(t_i(0), 0) \in \mathbb{R}$. Note that this response depends both on the value of the instrument assigned (2nd argument) and on the treatment value that units take up in response to that instrument value (1st argument). All quantities introduced so far are fixed for a given unit i but not always observed. In practice, we have a random variable $Z_i \in \{0, 1\}$ for each unit denoting the level of instrument that it was assigned, and observed realizations of Z_i are denoted with z_i . Whether a unit receives treatment is now a random variable (T_i), and the outcome is random (Y_i), and they take the form:

$$Y_i = y_i(t_i(1), 1)Z_i + y_i(t_i(0), 0)(1 - Z_i)$$

$$T_i = t_i(1)Z_i + t_i(0)(1 - Z_i).$$

Note that the only randomness in the observed variables comes from the instrument, all other quantities are fixed. We use y_i and t_i to denote observed realizations of Y_i and T_i respectively. We also observe a fixed vector of p covariates for each unit, $\mathbf{x}_i \in \mathcal{X}$, where \mathcal{X} is a space with p dimensions. In this paper we are interested in the case in which $\mathcal{X} = \{0, 1\}^p$, corresponding to categorical variables, where exact matching is well-defined.

Throughout we make the SUTVA assumption, that is (i) outcome and treatment assignment for each individual are unrelated to the instrument exposure of other individuals, and (ii) the outcome for each individual is unrelated to the treatment assignment of other individuals (Angrist et al., 1996). However, ignorability of treatment assignment is not required. We make use of the instrumental variable to estimate the causal effect of treatment on outcome. In order for a variable to be a valid instrument it must satisfy the following standard assumptions (see, e.g., Imbens and Angrist, 1994; Angrist et al., 1996; Imbens and Rubin, 2015):

(A1) Relevance: $\frac{1}{n} \sum_{i=1}^n t_i(1) - t_i(0) \neq 0$, that is, the variable does indeed have a non-zero causal effect on treatment assignment, on average.

(A2) Exclusion: If $z \neq z'$ and $t_i(z) = t_i(z')$ then $y_i(t_i(z), z) = y_i(t_i(z'), z')$ for each unit i . This assumption states that unit i 's potential outcomes are only af-

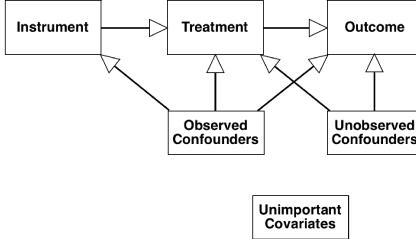


Figure 1: Causal DAG for instrumental variables. Arrows represent causal relationships between variables. The lack of a direct arrow from instrument to outcome represents Assumption A2 and the lack of a direct arrow from unobserved confounders to the instrument represents A3.

fected by the treatment it is exposed to, and not by the value of the instrument. Therefore, $y_i(t_i(z), z)$ can be denoted by: $y_i(t_i(z))$.

(A3) Ignorability: $Pr(Z_i = 1 | \mathbf{x}_i) = e(\mathbf{x}_i)$ for all units i , and some non-random function $e : \mathcal{X} \mapsto (0, 1)$. This assumption states that the instrument is assigned to all units that have covariate value \mathbf{x}_i with the same probability. It implies that if two units i and k have $\mathbf{x}_i = \mathbf{x}_k$, then $Pr(Z_i = 1 | \mathbf{x}_i) = Pr(Z_k = 1 | \mathbf{x}_k)$.

(A4) Strong Monotonicity: $t_i(1) \geq t_i(0)$ for each unit i . This assumption states that the instrument is seen as an encouragement to take up the treatment, this encouragement will only make it more likely that units take up the treatment and never less likely.

Figure 1 gives a graphical summary of the identification assumptions. An instrumental variable satisfying (A1, A2, A3 and A4) allows us to estimate the treatment effect for a subgroup that responds positively to exposure to the instrument (Imbens and Angrist, 1994). We note that these are not the only criteria for the use of instrumental variables, for example Brito and Pearl (2002) introduces a graphical criterion for identification with instrumental variables. These are units that would have undertaken the treatment only after administration of the instrument and never without (Angrist et al., 1996). Note that we cannot identify these units in our sample, given what we observe, but we can estimate the treatment effect on them (Imbens and Rubin, 2015). This treatment effect is known as Local Average Treatment Effect (LATE) and takes the following form (Imbens and Angrist, 1994; Angrist et al., 1996):

$$\begin{aligned} \lambda &= \frac{1}{n_c} \sum_{i: t_i(1) > t_i(0)} y_i(1) - y_i(0) \\ &= \frac{\sum_{\mathbf{x} \in \mathcal{X}} \omega_{\mathbf{x}} \text{ITT}_{y, \mathbf{x}}}{\sum_{\mathbf{x} \in \mathcal{X}} \omega_{\mathbf{x}} \text{ITT}_{t, \mathbf{x}}}, \end{aligned} \quad (1)$$

where n_c is the total number of units such that $t_i(1) >$

$t_i(0)$, $\omega_{\mathbf{x}} = n_{\mathbf{x}}/n$ is the weight associated with each value of \mathbf{x} , $n_{\mathbf{x}}$ is the number of units where $\mathbf{x}_i = \mathbf{x}$, and:

$$\begin{aligned} \text{ITT}_{y, \mathbf{x}} &= \frac{1}{n_{\mathbf{x}}} \sum_{i: \mathbf{x}_i = \mathbf{x}} y_i(t_i(1)) - y_i(t_i(0)) \\ \text{ITT}_{t, \mathbf{x}} &= \frac{1}{n_{\mathbf{x}}} \sum_{i: \mathbf{x}_i = \mathbf{x}} t_i(1) - t_i(0). \end{aligned}$$

The quantities above are also known as the Intent-To-Treat effects: they represent the causal effects of the instrument on the outcome and the treatment, respectively. Intuitively, these effects can be estimated in an unbiased and consistent way due to ignorability of instrument assignment (A3) conditional on units having the same value of \mathbf{x} .

Approximate matching comes into this framework because in practice we almost never have enough treated and control units with the same exact values of \mathbf{x} in our observed data to accurately estimate the quantities above. With approximate matching, we want to construct matched groups from observed \mathbf{x} such that A3 holds approximately within each group. This means that a good approximate matching algorithm is one that produces groups where, if i and j are grouped together, then $\mathbf{x}_i \approx \mathbf{x}_j$. In the next section, we propose the Almost-Matching Exactly with Instrumental Variables (AME-IV) framework to build good approximately matched groups from binary covariates.

3.1 ALMOST-MATCHING EXACTLY WITH INSTRUMENTAL VARIABLES (AME-IV PROBLEM)

The AME-IV framework has the goal of matching each instrumented (i.e., $z_i = 1$) unit to at least one non-instrumented unit (i.e., $z_k = 0$) as exactly as possible. (The entire set of calculations is symmetric when we match each non-instrumented unit, thus w.l.o.g. we consider only instrumented units.) When units are matched on all covariates, this is an exact match. When units can be matched on the most important covariates (but not necessarily all covariates), this is an almost-exact match. The importance of covariate j for matching is represented by a fixed nonnegative weight w_j . Thus, we consider the following problem for each instrumented unit i , which is to maximize the weighted sum of covariates on which we can create a valid matched group for i :

$$\begin{aligned} \boldsymbol{\theta}^{i*} &\in \operatorname{argmax}_{\boldsymbol{\theta} \in \{0,1\}^p} \boldsymbol{\theta}^T \mathbf{w}, \text{ such that} \\ &\exists k \text{ with } z_k = 0 \text{ and } \mathbf{x}_k \circ \boldsymbol{\theta} = \mathbf{x}_i \circ \boldsymbol{\theta}, \end{aligned}$$

where \circ denotes the Hadamard product, $\boldsymbol{\theta}$ is a binary vector to represent whether or not each covariate is used for matching, and \mathbf{w} is a nonnegative vector with a reward

value associated with matching on each covariate. The constraint in our optimization problem definition guarantees that the main matched group of each instrumented unit i contains at least one non-instrumented unit. The solution to this optimization problem is a binary indicator of the optimal set of covariates that unit i can be matched on. Note that, if all entries of θ^{i*} happen to be one, then the units in unit i 's main matched group will be exact matches for i .

We define i 's **main matched group** in terms of θ^{i*} as:

$$\text{MG}(\theta^{i*}, \mathbf{x}_i) = \{k : \theta^{i*} \circ \mathbf{x}_k = \theta^{i*} \circ \mathbf{x}_i\}.$$

We now theoretically connect Assumption A3 with solving the AME-IV problem, and show how approximate matches can lead to the assumption being approximately satisfied within each matched group. This makes IV estimation possible even when it is not possible to exactly match each unit. To do so, we introduce the notation $\mathbf{1}_{[\mathbf{x}_i \neq \mathbf{x}_k]}$ to denote a vector of length p where the j^{th} entry is one if $x_{ij} = x_{kj}$ and zero otherwise.

Lemma 3.1. *For any unit i where $z_i = 1$, with θ^{i*} as defined in the AME-IV problem, then for any unit k with $z_k \neq z_i$, if $\mathbf{x}_k \circ \theta^{i*} = \mathbf{x}_i \circ \theta^{i*}$, i.e., $k \in \text{MG}(\theta^{i*}, \mathbf{x}_i)$, we have:*

$$k \in \operatorname{argmin}_{\substack{l=1, \dots, n \\ z_l \neq z_i}} \mathbf{w}^T \mathbf{1}_{[\mathbf{x}_i \neq \mathbf{x}_l]}. \quad (2)$$

In particular, if θ^{i*} has all entries equal to one and $k \in \text{MG}(\theta^{i*}, \mathbf{x}_i)$ then $\mathbf{w}^T \mathbf{1}_{[\mathbf{x}_i \neq \mathbf{x}_k]} = 0$.

The detailed derivation of this lemma is in the supplement. This statement clarifies that by solving the AME-IV problem, we minimize the weighted hamming distance between each unit i and all other units with a different assignment of the instrument that belong to i 's main matched group. We now introduce a smoothness assumption under which we can formally link the matched groups created by AME-IV with the necessary conditions for causal estimation using instrumental variables.

(A5) Smoothness: For any two $\mathbf{x}_i, \mathbf{x}_k \in \{0, 1\}^p$, and $\delta > 0$, we have: $\mathbf{w}^T \mathbf{1}_{[\mathbf{x}_i \neq \mathbf{x}_k]} \leq \delta \implies |p(Z_i = 1|\mathbf{x}_i) - p(Z_k = 1|\mathbf{x}_k)| \leq \epsilon(\delta)$, where $\epsilon(\delta)$ is an increasing function of δ such that $\epsilon(0) = 0$.

Note that this is a variant of a standard assumption made in most matching frameworks (see, e.g., Rosenbaum, 2010). The following proposition follows immediately from Lemma 3.1 applied to A5.

Proposition 3.1. *If $k \in \text{MG}(\theta^{i*}, \mathbf{x}_i)$ with $z_i \neq z_k$, and A5 holds, then*

$$|P(Z_i = 1|\mathbf{x}_i) - P(Z_k = 1|\mathbf{x}_k)| \leq \epsilon \left(\min_{\substack{l=1, \dots, n \\ z_l \neq z_i}} \mathbf{w}^T \mathbf{1}_{[\mathbf{x}_i \neq \mathbf{x}_l]} \right).$$

In particular, if θ^{i*} is one in all entries, then $\Pr(Z_i = 1|\mathbf{x}_i) = \Pr(Z_k = 1|\mathbf{x}_k)$.

With this observation, we know that units matched together will have similar probabilities of being instrumented (in fact, as similar as possible, as finite data permits). This will allow us to produce reliable estimates of λ using our matched groups, provided that the data actually contain matches of sufficiently high quality.

3.2 FULL AME-IV PROBLEM

In the full version of the AME-IV problem, the weights are chosen so that the variables used for each matched group have a useful quality: these variables together can create a high-quality predictive model for the outcomes. The weights become variable importance measures for each of the variables.

In order to determine the importance of each variable j , we use variable importance techniques to analyze machine learning models trained on a separate training set. Specifically, the units $1, \dots, n$ are divided into a training and a holdout set, the first is used to create matched groups and estimate causal quantities, and the second to learn the importance of each of the variables for match quality. Formally define the empirical predictive error on the training set, for set of variables θ as:

$$\widehat{\text{PE}}_{\mathcal{F}}(\theta) = \min_{f \in \mathcal{F}} \sum_{a \in \text{training}} (f(\theta \circ \mathbf{x}_a^{tr}, z_a^{tr}) - y_a^{tr})^2,$$

where \mathcal{F} is some class of prediction functions. The empirical predictive error measures the usefulness of a set of variables. (The set of variables being evaluated are the ones highlighted by the indicator variables θ .)

We ensure that we always match using sets of variables θ that together have a low error $\widehat{\text{PE}}_{\mathcal{F}}$. In fact, for each unit, if we cannot match on all the variables, we will aim to match on the set of variables for which the lowest possible prediction error is attained. Because of this, all matched groups are matched on a set of variables that together can predict outcomes sufficiently well.

The Full-AME-IV problem can thus be stated as: for all instrumented units i ,

$$\begin{aligned} \theta^{i*} &\in \operatorname{argmin}_{\theta \in \{0, 1\}^p} \widehat{\text{PE}}_{\mathcal{F}}(\theta), \text{ such that:} \\ &\exists k \text{ with } z_k = 0 \text{ and } \mathbf{x}_k \circ \theta^{i*} = \mathbf{x}_i \circ \theta^{i*}, \end{aligned}$$

When importance weights are a linear function of the covariates, then solving the problem above is equivalent to solving the general AME-IV problem. An analogous result holds without IVs for the AME problem (Wang et al., 2019).

In the standard Full-AME problem, there is no instrument, and each matched group must contain *both treatment and control* units, whereas in the Full-AME-IV case, the key is to match units so that instrumented units are matched with non-instrumented units *regardless of treatment*. Intuitively, this makes sense because treatment uptake is in itself an outcome of instrumentation in the IV framework: a group with very large or very small numbers of treated or control units would imply that units with certain values of \mathbf{x} are either highly likely or highly unlikely to respond to the instrument by taking up the treatment.

3.3 FLAME-IV: AN APPROXIMATE ALGORITHM FOR THE FULL-AME-IV PROBLEM

We extend ideas from the Fast Large-scale Almost Matching Exactly (FLAME) algorithm introduced by Wang et al. (2019) to approximately solve the AME-IV problem. Our algorithm – FLAME-IV – uses instrumental variables to create matched groups that have at least one instrumented and one non-instrumented unit within them. The procedure starts with an exact matching that finds all exact main matched groups. Then at each iteration FLAME-IV iteratively chooses one covariate to drop, and creates matched groups on the remaining covariates. To decide which covariate to drop at each iteration, FLAME-IV loops through the possibilities: it temporarily drops one covariate and computes the *match quality* MQ after dropping this covariate. Then FLAME-IV selects the covariate for which MQ was maximized during this loop. Match quality MQ is defined as a trade-off between prediction error, \widehat{PE} (which is defined in Section 3.2) and a balancing factor, which is defined as:

$$BF = \frac{\# \text{ matched non-instrumented}}{\# \text{ available non-instrumented}} + \frac{\# \text{ matched instrumented}}{\# \text{ available instrumented}}$$

MQ is computed on the holdout training dataset. In practice, the balancing factor improves the quality of matches by preventing FLAME-IV from leaving too many units stranded without matched groups. That is, it could prevent all treated units from being matched to the same few control units when more balanced matched groups were possible. More details about the FLAME-IV algorithm are in the supplement.

It is recommended to early-stop the algorithm before the MQ drops by 5% or more (Wang et al., 2019). This way, the set of variables defining each matched group is sufficient to predict outcomes well (on the training set). The details about early-stopping are in the supplement.

4 ESTIMATION

Assuming that (A1) through (A5) and SUTVA hold, the LATE, λ , can be estimated in a consistent way (Imbens and Angrist, 1994; Angrist et al., 1996); in this section we adapt common estimators for λ to our matching framework. Consider a collection of m matched groups, MG_1, \dots, MG_m , each associated with a different value of (θ, \mathbf{x}) . We estimate the average causal effect of the instrument on the treatment, $\widehat{ITT}_{t,\ell}$ and on the outcome, $\widehat{ITT}_{y,\ell}$, within each matched group, ℓ , and then take the ratio of their weighted sums over all groups to estimate λ .

We start with the canonical estimator for $\widehat{ITT}_{y,\ell}$:

$$\widehat{ITT}_{y,\ell} = \frac{\sum_{i \in MG_\ell} y_i z_i}{\sum_{i \in MG_\ell} z_i} - \frac{\sum_{i \in MG_\ell} y_i (1 - z_i)}{\sum_{i \in MG_\ell} (1 - z_i)}. \quad (3)$$

Similarly, the estimator for the causal effect of the instrument on the treatment, $\widehat{ITT}_{t,\ell}$, can be written as:

$$\widehat{ITT}_{t,\ell} = \frac{\sum_{i \in MG_\ell} t_i z_i}{\sum_{i \in MG_\ell} z_i} - \frac{\sum_{i \in MG_\ell} t_i (1 - z_i)}{\sum_{i \in MG_\ell} (1 - z_i)}. \quad (4)$$

From the form of λ in Equation (1) it is easy to see that, if the estimators in (3) and (4) are unbiased for $\widehat{ITT}_{y,\ell}$ and $\widehat{ITT}_{t,\ell}$ respectively (which is true, for instance, when matches are made exactly for all units), then the ratio of their weighted average across all matched groups is a consistent estimator for λ :

$$\hat{\lambda} = \frac{\sum_{\ell=1}^m n_\ell \widehat{ITT}_{y,\ell}}{\sum_{\ell=1}^m n_\ell \widehat{ITT}_{t,\ell}}, \quad (5)$$

where n_ℓ denotes the number of units in matched group ℓ . A natural extension of this framework allows us to estimate the LATE within matched group ℓ , defined as:

$$\lambda_\ell = \frac{1}{n_\ell} \sum_{\substack{i \in MG_\ell: \\ t_{i\ell}(1) > t_{i\ell}(0)}} y_i(1) - y_i(0). \quad (6)$$

This can be accomplished with the following estimator:

$$\hat{\lambda}_\ell = \frac{\widehat{ITT}_{y,\ell}}{\widehat{ITT}_{t,\ell}}. \quad (7)$$

We quantify uncertainty around our estimates with asymptotic Confidence Intervals (CIs). To compute CIs for these estimators we adapt the approach laid out in Imbens and Rubin (2015). Details on variance estimators and computations are given in the supplement.

In the following section, we present simulations that employ these estimators in conjunction with the algorithms presented in the previous section to estimate λ and λ_ℓ . The performance of our methodology is shown to surpass that of other existing approaches.

5 SIMULATIONS

We evaluate the performance of our method using simulated data. We compare our approach to several other methods including **two-stage least squares** (Angrist and Keueger, 1991; Card, 1993; Wooldridge, 2010), and two other state-of-the-art nonparametric methods for instrumental variables, **full matching** (Kang et al., 2016) and **nearfar matching** (Baiocchi et al., 2010). Full matching and nearfar matching find units that differ on the instrument while being close in covariate space according to a predefined distance metric. Both algorithms rely on a sample-rank Mahalanobis distance with an instrument propensity score caliper. We implement FLAME-IV using *bit-vector calculations*. More details about the implementation are in the supplementary materials.

In the first set of experiments, we compare the performance of the different methods on the estimation of local average treatment effects. In Experiment 5.2 we demonstrate the power of FLAME-IV for estimating individualized local average treatment effects. Experiment 5.3 describes the scalability of the approach in terms of the number of covariates and number of units.

Throughout, we generate instruments, covariates and continuous exposures based on the following structural equation model (Wooldridge, 2010):

$$T_i^* = k + \pi Z_i + \rho^T X_i + \xi_i \quad (8)$$

where $Z_i \sim \text{Bernoulli}(0.5)$, and $\xi_i \sim N(0, 0.8)$. For important covariates, $X_{ij} \sim \text{Bernoulli}(0.5)$. For unimportant covariates, $X_{ij} \sim \text{Bernoulli}(0.1)$ in the control group, and $X_{ij} \sim \text{Bernoulli}(0.9)$ in the treatment group. We discretize the exposure values T_i^* by defining:

$$\begin{aligned} T_i &= \mathbf{I}_{[0.3 < T_i^* \leq 0.6]} \\ &+ 2 \times \mathbf{I}_{[0.6 < T_i^* \leq 1.0]} + 3 \times \mathbf{I}_{[T_i^* > 1.0]}. \end{aligned}$$

5.1 ESTIMATION OF λ

In this experiment, outcomes are generated based on one of two homogeneous treatment effect models: a linear and a nonlinear model, respectively defined as:

$$Y_i = \sum_{j=1}^{10} \alpha_j X_{ij} + 10T_i \quad (9)$$

$$Y_i = \sum_{j=1}^{10} \alpha_j X_{ij} + 10T_i + \sum_{1 \leq j < \gamma \leq 5} X_{ij} X_{i\gamma}. \quad (10)$$

Under both generation models, the true treatment effect is 10 for all individuals. There are 10 confounding covariates, 8 of which are important and 2 are unimportant.

The importance of the variables is exponentially decaying with $\alpha_j = 0.5^j$.

We measure performance using the **absolute bias of the median**, i.e., the absolute value of the bias of the median estimate of 500 simulations and **median absolute deviation**, i.e., the median of the absolute deviations from the true effect, for each simulation. We present simulation results at varying levels of strength of the instrumental variable. This is measured by a concentration parameter, defined as the influence that the instrument has on treatment take-up. This is represented by the concentration parameter π in Eq. (8). Usually a concentration parameter below 10 suggests that instruments are weak (Stock et al., 2002).

We also assess the performance of our methods by varying the size of training and holdout data. We generate two training and holdout datasets of different sizes: one with 1000 instrumented units and 1000 non-instrumented units, and one with 50 instrumented units and 50 non-instrumented units. For each case, we run each experiment 500 times for each of the algorithms.

Figures 2 and 3 show the results of this experiment. All algorithms achieve better estimation accuracy when the instrument is stronger (i.e., more instrumented units take up the treatment). Figure 2 shows results for the linear generation model, and Figure 3 shows results for the nonlinear generation model. As both figures show, FLAME-IV with and without early-stopping generally outperform all other algorithms in terms of bias and deviation. This is likely because our methodology does not rely on a parametric outcome model and uses a discrete learned distance metric. The only exceptions are the left-upper plot on Figure 2 and Figure 3, which represents the bias results on small datasets (50 instrumented & 50 noninstrumented). 2SLS has advantages here, because the amount of data is too small for powerful nonparametric methods like FLAME-IV to fit reliably. FLAME-IV's matching estimates lead to slightly larger bias than 2SLS. In the supplementary materials, we report results of similar experiments but with the additional inclusion of observed confounders of instrument assignment. We see no degradation in the performance. Result patterns with confounded instruments mimic those in Figures 2 and 3.

Next, we compare 95% confidence intervals for each algorithm. The results are reported in Table 1. FLAME-IV performs well on the nonlinear generation model, leading to the narrowest 95% CI of all the methods. For the linear generation model, the 95% CI for FLAME-IV is narrower than the equivalent CIs for full matching and nearfar matching, but wider than 2SLS. Again, this is expected, and due to the correct parameterization of 2SLS with the linear generation model. More details about the

confidence intervals are available in the supplement.

5.2 ESTIMATION OF λ_ℓ

One advantage of the AME-IV methodology is that it allows us to estimate LATE's on compliers (units for whom $t_i(1) > t_i(0)$) within each matched group. This results in more nuanced estimates of the LATE and in overall better descriptions of the estimated causal effects. We evaluate performance of FLAME-IV in estimating matched group-level effects in a simulation study, with the estimators described in Section 4. To study how well FLAME-

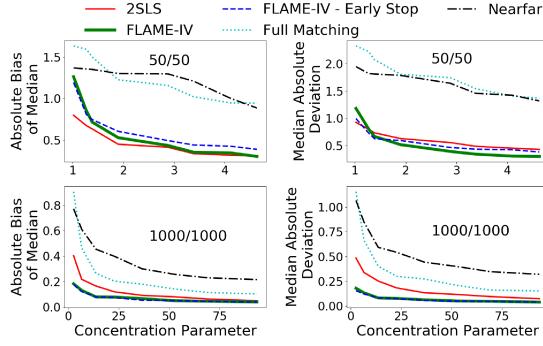


Figure 2: Performance for linear generation model with various sample sizes. Here, 2SLS has an advantage because the data are generated according to a 2SLS model. FLAME-IV (either early-stopping or run-until-no-more-matches) outperforms other methods on the large dataset, with smaller absolute bias of the median and median absolute deviation. On the smaller datasets, FLAME-IV has a slightly larger bias than 2SLS but the smallest median absolute deviation among all methods.

IV estimates individual causal effects, we generate data with heterogeneous treatment effects. The new generation models, (11) and (12) below, are unlike the generation models in (9) and (10), in that different individuals have different treatment effects. The two heterogeneous treatment effect data generation models are:

$$Y_i = \sum_{j=1}^T \alpha_j X_{ij} + T_i \sum_{j=1}^{10} \beta_j X_{ij} \quad (11)$$

$$Y_i = \sum_{j=1}^T \alpha_j X_{ij} + T_i \sum_{j=1}^{10} \beta_j X_{ij} + \sum_{\substack{j=1 \dots 5 \\ \gamma=1 \dots 5 \\ \gamma > j}} X_{ij} X_{i\gamma}. \quad (12)$$

Here $\alpha_i \sim N(10s, 1)$ with $s \sim \text{Uniform}\{-1, 1\}$, $\beta_j \sim N(1.5, 0.15)$. We generate 1000 treatment and 1000 control units from both models. We increased the value of the concentration parameter π in Eq. (8) so that Z has a strong effect on T for the whole dataset. This is done to ensure appropriate treatment take-up within each group. Even with this adjustment, a few groups did not have any units take up treatment in the simulation. Results for these

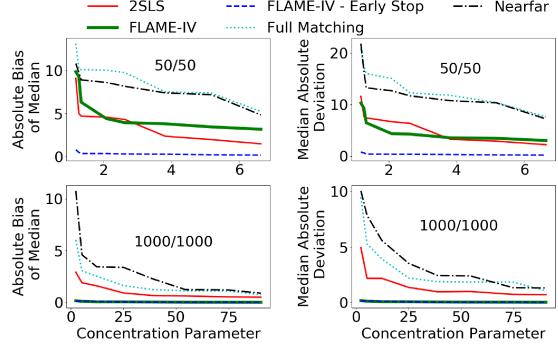


Figure 3: Performance for nonlinear generation model with different sample sizes. Here, the 2SLS model is misspecified. FLAME-IV (either early-stop or run-until-no-more-matches) outperforms other methods on both datasets, having smaller absolute bias of median and median absolute deviation.

groups were not computed and are not reported in Figure 4. We estimate the LATE within each matched group (λ_ℓ). Note that in groups where the instrument is very strong, the LATE will approximately equal the average treatment effect on the treated. Experimental results for both data generation models are shown in Figure 4. As we can see, our estimated effects almost align with true treatment effects and lead to relatively small estimation error for both linear and nonlinear generation models. Our algorithm performs slightly better when the generation model is linear.

5.3 RUNNING TIME EVALUATION

For the synthetic data generated by Section 5.2, Figure 5 compares the runtime of our algorithm against full matching. We computed the runtime by varying number of units (Figure 5, left panel) and by varying number of covariates (Figure 5, right panel). Each runtime is the average of five experiment results. The plot suggests that our algorithm scales well with both the number of units and number of covariates. Full matching depends on a Mahalanobis distance metric, which is costly to compute in terms of time. FLAME-IV scales even better than full matching on a larger dataset with more units or covariates. Experimental results about larger datasets are in the supplement. We note that the maximum number of units and covariates of full matching is also limited to the maximum size of vectors in R. Experiments were run on an Intel Core i7-4790 @ 3.6 GHz with 8 GB RAM and Ubuntu 16.04.01.

6 WILL A FIVE-MINUTE DISCUSSION CHANGE YOUR MIND?

In this section, we demonstrate the practical utility of our method by applying it to a real-world dataset. Since

Table 1: Point Estimates for Linear and Nonlinear Models

| | FLAME-IV | 2SLS | Full-Matching | Nearfar Matching |
|-----------------|------------------------|------------------------|-------------------------|-------------------------|
| Linear Model | 10.15 (9.72, 10.58) | 10.16 (9.92, 10.40) | 10.96 (10.14, 12.68) | 11.23 (10.23, 12.89) |
| Nonlinear Model | 9.95 (9.47, 10.43) | 10.11 (6.96, 13.25) | 18.97 (11.35, 41.44) | 21.67 (12.96, 45.71) |

95% confidence interval for each estimate is given in parentheses. The value of concentration parameter for linear model is 36.64, whereas the same for nonlinear model is 15.57.

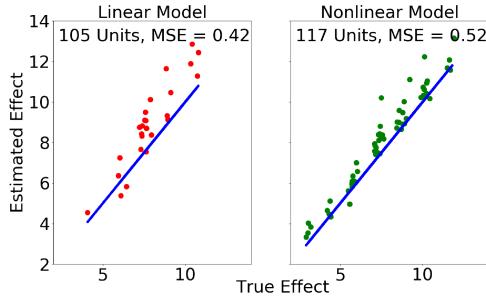


Figure 4: True Individual Causal Effect vs. Estimated Individual Causal Effect. The numbers on each plot represent the total number of instrumented units for calculating unit-level LATE, and MSE of our predictions. The concentration parameter is the same for the whole dataset, set to 288.84 for the linear outcome model, and 272.92 for the nonlinear outcome model.

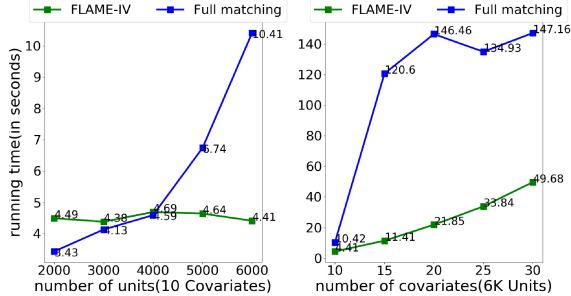


Figure 5: Running Time for FLAME-IV and Full Matching. Left panel presents run time by varying number of units, whereas the right panel presents run time by varying number of covariates.

we do not observe the ground truth, we cannot evaluate the performance in terms of predictions, instead, we determine whether we can replicate the results of a previously published study. Specifically, we examine how door-to-door canvassing affects actual electoral outcomes; using experimental data generated by running a country-wide experiment during the 2012 French general election (Pons, 2018). The original study estimates the effects of a door-to-door campaign in favor of François Hollande’s *Parti Socialiste* (PS) on two outcomes: voter turnout and share of votes for PS. The two outcomes are measured twice: once for each of the two rounds of voting that took

place during the 2012 election. The units of analysis are geographically defined electoral precincts, often, but not always, comprised of different municipalities.

The instrument in this case is pre-selection into campaign precincts: the 3,260 electoral precincts were clustered into strata of 5, among which 4 were randomly chosen and made available to conduct a campaign. The treatment is the decision of which of these four instrumented precincts to actually run campaigns in, as not all of the four instrumented precincts were actually chosen for door-to-door campaigns. The decision was based on the proportion of PS votes at the previous election within each precinct and the target number of registered citizens for each territory. These deciding factors evidently confound the causal relationship between treatment and outcomes. This setup provides an ideal setting for an Instrumental Variable design, where random pre-selection into campaign districts can be used to estimate the LATE of actual door-to-door campaigns on both turnout and PS vote share.

We replicate the original study’s results by running our algorithm on the data without explicitly accounting for the strata defined by the original experiment. Since some of the covariates used for matching are continuous, we coarsen them into 5 ordinal categories. We coarsen turnout at the previous election and PS vote share at the previous election into 10 categories instead, as these variables are particularly important for matching and we would like to make more granular matches on them. Results from applying our methods to the data from the study are presented in Table 2. Columns 2 and 3 shows results for PS vote share as an outcome, and the last two columns for voter turnout as an outcome. Results are presented disaggregated by each round of election.

Panel A provides LATE estimates from FLAME-IV. Unlike the earlier study (Pons, 2018), our estimates are independent of the strong parametric assumptions of 2SLS. We reach conclusions similar to those of the original paper, finding no positive effect of canvassing on voter turnout and a positive statistically significant effect on vote share for PS. In general, our standard error estimates are similar to those obtained with 2SLS, however more conservative due to the non-parametric nature of the estimators we

Table 2: Effect of Door-to-Door Canvassing on Electoral Outcomes

| | Vote Share | | Voter Turnout | |
|--|----------------------|-----------------------|-----------------------|------------------------|
| | First round | Second round | First round | Second round |
| <i>Panel A: All Precincts</i> | | | | |
| | 0.02280 (0.00683) | 0.01593 (0.00827) | -0.00352 (0.00163) | -0.00634 (0.00158.) |
| <i>Panel B: Precincts by Income Levels</i> | | | | |
| Low | 0.02844 (0.00429) | 0.03903 (0.00562) | -0.00666 (0.00228) | -0.01505 (0.00254) |
| Medium | 0.01772 (0.00388) | 0.02090 (0.00434) | -0.00311 (0.00287) | -0.00070 (0.00333) |
| High | 0.02560 (0.02780) | 0.04313 (0.02752) | -0.02717 (0.01217) | -0.01367 (0.00538) |
| <i>Panel C: Precincts by Gender Majority</i> | | | | |
| Male | 0.05619 (0.00879) | -0.00442 (0.00995) | 0.00973 (0.00376) | -0.00056 (0.00346) |
| Female | 0.01640 (0.00834) | 0.00777 (0.00719) | -0.00692 (0.00237) | -0.00675 (0.00239) |

Columns 2 and 3 correspond to causal effects on vote share for PS, whereas Columns 4 and 5 reports causal effects on voter turnout. Panel A accounts for all the precincts and reports population causal effects. Panel B divides precincts by median income level and reports causal effect for each subgroup. Panel C divides precincts by gender-majority and reports associated causal effects. We use 15% of the data as holdout training data and use a 5% change in match quality as an early stopping rule. Differences between our approach and the original paper’s approach in estimated variances are mainly due to the strata used by the authors being marginally different from those produced with our methodology.

employ. Interestingly, our estimate of the effect of canvassing on vote share has a *greater* magnitude than the original analysis, while our estimate for the effect of canvassing on voter turnout is nearly the same as the original paper’s.

Our methodology also allows an improvement on the original analysis by estimating effects of door-to-door campaigns on the two outcomes for particular subgroups of interest. LATE estimates for income and gender subgroups are reported in *Panel B* and *Panel C* of Table 2. The income subgroups are defined by median income, whereas gender subgroups are defined by share of female population in each precinct. We find that canvassing was more effective in increasing the vote share for PS, in the first round of the election, in precincts where male population is in the majority. We also find that canvassing had negative effect on voter turnout in low income precincts, but positive effect on voter share for PS. The combination of these results show that canvassing was successful in convincing voters to switch their votes in favour of François Hollande.

In the supplement we show two example matched groups output by FLAME-IV. In this case the algorithm was successful in separating localities with low support for PS from localities in which support for PS was greater. These examples highlight how the algorithm can produce meaningful and interpretable groups, while reducing potential for confounding by observed covariates.

In conclusion, the results of our analysis of the voter turnout data clearly show that our method produces novel and interesting results when applied to real-world scenarios,

independently of strong parametric assumptions, and with a simple interpretable framework.

7 CONCLUSION

Matching methods can be extremely powerful: they are both highly nonparametric and interpretable to users, allowing them to trust and troubleshoot their models more easily. Our approach to matching for instrumental variables accounts for the limitations faced by existing methods. We improve on 2SLS by using a highly non-parametric powerful modeling approach. We retain interpretability unlike traditional machine learning approaches by using matching. We improve on existing matching methods by *learning* an interpretable distance metric on a training set. Our methodology also provides a systematic way to account for nuisance variables, and to achieve consistently high quality matching outcomes. The algorithm can be implemented easily within most common database systems for optimal performance. It scales well to large datasets. It achieves a balance between interpretability, scalability, trustworthiness, and modeling power that is unsurpassed by any other method for IV analysis. Code is publicly available at: <https://github.com/almost-matching-exactly>

Acknowledgements

This work was supported in part by NIH award 1R01EB025021-01, NSF awards IIS-1552538 and IIS1703431, a DARPA award under the L2M program, and a Duke University Energy Initiative Energy Research Seed Fund (ERSF) grant.

References

Alberto Abadie. Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics*, 113(2):231–263, 2003.

Daron Acemoglu, Simon Johnson, and James A Robinson. The colonial origins of comparative development: An empirical investigation. *American Economic Review*, 91(5):1369–1401, 2001.

Joshua D Angrist and Alan B Keueger. Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, 106(4):979–1014, 1991.

Joshua D Angrist, Guido W Imbens, and Donald B Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455, 1996.

David H. Autor, David Dorn, and Gordon H. Hanson. The china syndrome: Local labor market effects of import competition in the united states. *American Economic Review*, 103(6):2121–68, 2013.

Mike Baiocchi, Dylan S Small, Scott Lorch, and Paul R Rosenbaum. Building a stronger instrument in an observational study of perinatal care for premature infants. *Journal of the American Statistical Association*, 105(492):1285–1296, 2010.

Carlos Brito and Judea Pearl. Generalized instrumental variables. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, pages 85–93. Morgan Kaufmann Publishers Inc., 2002.

David Card. Using geographic variation in college proximity to estimate the return to schooling. Technical report, National Bureau of Economic Research, 1993.

Bryant Chen, Judea Pearl, and Elias Bareinboim. Incorporating knowledge into structural equation models using auxiliary variables. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 3577–3583. AAAI Press, 2016.

Awa Dieng, Yameng Liu, Sudeepa Roy, Cynthia Rudin, and Alexander Volfovsky. Interpretable almost-exact matching for causal inference. In *Proceedings of Artificial Intelligence and Statistics (AISTATS)*, pages 2445–2453, 2019.

Markus Frölich. Nonparametric IV estimation of local average treatment effects with covariates. *Journal of Econometrics*, 139(1):35–75, July 2007.

Alan S Gerber and Donald P Green. The effects of canvassing, telephone calls, and direct mail on voter turnout: A field experiment. *American Political Science Review*, 94(3):653–663, 2000.

Jason Hartford, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. Deep IV: A flexible approach for counterfactual prediction. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 1414–1423, 2017.

Hidehiko Ichimura and Christopher Taber. Propensity-score matching with instrumental variables. *American Economic Review*, 91(2):119–124, 2001.

Guido W Imbens and Joshua D Angrist. Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475, Mar. 1994.

Guido W Imbens and Donald B Rubin. Bayesian inference for causal effects in randomized experiments with noncompliance. *The Annals of Statistics*, pages 305–327, 1997.

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

Paul L Joskow. Contract duration and relationship-specific investments: Empirical evidence from coal markets. *The American Economic Review*, pages 168–185, 1987.

Hyunseung Kang, Benno Kreuels, Ohene Adjei, Ralf Krumkamp, Jürgen May, and Dylan S Small. The causal effect of malaria on stunting: a mendelian randomization and matching approach. *International Journal of Epidemiology*, 42(5):1390–1398, 2013.

Hyunseung Kang, Benno Kreuels, Jürgen May, Dylan S Small, et al. Full matching approach to instrumental variables estimation with application to the effect of malaria on stunting. *The Annals of Applied Statistics*, 10(1):335–364, 2016.

Whitney K Newey and James L Powell. Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578, 2003.

Elizabeth L Ogburn, Andrea Rotnitzky, and James M Robins. Doubly robust estimation of the local average treatment effect curve. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(2):373–396, 2015.

Vincent Pons. Will a five-minute discussion change your mind? a countrywide experiment on voter choice in france. *American Economic Review*, 108(6):1322–63, 2018.

Paul R Rosenbaum. *Design of observational studies*, volume 10. Springer, 2010.

Ricardo Silva and Shohei Shimizu. Learning instrumental variables with structural and non-gaussianity assumptions. *Journal of Machine Learning Research*, 18(120):1–49, 2017. URL <http://jmlr.org/papers/v18/17-014.html>.

James H Stock, Jonathan H Wright, and Motohiro Yogo. A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics*, 20(4):518–529, 2002.

Zhiqiang Tan. Regression and weighting methods for causal inference using instrumental variables. *Journal of the American Statistical Association*, 101(476):1607–1618, 2006.

Tianyu Wang, Marco Morucci, M. Usaid Awan, Yameng Liu, Sudeepa Roy, Cynthia Rudin, and Alexander Volfovsky. FLAME: A fast large-scale almost matching exactly approach to causal inference. *arXiv:1707.06315*, 2019.

Jeffrey M Wooldridge. *Econometric analysis of cross section and panel data*. MIT press, 2010.