Interpretable Almost-Matching Exactly with Instrumental Variables

by

Yameng Liu

Department of Computer Science
Duke University

Date: _____

Approved:

_____
Cynthia Rudin, Co-Supervisor

_____
Sudeepa Roy, Co-Supervisor

_____
Alexander Volfovsky

A thesis submitted in partial fulfillment of the
requirements for the degree of Master of Science
in the Department of Computer Science
in the Graduate School of
Duke University

2019

ABSTRACT

Interpretable Almost-Matching Exactly with Instrumental Variables

by

Yameng Liu

Department of Computer Science
Duke University

Date: _____

Approved:

_____
Cynthia Rudin, Co-Supervisor

_____
Sudeepa Roy, Co-Supervisor

_____
Alexander Volfovsky

An abstract of a thesis submitted in partial fulfillment of the
requirements for the degree of Master of Science
in the Department of Computer Science
in the Graduate School of
Duke University

2019

Abstract

We aim to create the highest possible quality of treatment-control matches for categorical data in the potential outcomes framework. The method proposed in this work aims to match units on a weighted Hamming distance, taking into account the relative importance of the covariates; To match units on as many relevant variables as possible, the algorithm creates a hierarchy of covariate combinations on which to match (similar to downward closure), in the process solving an optimization problem for each unit in order to construct the optimal matches. The algorithm uses a single dynamic program to solve all of the units' optimization problems simultaneously. Notable advantages of our method over existing matching procedures are its high-quality interpretable matches, versatility in handling different data distributions that may have irrelevant variables, and ability to handle missing data by matching on as many available covariates as possible. We also adapt the matching framework by using instrumental variables (IV) to the presence of observed categorical confounding that breaks the randomness assumptions and propose an approximate algorithm which speedily generates high-quality interpretable solutions.We show that our algorithms construct better matches than other existing methods on simulated datasets, produce interesting results in applications to crime intervention and political canvassing.

Contents

Abstract **iv**

List of Tables **viii**

List of Figures **ix**

1 Introduction **1**

2 Related Work **6**

3 Almost-Matching Exactly (AME) Framework **9**

4 The Dynamic AME (DAME) Algorithm **13**

5 Almost-Matching Exactly with Instrumental Variables (AME-IV) **19**

5.1 Basic Instrumental Variables Assumptions 19

5.2 Almost-Matching Exactly with Instrumental Variables (AME-IV Problem) 24

5.3 Full AME-IV Problem 27

5.4 FLAME-IV: An Approximate Algorithm for the Full-AME-IV Problem 29

5.4.1 Estimation 31

6 Simulations **33**

6.1 Simulations with DAME Algorithm 33

6.1.1 Presence of Irrelevant Covariates 34
6.1.2 Exponentially Decaying Covariates 35
6.1.3 Imbalanced Data 36
6.1.4 Run Time Evaluation 36
6.2 Simulations with FLAME-IV Algorithm 38
6.2.1 Implementation of FLAME-IV 39
6.2.2 Estimation of λ 40
6.2.3 Estimation of λ_ℓ 43
6.2.4 Running Time Evaluation 46

7 Real Data Experiment **49**

7.1 Breaking the Cycle of Drugs and Crime in the United States 49
7.2 Will a Five-Minute Discussion Change Your Mind? 50

8 Appendix **55**

8.1 Naïve AME solutions 55
8.2 Proof of Proposition 4.0.1 57
8.3 Proof of Theorem 4.0.2 57
8.4 Proof of Lemma 5.2.1 60
8.5 Asymptotic Variance and Confidence Intervals for LATE Estimates . 61

8.6 The FLAME-IV Algorithm 66

8.7 Implementation of `GroupedMR` using Database (SQL) Queries 70

8.8 Implementation of `GroupedMR` using Bit Vectors 72

8.9 Sample Matched Groups 74

Bibliography **76**

List of Tables

6.1 MSE for different imbalance ratios 36

6.2 Point Estimates for Linear and Nonlinear Models. 42

7.1 Effect of Door-to-Door Canvassing on Electoral Outcomes. 51

8.1 Example population table illustrating the *bit-vector* implementation. 73

8.2 Two sample matched groups generated by FLAME on the application data described in Section 8.9. 75

List of Figures

5.1 Causal DAG for instrumental variables. 22

6.1 Estimated CATT vs. True CATT. DAME and FLAME perfectly estimate the CATTs before dropping important covariates. 35

6.2 DAME makes higher quality matches early on. DAME matches on more covariates than FLAME, yielding lower MSE from matched groups. 37

6.3 Run-time comparison between DAME FLAME, and brute force. 38

6.4 Performance for linear generation model with various sample sizes. Here, 2SLS has an advantage because the data are generated according to a 2SLS model. 43

6.5 Performance for nonlinear generation model with different sample sizes. Here, the 2SLS model is misspecified. 44

6.6 True Individual Causal Effect vs. Estimated Individual Causal Effect. 47

6.7 Running Time for FLAME-IV and Full Matching. 48

8.1 Running Time for FLAME-IV on large dataset. 73

Chapter 1

Introduction

In observational causal inference where the scientist does not control the randomization of individuals into treatment, an ideal approach matches each treatment unit to a control unit with identical covariates. However, in high dimensions, few such "identical twins" exist, since it becomes unlikely that any two units have identical covariates in high dimensions. In that case, how might we construct a match assignment that would lead to accurate estimates of conditional average treatment effects (CATEs)?

For categorical variables, we might choose a Hamming distance to measure similarity between covariates. Then, the goal is to find control units that are similar to the treatment units on as many covariates as possible. However, the fact that not all covariates are equally important has serious implications for CATE estimation. Matching methods generally suffer in the presence of many irrelevant covariates (covariates that are not related to either treatment or outcome): the irrelevant variables would dominate the Hamming distance calculation, so that the treatment units would mainly be matched to the control units on the irrelevant variables. This

means that matching methods do not always pass an important sanity check in that irrelevant variables should be irrelevant. To handle this issue with irrelevant covariates, in this work we choose to match units based on a *weighted* Hamming distance, where the weights can be learned from machine learning on a hold-out training set. These weights act like variable importance measures for defining the Hamming distance.

The choice to optimize matches using Hamming distance leads to a serious computational challenge: how does one compute optimal matches on Hamming distance? In this work, we define a matched group for a given unit as the solution to a constrained discrete optimization problem, which is to find the weighted Hamming distance of each treatment unit to the nearest control unit (and vice versa). There is one such optimization problem for each unit, and we solve all of these optimization problems efficiently with a single dynamic program. Our dynamic programming algorithm has the same basic monotonicity property (downwards closure) as that of the apriori algorithm [AS94] used in data mining for finding frequent itemsets. However, frequency of itemsets is irrelevant here, instead the goal is to find a largest (weighted) set of covariates that both a treatment and control unit have in common. The algorithm, Dynamic Almost Matching Exactly – `DAME` – is efficient, owing to the use of bit-vector computations to match units in groups, and does not require an integer programming solver.

A more general version of our formulation (Full Almost Matching Exactly) adaptively chooses the features for matching in a data-driven way. Instead of using a fixed weighted Hamming distance, it uses the hold-out training set to determine how useful a *set* of variables is for prediction out of sample. For each treatment unit, it finds a set of variables that (i) allows a match to at least one control unit; (ii) together have the best out-of-sample prediction ability among all subsets of variables for which a match can be created (to at least one control unit). Again, even though for each unit we are searching for the best subset of variables, we can solve all of these optimization problems at once with our single dynamic program.

In this research, we also study the case where the randomness assumption of treatment assignment is broken in practical 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., [Jos87, GG00, AJR01, DDH13]).

The most popular existing method that uses instrumental variables to conduct

causal inference is Two-Stage Least Squares Regression (2SLS) [AK91, Car93, Woo10]. 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 [NP03, Frö07], 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 [KKM+16]: 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 of potentially different type. 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 matching groups. We create an Almost-Matching Exactly framework [WRVR17] 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 non-linear models. It improves over existing matching methods for instrumental variables when covariates are discrete, as it

employs an appropriate distance metric for this type of data. Finally, our proposed method 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 (FLAME-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.

Chapter 2

Related Work

As mentioned earlier, exact matching is not possible in high dimensions, as "identical twins" in treatment and control samples are not likely to exist. Early on, this led to techniques that reduce dimension using propensity score matching [Rub73b, Rub73a, Rub76, CR73], which extend to penalized regression approaches [SRG+09, RS12, BCH14, Far15]. Propensity score matching methods project the entire dataset to one dimension and thus cannot be used for estimating CATE (conditional average treatment effect), since units within the matched groups often differ on important covariates. In "optimal matching," [Ros16], an optimization problem is formed to choose matches according to a pre-defined distance measure, though as discussed above, this distance measure can be dominated by irrelevant covariates, leading to poor matched groups and biased estimates. Coarsened exact matching [IKP12a, IKP11] has the same problem, since again, the distance metric is pre-defined, rather than learned. Recent integer-programming-based methods considers extreme matches for all possible reasonable distance metrics, but this is computationally expensive and relies on manual effort to create the ranges [MNEAR18, NEAR15c]; in

contrast we use machine learning to create a single good match assignment.

In the framework of *almost-matching exactly* [WRRV17], each matched group contains units that are close on covariates that are important for predicting outcomes. For example, Coarsened Exact Matching [IKP12a, IKP11] is almost-exact if one were to use an oracle (should one ever become available) that bins covariates according to importance for estimating causal effects. `DAME`'s predecessor, the FLAME algorithm [WRRV17] is an almost-exact matching method that adapts the distance metric to the data using machine learning. It starts by matching "identical twins," and proceeds by eliminating less important covariates one by one, attempting to match individuals on the largest set of covariates that produce valid matched groups.

FLAME can handle huge datasets, even datasets that are too large to fit in memory, and scales well with the number of covariates, but removing covariates in exactly one order (rather than all possible orders as in `DAME`) means that many high-quality matches will be missed. `DAME` tends to match on more covariates than FLAME; the distances between matched units are smaller in `DAME` than in FLAME, thus its matches are distinctly higher quality. This has implications for missing data, where `DAME` can find matched groups that FLAME cannot.

IVs are widely used to handle uncontrolled confoundedness. Definition and identification of IVs are given in [IR97, AIR96], and generalized in [BP02, CPB16].

Methods for discovery of IVs in observational data are developed in [SS17].The most popular method for IV estimation in the presence of observed confounders is two-stage least squares (2SLS) [Car93]. 2SLS estimators are consistent and efficient under linear single-variable structural equation models with a constant treatment effect [Woo10]. 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 [IT01]. These matching methods can be very powerful nonparametric estimators; full matching [KKA+13] is one such approach, but has a limitation in that its distance metric between covariates is fixed, whereas ours is learned.Other IV methods in the presence of measured covariates include Bayesian methods [IR97], semiparametric methods [Aba03, Tan06, ORR15], nonparametric methods [Frö02] and deep learning methods [HLLBT17], but these methods do not enjoy the benefits of interpretability that matching provides.

Chapter 3

Almost-Matching Exactly (AME) Framework

Consider a dataframe $D = [X, Y, T]$ where $X \in \{0, 1, \ldots, k\}^{n \times p}$, $Y \in \mathbb{R}^n$, $T \in \{0, 1\}^n$ respectively denote the categorical covariates for all units, the outcome vector and the treatment indicator (1 for treated, 0 for control). The j-th covariate X of unit i is denoted $x_{ij} \in \{0, 1, \ldots, k\}$. Notation $\mathbf{x}_i \in \{0, 1, \ldots, k\}^p$ indicates covariates for the ith unit, and $T_i \in \{0, 1\}$ is an indicator for whether or not unit i is treated.

Throughout we make SUTVA and ignorability assumptions [Rub80a]. The goal is to match treatment and control units on as many relevant covariates as possible. Relevance of covariate j is denoted by $w_j \geq 0$ and it is determined using a hold-out training set. w_j's can either be fixed beforehand or adjusted dynamically inside the algorithm, which is called Full-AME.

For now, assuming that we have a fixed weight w_j for each covariate j, we would like to find a match for each treatment unit t that *matches at least one control unit on as many relevant covariates as possible.* Thus we consider the following matching problem:

Almost-Matching Exactly with Fixed Weights (AME): *For each treatment unit* t,

$$\boldsymbol{\theta}^{t\star} \in \text{argmax}_{\boldsymbol{\theta}\in\{0,1\}^p} \boldsymbol{\theta}^T \mathbf{w} \text{ such that}$$

$$\exists\, \ell \text{ with } T_\ell = 0 \text{ and } \mathbf{x}_\ell \circ \boldsymbol{\theta} = \mathbf{x}_t \circ \boldsymbol{\theta},$$

where $\circ$ denotes Hadamard product. The solution to the AME problem is an indicator of the optimal set of covariates for the matched group of treatment unit t. The constraint says that the optimal matched group contains at least one control unit. When the solution of the AME problem is the same for multiple treatment units, they form a single matched group. For treatment unit t, the **main matched group** for t contains all units ℓ so that $\mathbf{x}_t \circ \boldsymbol{\theta}^{t\star} = \mathbf{x}_\ell \circ \boldsymbol{\theta}^{t\star}$. If any unit ℓ within t's main matched group has its own different main matched group, then t's matched group is an **auxiliary matched group** for ℓ.

The formulation of the AME and main matched group formulation is symmetric for control units. There are two straightforward (but inefficient) approaches to solving the AME problem for all units.

AME Solution 1 (quadratic in n, linear in p): Brute force pairwise comparison of treatment points to control points. (Detailed in the appendix.)

AME Solution 2 (order $n \log n$, exponential in p): Brute force iteration over all

2^p subsets of the p covariates. (Detailed in the appendix.)

If n is in the millions, the first solution, or any simple variation of it, is practically infeasible. A straightforward implementation of the second solution is also inefficient. However, a monotonicity property (downward closure) allows us to prune the search space so that the second solution can be modified to be completely practical. The `DAME` algorithm does not enumerate all $\boldsymbol{\theta}$'s, monotonicity reduces the number of $\boldsymbol{\theta}$'s it considers.

Proposition 3.0.1. (Monotonicity of $\boldsymbol{\theta}^*$ in AME solutions) *Fix treatment unit t. Consider feasible $\boldsymbol{\theta}$, meaning $\exists\ \ell$ with $T_\ell = 0$ and $\mathbf{x}_\ell \circ \boldsymbol{\theta} = \mathbf{x}_t \circ \boldsymbol{\theta}$. Then,*

- *Any feasible $\boldsymbol{\theta}'$ such that $\boldsymbol{\theta}' < \boldsymbol{\theta}$ elementwise will have $\boldsymbol{\theta}'^T\mathbf{w} \leq \boldsymbol{\theta}^T\mathbf{w}$.*
- *Consequently, consider feasible vectors $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$. Define $\tilde{\boldsymbol{\theta}}$ as the elementwise $\min(\boldsymbol{\theta}, \boldsymbol{\theta}')$. Then $\tilde{\boldsymbol{\theta}}^T\mathbf{w} < \boldsymbol{\theta}^T\mathbf{w}$, and $\tilde{\boldsymbol{\theta}}^T\mathbf{w} < \boldsymbol{\theta}'^T\mathbf{w}$.*

These follow from the fact that the elements of $\boldsymbol{\theta}$ are binary and the elements of $\mathbf{w}$ are non-negative. The first property means that if we have found a feasible $\boldsymbol{\theta}$, we do not need to consider any $\boldsymbol{\theta}'$ with fewer 1's as a possible solution of the AME for unit t. Thus, the `DAME` algorithm starts from $\boldsymbol{\theta}$ being all 1's (consider all covariates). It systematically drops one element of $\boldsymbol{\theta}$ to zero at a time, then two, then three, ordered according to values of $\boldsymbol{\theta}^T\mathbf{w}$. The second property implies that we must evaluate both

$\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ as possible AME solutions before evaluating $\tilde{\boldsymbol{\theta}}$. Conversely, a new subset of variables defined by $\tilde{\boldsymbol{\theta}}$ cannot be considered unless all of its supersets have been considered. These two properties form the basis of the `DAME` algorithm.

The algorithm must be stopped early in order to avoid creating low quality matches. A useful stopping criterion is if the weighted sum of covariates $\boldsymbol{\theta}\mathbf{w}$.

Note that the matching does not produce estimates, it produces a partition of the covariate space, based on which we can estimate CATEs. Within each matched group, we use the difference between the average outcome of the treated units and the average outcome of the control units as an estimate of the CATE value given the covariate values associated with that group. Smoothing the CATE estimates could be useful after matching.

Chapter 4

The Dynamic AME (DAME) Algorithm

We call a *covariate-set* any set of covariates. We denote by $\mathcal{J}$ the original set of all covariates from the input dataset, where $p = |\mathcal{J}|$. When we *drop* a set of covariates s, it means we will match on $\mathcal{J} \setminus s$. For any covariate-set s, we associate an **indicator-vector** $\boldsymbol{\theta}_s \in \{0,1\}^p$ defined as follows:

$$\boldsymbol{\theta}_{s,i} = \mathbb{1}_{\{i \notin s\}} \qquad \forall\ i \in \{1,..,p\} \tag{4.1}$$

i.e. the value is 1 if the covariate is *not in* s implying that it is being used for matching.

Algorithm 1 gives the pseudocode of the `DAME` algorithm. Instead of looping over all possible 2^p vectors to solve the AME, it considers a covariate-set s for dropping only if satisfies the monotonicity property of Proposition 3.0.1. For example, if $\{1\}$ has been considered for dropping to form matched groups, it would not process $\{1,2,3\}$ next because the monotonicity property requires $\{1,2\}$, $\{1,3\}$ and $\{2,3\}$ to have been considered previously for dropping.

The `DAME` algorithm uses the `GroupedMR` (*Grouped Matching with Replacement*) subroutine given in Algorithm 2 to form all valid main matched groups having at

Algorithm 1: The `DAME` algorithm

Input : Data D, precomputed weight vector w for all covariates (see appendix)

Output : $\{D^m_{(i)}, \mathcal{MG}_{(i)}\}_{i\geq 1}$ all matched units and all the matched groups from all iterations i

Notation: i: iterations, $D_{(i)}$ ($D^m_{(i)}$) = unmatched (matched) units at the end of iteration i, $\mathcal{MG}_{(i)}$ = matched groups at the end of iteration i, $\Lambda_{(i)}$ = set of active covariate-sets at the end of iteration i that are eligible be dropped to form matched groups, $\Delta_{(i)}$ = set of covariate-sets at the end of iteration i that have been processed (considered for dropping and formulation of matched groups).

Initialize: $D_{(0)} = D, D^m_{(0)} = \varnothing, \mathcal{MG}_{(0)} = \varnothing, \Lambda_{(0)} = \{\{1\}, ..., \{p\}\}, \Delta_{(0)} = \varnothing$, $i = 1$

while *there is at least one treatment unit to match in* $D_{(i-1)}$ **do**

- `(find the 'best' covariate-set to drop from the set of active covariate-sets)`
- Let $s^*_{(i)} = \arg\max_{s\in\Lambda_{i-1}} \boldsymbol{\theta}_s^T \mathbf{w}$ ($\boldsymbol{\theta}_s \in \{0,1\}^p$ denotes the indicator-vector of s as in (4.1))
- $(D^m_{(i)}, \mathcal{MG}_{(i)}) = GroupedMR(D, D_{(i-1)}, J \setminus s^*_{(i)})$ `(find matched units and main groups)`
- $Z_{(i)} =$ `GenerateNewActiveSets`$(\Delta_{(i-1)}, s^*_{(i)})$ (`generate new active covariate-sets`)
- $\Lambda_{(i)} = \Lambda_{(i-1)} \setminus \{s^*_{(i)}\}$ (`remove` $s^*_{(i)}$ `from the set of active sets`)
- $\Lambda_{(i)} = \Lambda_{(i)} \cup Z_{(i)}$ (`update the set of active sets`)
- $\Delta_{(i)} = \Delta_{(i-1)} \cup \{s^*_{(i)}\}$ (`update the set of already processed covariate-sets`)
- $D_{(i)} = D_{(i-1)} \setminus D^m_{(i-1)}$ `(remove matches)`
- $i = i + 1$

return $\{D^m_{(i)}, \mathcal{MG}_{(i)}\}_{i\geq 1}$

least one treated and one control unit. `GroupedMR` takes a given subset of covariates and finds all subsets of treatment and control units that have identical values of those covariates. We use an efficient implementation of the `group-by` operation used in the algorithm using *bit-vectors*, which is discussed in the appendix. To keep track of main and auxiliary matched groups, `GroupedMR` takes the entire set of units D as well as the set of unmatched units from the previous iteration $D_{(i-1)}$ as input along

with the covariate-set $J \setminus s^*_{(i)}$ to match on in this iteration. Instead of matching only the unmatched units in $D_{(i-1)}$ using the group-by operator, it matches all units in D to allow for matching with replacement as in the AME objective. It keeps track of the main matched groups for the unmatched units $D_{(i-1)}$ that are matched for the first time, and auxiliary groups (see previous section) for the other matched units.

Algorithm 2: Procedure `GroupedMR`

Input : Data D, unmatched Data $D^{um} \subseteq D = (X, Y, T)$, subset of indexes of covariates $J^s \subseteq \{1, ..., p\}$.

Output : Newly matched units D^m using covariates indexed by J^s where groups have at least one matched and one control unit, and main matched groups for D^m.

M_{raw}=`group-by` (D, J^s) (`form groups on` D `by exact matching on` J^s)

M=`prune(`M_{raw}`)` (`remove groups without at least one treatment and control unit`)

D^m=`Get subset of` D^{um} `where the covariates match with some group in` M (`recover newly matched units (form main matched groups)`)

`For units in` $D \setminus D^m$`, if the covariates match with` M`, record` M `as an auxialiary matched group.`

return $\{D^m, M\}$. (`newly matched units, and main matched groups`)

`DAME` keeps track of two sets of covariate-sets: (1) The set of **processed sets** Δ contains the covariate-sets whose main matched groups (if any exist) have already been formed. That is, Δ contains s if matches have been constructed on $\mathcal{J} \setminus s$ by calling the `GroupedMR` procedure. (2) The set of **active sets** Λ contains the covariate-sets s that are eligible to be dropped according to Proposition 3.0.1. For any iteration i, $\Lambda_{(i)} \cap \Delta_{(i)} = \varnothing$, i.e., the sets are disjoint, where $\Lambda_{(i)}, \Delta_{(i)}$ denote the states of Λ, Δ at the end of iteration i. Due to the monotonicity property Proposition 3.0.1, if $s \in \Lambda_{(i)}$,

then each proper subset $r \subset s$ belonged to $\Lambda_{(j)}$ in an earlier iteration $j < i$. Once an active set $s \in \Lambda_{(i-1)}$ is chosen as the optimal subset to drop $s^*_{(i)}$ in iteration i, s is excluded from $\Lambda_{(i)}$ (it is no longer active) and is included in $\Delta_{(i)}$ as a processed set. In that sense, the active sets are generated and included in $\Lambda_{(i)}$ in a hierarchical manner similar to the apriori algorithm. A set s is included in $\Lambda_{(i)}$ only if all of its proper subsets of one less size $r \subset s$, $|r| = |s| - 1$, have been processed.

The procedure `GenerateNewActiveSets` gives an efficient implementation of generation of new active sets in each iteration of `DAME`, and takes the currently processed sets $\Delta = \Delta_{(i-1)}$ and a newly processed set $s = s^*_{(i)}$ as input. Let $|s| = k$. In this procedure, $\Delta^k \subseteq \Delta \cup \{s\}$ denotes the set of all processed covariate-sets in Δ of size k, and also includes s. Inclusion of s in Δ^k may lead to generation of new active sets of size $k+1$ if all of its subsets of size k (one less) have been already processed. The new active sets triggered by inclusion of s in Δ^k would be supersets r of s of size $k+1$ if all subsets $s' \subset r$ of size $|s'| = k$ belong to Δ^k. To generate such candidate superset r, we can append s with all covariates appearing in some covariate-set in Δ except those in s. However, this naive approach would iterate over many superfluous candidates for active sets. Instead, `GenerateNewActiveSets` safely prunes some such candidates that cannot be valid active sets using **support** of each covariate e in Δ^k, which is the number of sets in Δ^k containing e. Indeed, for any covariate that is not

frequent enough in Δ^k, the monotonicity property ensures that any covariate-set that contains that covariate cannot be active. The following proposition shows that this pruning step does not eliminate any valid active set (proof is in the appendix):

Proposition 4.0.1. *If for a superset r of a newly processed set s where $|s| = k$ and $|r| = k+1$, all subsets s' of r of size k have been processed (i.e. r is eligible to be active after s is processed), then r is included in the set Z returned by* `GenerateNewActiveSets`.

The explicit verification step of whether all possible subsets of r of one less size belongs to Δ^k is necessary, i.e., the above optimization only prunes some candidate sets that are guaranteed not to be active. For instance, consider $s = \{2, 3\}$, $k = 2$, and $\Delta^2 = \{\{1, 2\}, \{1, 3\}, \{3, 5\}, \{5, 6\}\} \cup \{\{2, 3\}\}$. For the superset $r = \{2, 3, 5\}$ of s, all of $2, 3, 5$ have support of ≥ 2 in Δ^2, but this r cannot become active yet, since the subset $\{2, 5\}$ of r does not belong to Δ^2.

Finally, the following theorem states the correctness of the `DAME` algorithm (proof is in the appendix).

Theorem 4.0.2. ***(Correctness)*** *The algorithm* `DAME` *solves the AME problem.*

Algorithm 3: Procedure `GenerateNewActiveSets`

1. **Input :** s a newly dropped set of size k,
   2. Δ the set of previously processed sets
3. **Initialize:** `new active sets` $Z = \varnothing$

– compute all subsets of Δ of size k and include s

4. $\Delta^k = \{\delta \in \Delta \mid size(\delta) = k\} \cup \{s\}$

Notation: $\mathcal{S}_e$ = support of covariate e in Δ^k

– get all the covariates contained in sets in Δ^k

5. $\Gamma = \{\gamma \mid \gamma \in \delta \text{ and } \delta \in \Delta^k\}$

– get the covariates that have enough support minus s

6. $\Omega = \{\alpha \mid \alpha \in \Gamma \text{ and } \mathcal{S}_e \geq k\} \setminus s$

– if all covariates in s have enough support in Δ^k

7. **if** $\{\forall e \in s : \mathcal{S}_e \geq k\}$ **then**
   – generate new active set
   8. **for** *all* $\alpha \in \Omega$ **do**
      9. $r = s \cup \{\alpha\}$
      10. **if** *all subsets* $s' \subset r$, $|s'| = k$, *belong to* Δ^k **then**
          – add newly active set r to Z
          11. add r to Z

return Z

Example (follow line number correspondence)

1. $s = \{2, 3\}$, $k = 2$,
2. $\Delta = \{\{1\}, \{2\}, \{3\}, \{5\}, \{1, 2\}, \{1, 3\}, \{1, 5\}\}$
3. $Z = \varnothing$
4. $\Delta^2 = \{\{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 5\}\}$
5. $\mathcal{S}_1 = 3, \mathcal{S}_2 = 2, \mathcal{S}_3 = 2, \mathcal{S}_5 = 1$
6. $\Gamma = \{1, 2, 3, 5\}$
7. $\Omega = \{1, 2, 3\} \setminus \{2, 3\} = \{1\}$
8. $True$:

$\alpha = 1$

9. $r = \{2, 3\} \cup \{1\} = \{1, 2, 3\}$
10. $True$ (subsets of r of size 2 are $\{1, 2\}, \{1, 3\}, \{2, 3\}$)
11. $Z = \{\{1, 2, 3\}\}$

return $Z = \{\{1, 2, 3\}\}$

Chapter 5

Almost-Matching Exactly with Instrumental Variables (AME-IV)

In this section, we adapt the AME framework to the case that the randomness assumption is broken. We introduce an adaptive framework with instrumental variables (AME-IV) and solve it by using an approximate algorithm (FLAME-IV).

5.1 Basic Instrumental Variables Assumptions

We consider the problem of instrumental variable estimation for a set of n units indexed by $i = 1, \ldots, n$. 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 [AIR96]. 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., [IA94, AIR96, IR15]):

(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 affected 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.

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 [IA94]. We note that these are not the only criteria for the use of instrumental variables,

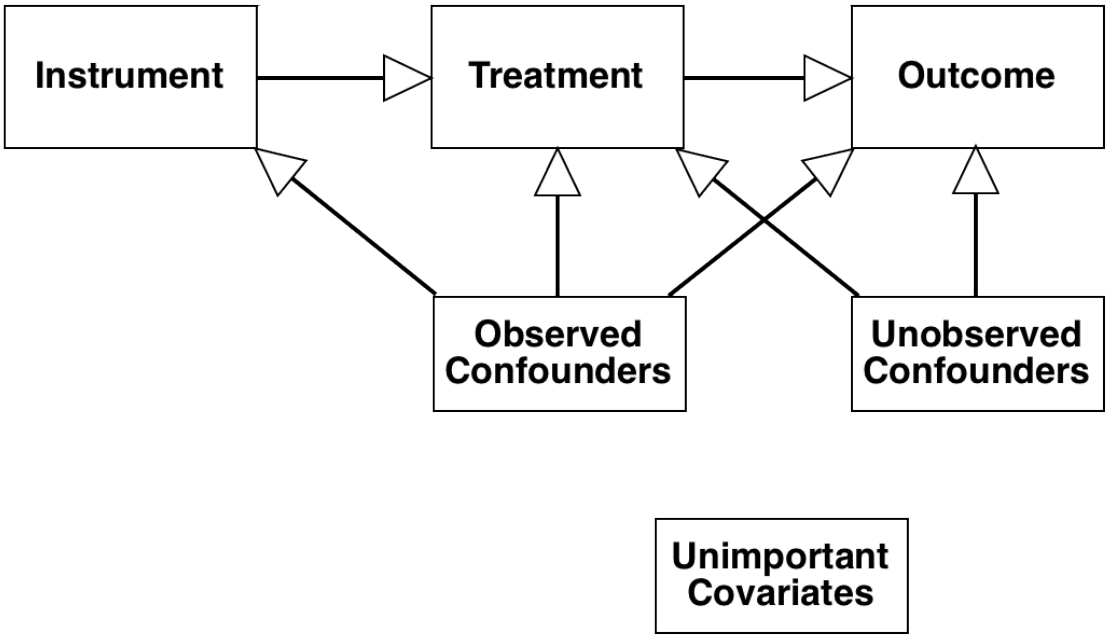

Figure 5.1: Causal DAG for instrumental variables.
Notes: Arrows represent causal relationships between variables. The lack of a direct arrow from instrument to outcome represents Assumption A2 and the lack of a direct arrow from unobserved confounders to the instrument represents A3.

for example [BP02] introduces a graphical criterion for identification with instrumental variables. Figure 5.1 gives a graphical summary of the identification assumptions introduced before. These are units that would have undertaken the treatment only after administration of the instrument and never without [AIR96]. Note that we cannot identify these units in our sample, given what we observe, but we can estimate the treatment effect on them [IR15]. This treatment effect is known as Local Average

Treatment Effect (LATE) and takes the following form [IA94, AIR96]:

$$\begin{aligned}\lambda &= \frac{1}{n_c} \sum_{i:t_i(1)>t_i(0)} y_i(1) - y_i(0) \\ &= \frac{\sum_{\mathbf{x}\in\mathcal{X}} \omega_{\mathbf{x}} ITT_{y,\mathbf{x}}}{\sum_{\mathbf{x}\in\mathcal{X}} \omega_{\mathbf{x}} ITT_{t,\mathbf{x}}},\end{aligned} \tag{5.1}$$

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:

$$ITT_{y,\mathbf{x}} = \frac{1}{n_{\mathbf{x}}} \sum_{i:\mathbf{x}_i=\mathbf{x}} y_i(t_i(1)) - y_i(t_i(0))$$

$$ITT_{t,\mathbf{x}} = \frac{1}{n_{\mathbf{x}}} \sum_{i:\mathbf{x}_i=\mathbf{x}} t_i(1) - t_i(0).$$

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.

5.2 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 , which is to maximize the weighted sum of covariates on which we can create a valid matched

group for :

$$\boldsymbol{\theta}^* \in \underset{\boldsymbol{\theta}\in\{0,1\}^p}{\arg\max}\, \boldsymbol{\theta}^T \mathbf{w}, \;\; such\ that$$

$$\exists\, k \;\; with\ z_k = 0 \;\; and\ \mathbf{x}_k \circ \boldsymbol{\theta} = \mathbf{x} \circ \boldsymbol{\theta},$$

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 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 $\boldsymbol{\theta}^{i*}$ happen to be one, then the units in unit i's main matched group will be exact matches for i.

We define 's **main matched group** in terms of $\boldsymbol{\theta}^*$ as:

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

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 $\mathbb{1}_{[\mathbf{x}_i \neq \mathbf{x}_k]}$ to denote a vector of length p where the j^{th} entry is one if $x_{ij} = x_{kj}$ and zero otherwise.

Lemma 5.2.1. *For any unit where $z = 1$, with $\boldsymbol{\theta}^*$ as defined in the AME-IV problem, then for any unit k with $z_k \neq z$, if $\mathbf{x}_k \circ \boldsymbol{\theta}^* = \mathbf{x} \circ \boldsymbol{\theta}^*$, i.e., $k \in MG(\boldsymbol{\theta}^{i*}, \mathbf{x}_i)$, we have:*

$$k \in \underset{\substack{l=1,\ldots,n \\ z_l \neq z}}{\arg\min}\, \boldsymbol{w}^T \mathbb{1}_{[\mathbf{x}_i \neq \mathbf{x}_l]}. \tag{5.2}$$

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

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 and all other units with a different assignment of the instrument that belong to '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: $\mathbf{w}^T \mathbb{1}_{[\mathbf{x}_i \neq \mathbf{x}_k]} \leq \delta \implies |p(Z_i = 1|\mathbf{x}_i) - p(Z_k = 1|\mathbf{x}_k)| \leq \epsilon(\delta)$, 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 [Ros10]. The following proposition follows immediately from Lemma 5.2.1 applied to A5.

Proposition 5.2.2. *If* $k \in MG(\boldsymbol{\theta}^{i*}, \mathbf{x}_i)$ *with* $z_i \neq z_k$*, and A5 holds, then*

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

In particular, if $\boldsymbol{\theta}^{i*}$ *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.

5.3 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 *machine learning*

on a training set. Specifically, the units $1, \ldots, n$ 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 $\boldsymbol{\theta}$ as:

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

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 $\boldsymbol{\theta}$.)

We ensure that we always match using sets of variables $\boldsymbol{\theta}$ that together have a low error $\widehat{PE}_{\mathcal{F}}$. 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 ,

$$\boldsymbol{\theta}^* \in \arg\min_{\boldsymbol{\theta} \in \{0,1\}^p} \widehat{PE}_{\mathcal{F}}(\boldsymbol{\theta}), \text{ such that:}$$

$$\exists\, k \;\; \textit{with } z_k = 0 \;\; \textit{and } \mathbf{x}_k \circ \boldsymbol{\theta}^* = \mathbf{x} \circ \boldsymbol{\theta}^*,$$

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 [WRVR17].

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.

5.4 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 [WRVR17] 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 5.3) 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. 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.

5.4.1 Estimation

Assuming that (A1) through (A5) and SUTVA hold, the LATE, λ, can be estimated in a consistent way [IA94, AIR96]; in this section we adapt common estimators for λ to our matching framwork. Consider a collection of m matched groups, $\mathcal{MG}_1, \ldots, \mathcal{MG}_m$, each associated with a different value of $(\boldsymbol{\theta}, \mathbf{x})$. We estimate the average causal effect of the instrument on the treatment, $ITT_{t,\ell}$ and on the outcome, $ITT_{y,\ell}$, 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 $ITT_{y,\ell}$:

$$\widehat{ITT}_{y,\ell} = \frac{\sum_{i \in \mathcal{MG}_\ell} y_i z_i}{\sum_{i \in \mathcal{MG}_\ell} z_i} - \frac{\sum_{i \in \mathcal{MG}_\ell} y_i (1 - z_i)}{\sum_{i \in \mathcal{MG}_\ell} (1 - z_i)}. \tag{5.3}$$

Similarly, the estimator for the causal effect of the instrument on the treatment, $ITT_{t,j}$, can be written as:

$$\widehat{ITT}_{t,\ell} = \frac{\sum_{i \in \mathcal{MG}_\ell} t_i z_i}{\sum_{i \in \mathcal{MG}_\ell} z_i} - \frac{\sum_{i \in \mathcal{MG}_\ell} t_i (1 - z_i)}{\sum_{i \in \mathcal{MG}_\ell} (1 - z_i)}. \tag{5.4}$$

From the form of λ in Equation (5.1) it is easy to see that, if the estimators in (5.3) and (5.4) are unbiased for $ITT_{y,\ell}$ and $ITT_{t,\ell}$ 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:

$$\hat{\lambda} = \frac{\sum_{\ell=1}^{m} n_\ell \widehat{ITT}_{y,\ell}}{\sum_{\ell=1}^{m} n_\ell \widehat{ITT}_{t,\ell}}, \tag{5.5}$$

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:

$$\lambda_\ell = \frac{1}{n_\ell} \sum_{\substack{i \in \mathcal{MG}_\ell: \\ t_{i\ell}(1) > t_{i\ell}(0)}} y_i(1) - y_i(0). \tag{5.6}$$

This can be accomplished with the following estimator:

$$\hat{\lambda}_\ell = \frac{\widehat{ITT}_{y,\ell}}{\widehat{ITT}_{t,\ell}}. \tag{5.7}$$

We quantify uncertainty around our estimates with asymptotic Confidence Intervals (CIs). To compute CIs for these estimators we adapt the approach laid out in [IR15]. 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.

Chapter 6

Simulations

In this section, we show the performances of `DAME` and FLAME-IV on simulated datasets.

6.1 Simulations with DAME Algorithm

We firstly present results of `DAME` under several data generating processes. We show that `DAME` produces higher quality matches than popular matching methods such as 1-PSNNM (propensity score nearest neighbor matching) and Mahalanobis distance nearest neighbor matching, and better treatment effect estimates than black box machine learning methods such as Causal Forest (which is not a matching method, and is not interpretable). The 'MatchIt' R-package [HIKS11a] was used to perform 1-PSNNM and Mahalanobis distance nearest neighbor matching ('Mahalanobis'). For Causal Forest, we used the 'grf' R-package [ATW19]. `DAME` also improves over FLAME [WRRV17] with regards to the quality of matches. Other matching methods (optmatch, cardinality match) do not scale to large problems and thus needed to be omitted.

Throughout this section, the outcome is generated with $y = \sum_i \alpha_i x_i + T \sum_{i=1} \beta_i x_i + T \cdot U \sum_{i,\gamma,\gamma>i} x_i x_\gamma$ where $T \in \{0, 1\}$ is the binary treatment indicator. This generation process includes a baseline linear effect, linear treatment effect, and quadratic (non-linear) treatment effect. We vary the distribution of covariates, coefficients (α's, β's, U), and the fraction of treated units.

6.1.1 Presence of Irrelevant Covariates

A basic sanity check for matching algorithms is how sensitive they are to irrelevant covariates. To that end, we run experiments with a majority of the covariates being irrelevant to the outcome. For important covariates $1 \le i \le 5$ let $\alpha_i \sim N(10s, 1)$ with $s \sim \text{Uniform}\{-1, 1\}$, $\beta_i \sim N(1.5, 0.15)$, $x_i \sim \text{Bernoulli}(0.5)$. For unimportant covariates $5 < i \le 15$, $x_i \sim \text{Bernoulli}(0.1)$ in the control group and $x_i \sim \text{Bernoulli}(0.9)$ in the treatment group so there is little overlap between treatment and control distributions. This simulation generates 15000 control units, 15000 treatment units, 5 important covariates and 10 irrelevant covariates. **Results:** As Figure 6.1 shows, `DAME` and FLAME achieve the optimal result before dropping any important covariates. When imposing that FLAME and `DAME` find all possible matches even if important covariates are dropped (not recommended), poor matches can sometimes be introduced. In the simulation, `DAME` never ends up dropping important covariates, but FLAME does.

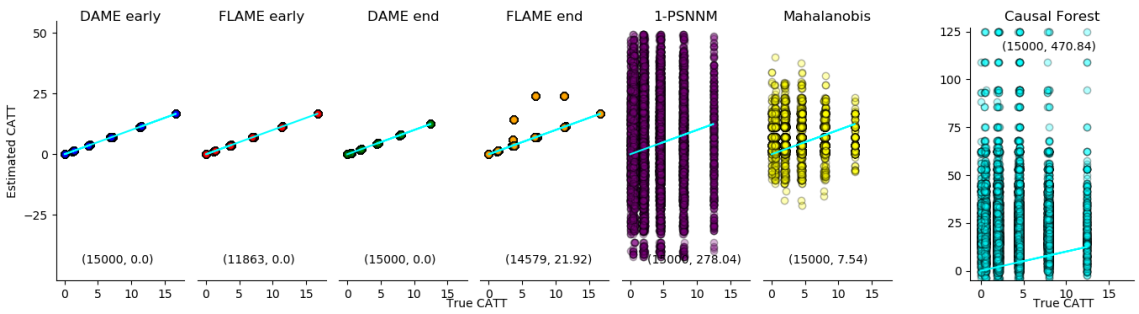

Figure 6.1: Estimated CATT vs. True CATT. `DAME` and FLAME perfectly estimate the CATTs before dropping important covariates.
Notes: `DAME` matches all units without dropping important covariates, but FLAME needs to stop early in order to avoid poor matches. All other methods are sensitive to irrelevant covariates and give poor estimates. The two numbers on each plot are the number of matched units and MSE.

However, even FLAME's worst case scenario is better than the comparative methods, all of which perform poorly in the presence of irrelevant covariates. Causal Forest is especially ill suited for this case.

6.1.2 Exponentially Decaying Covariates

A notable advantage of `DAME` over FLAME is that it produces more high quality matches before resorting to lower quality matches. To test this, we consider covariates of decaying importance, letting the α parameters decrease exponentially as $\alpha_i = 64 \times \left(\frac{1}{2}\right)^i$. We evaluated performance when $\approx 30\%$ and 50% of the units are matched.

Results: As Figure 6.2 shows, `DAME` matches on more covariates, yielding better estimates than FLAME.

Table 6.1: MSE for different imbalance ratios

	Mean Squared Error (MSE)		
	Ratio 1	Ratio 2	Ratio 3
D-AEM	**0.47**	**0.83**	**1.39**
FLAME	0.52	0.88	1.55
Mahalanobis	26.04	48.65	64.80
1-PSNNM	246.08	304.06	278.87

6.1.3 Imbalanced Data

Imbalance is common in observational studies: there are often substantially more control than treatment units. The data for this experiment has covariates with decreasing importance. A fixed batch of 2000 treatment and 40000 control units were generated. We sampled from the controls to construct different imbalance ratios: 40000 in the most imbalanced case (Ratio 1), then 20000 (Ratio 2), and 10000 (Ratio 3).

6.1.4 Run Time Evaluation

We compare the running time of `DAME` with a brute force solution (AEM Solution 1 described in Section 3 and in the appendix). All experiments were run on an Ubuntu 16.04.01 system with Intel Core i7 Processor (Cores: 8, Speed: 3.6 GHz), 8 GB RAM. **Results:** As shown in Figure 6.3, FLAME provides the best run-time performance because it incrementally reduces the number of covariates, rather than

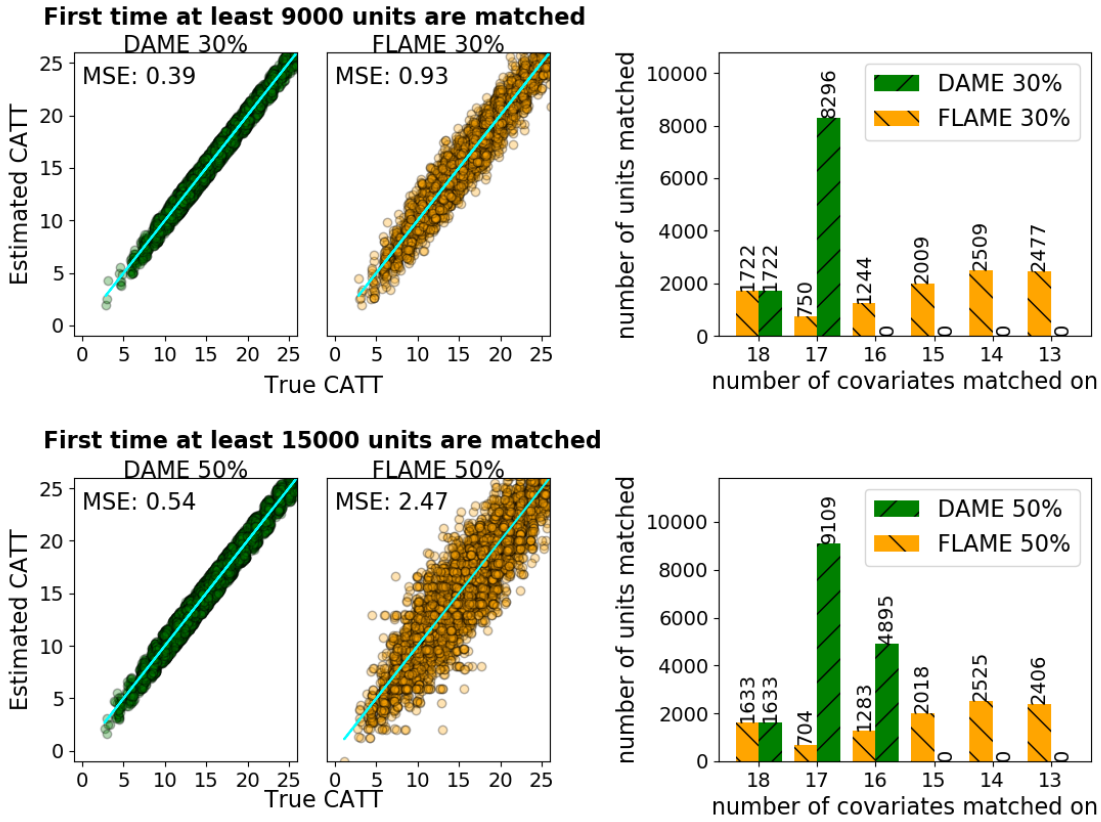

Figure 6.2: `DAME` makes higher quality matches early on. `DAME` matches on more covariates than FLAME, yielding lower MSE from matched groups.

Notes: Rows correspond to stopping thresholds (TH1=30%, TH2=50%).

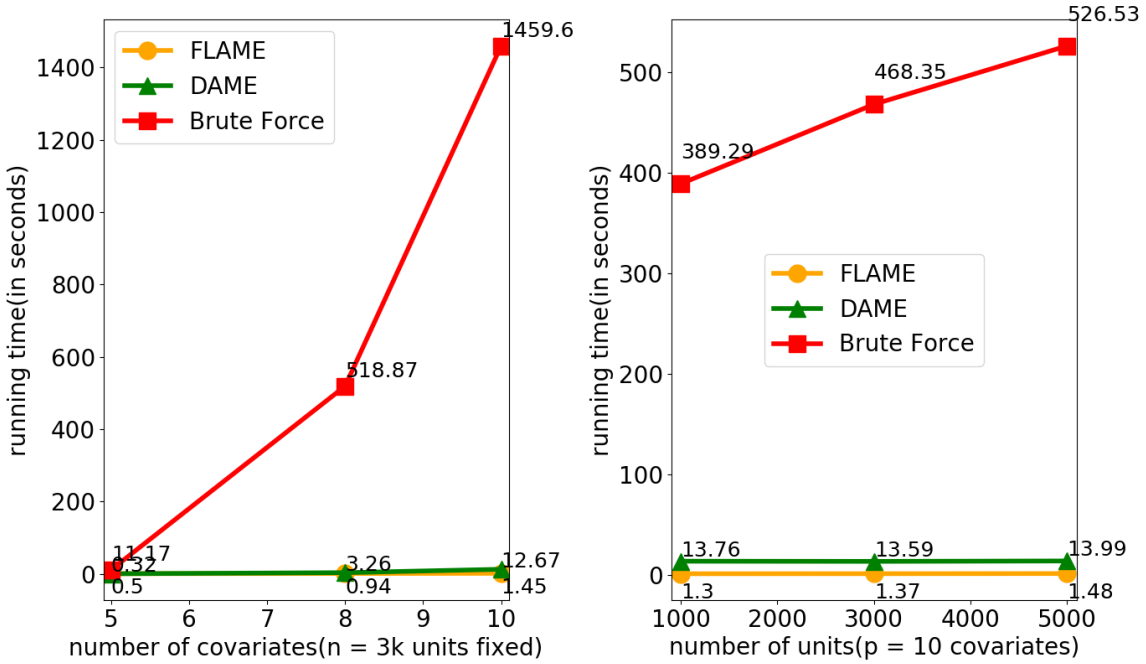

Figure 6.3: Run-time comparison between `DAME` FLAME, and brute force. *Notes:* *Left:* varying number of units. *Right:* varying number of covariates.

solving Full-AEM. On the other hand, as shown in the previous simulations, `DAME` produces high quality matches that the other methods do not. It solves the AEM much faster than brute force. The running time for `DAME` could be further optimized through simple parallelization of the checking of active sets.

6.2 Simulations with FLAME-IV Algorithm

In this section, We evaluate the performance of FLAME-IV using simulated data. We compare our approach to several other methods including **two-stage least squares** [AK91, Car93, Woo10], and two other state-of-the-art nonparametric meth-

ods for instrumental variables, **full matching** [KKM+16] and **nearfar matching** [BSLR10]. 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.

6.2.1 Implementation of FLAME-IV

We implement FLAME-IV using both *bit-vector calculations* and *database queries* (`GROUP-BY` operators). More details about the implementation are in the supplement.

In the first set of experiments, we compare the performance of the different methods on the estimation of local average treatment effects. In Experiment 6.2.3 we demonstrate the power of FLAME-IV for estimating individualized local average treatment effects. Experiment 6.2.4 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 [Woo10]:

$$T_i^\star = k + \pi Z_i + \rho^T X_i + \xi_i \tag{6.1}$$

where $Z_i \sim \text{Bernoulli}(0.5)$, and $\xi_i \sim N(0, 0.8)$. For important covariates, $X_{ij} \sim$

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 $T_i^\star$ by defining:

$$T_i = \mathbb{1}_{[0.3<T_i^\star \leq 0.6]} + 2 \times \mathbb{1}_{[0.6<T_i^\star \leq 1.0]} + 3 \times \mathbb{1}_{[T_i^\star > 1.0]}.$$

Experiments were run on an Ubuntu 16.04.01 system with eight-core CPU (Intel Core i7-4790 @ 3.6 GHz) and 8 GB RAM.

6.2.2 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:

$$Y_i = \sum_{j=1}^{10} \alpha_j X_{ij} + 10 T_i \tag{6.2}$$

$$Y_i = \sum_{j=1}^{10} \alpha_j X_{ij} + 10 T_i + \sum_{1 \leq j < \gamma \leq 5} X_{ij} X_{i\gamma}. \tag{6.3}$$

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 $\alpha_j = 0.5^j$. (Variables with earlier indices are more important.)

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. (6.1). Usually a concentration parameter below 10 suggests that instruments are weak stock2002survey.

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 6.4 and 6.5 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 6.4 shows results for the linear generation model, and Figure 6.5 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

Table 6.2: Point Estimates for Linear and Nonlinear Models.

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

Notes: 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.

on a parametric outcome model and uses a discrete learned distance metric. The only exceptions are the left-upper plot on Figure 6.4 and Figure 6.5, 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.

Next, we compare 95% confidence intervals for each algorithm. The results are reported in Table 6.2. 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 computation of confidence intervals are available in the supplement.

6.2.3 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 5.4.1.

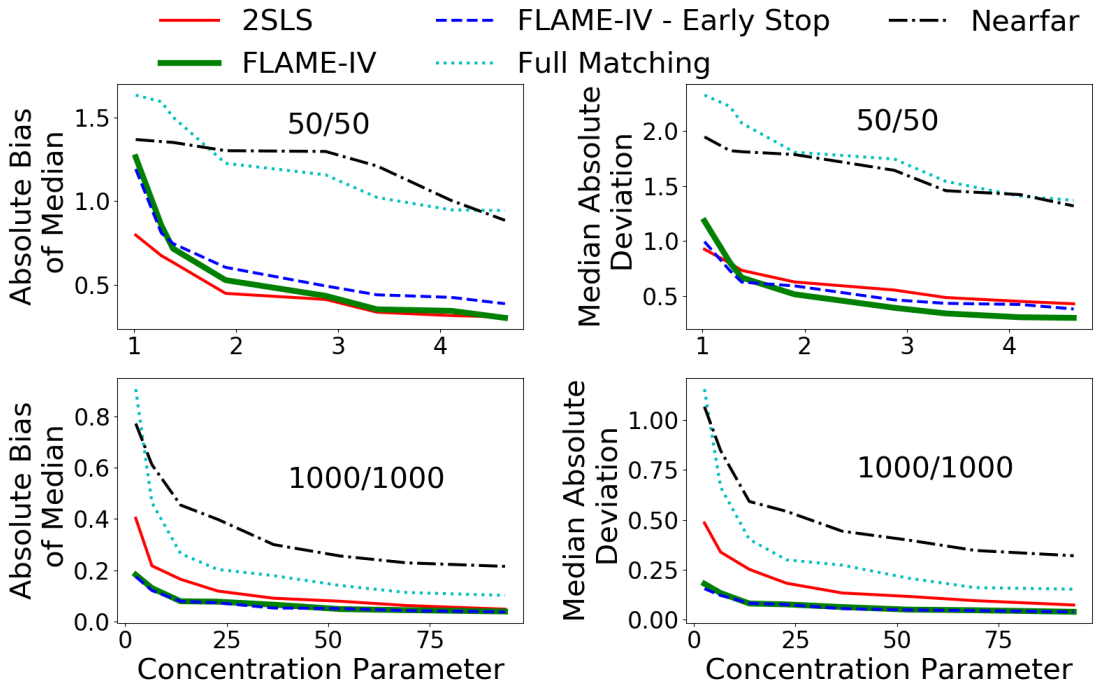

Figure 6.4: Performance for linear generation model with various sample sizes. Here, 2SLS has an advantage because the data are generated according to a 2SLS model. *Notes:* FLAME-IV (either early-stopping or run-until-no-more-matches) outperforms other methods on the large dataset, with smaller absolute 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.

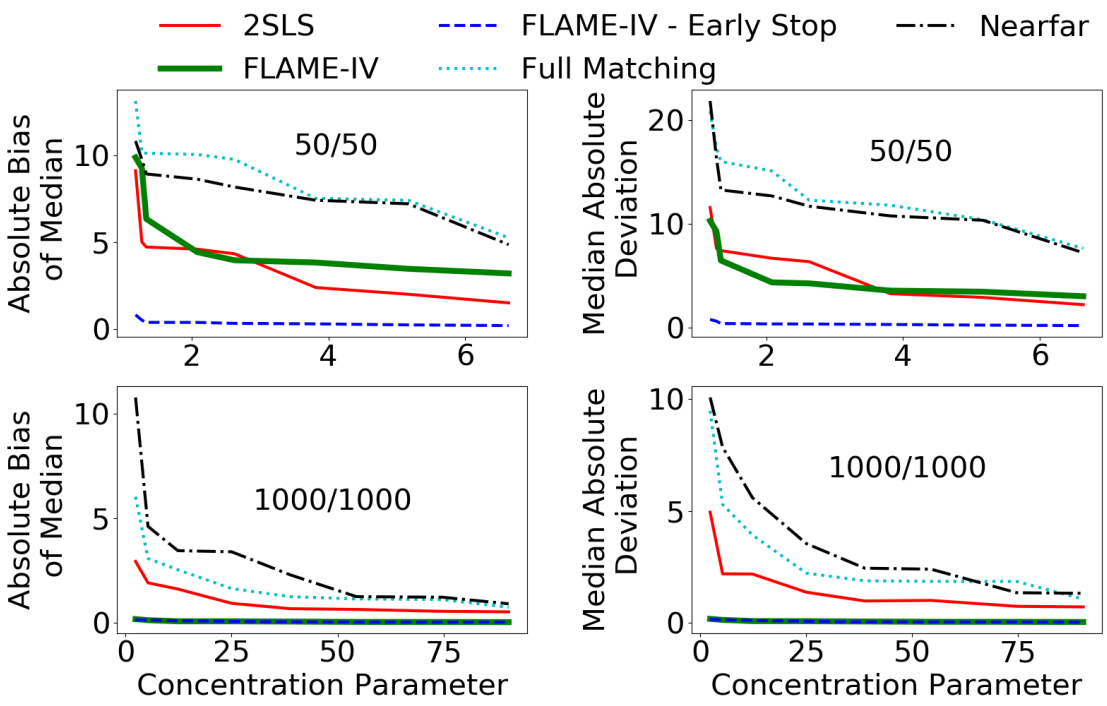

Figure 6.5: Performance for nonlinear generation model with different sample sizes. Here, the 2SLS model is misspecified.

Notes: 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.

To study how well FLAME-IV estimates individual causal effects, we generate data with heterogeneous treatment effects. The new generation models, (6.4) and (6.5) below, are unlike the generation models in (6.2) and (6.3), in that different individuals have different treatment effects. The two heterogeneous treatment effect

data generation models are:

$$Y_i = \sum_{j=1}^{T} \alpha_j X_{ij} + T_i \sum_{j=1}^{10} \beta_j X_{ij} \tag{6.4}$$

$$Y_i = \sum_{j=1}^{T} \alpha_j X_{ij} + T_i \sum_{j=1}^{10} \beta_j X_{ij} + \sum_{\substack{j=1\ldots5 \\ \gamma=1\ldots5 \\ \gamma>j}} X_{ij} X_{i\gamma}. \quad . \tag{6.5}$$

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. (6.1) 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 groups were not computed and are not reported in Figure 6.6. 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 6.6. 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.

6.2.4 Running Time Evaluation

For the synthetic data generated by Section 5.2, Figure 6.7 compares the runtime of our algorithm against full matching. We computed the runtime by varying number of units (Figure 6.7, left panel) and by varying number of covariates (Figure 6.7, 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.

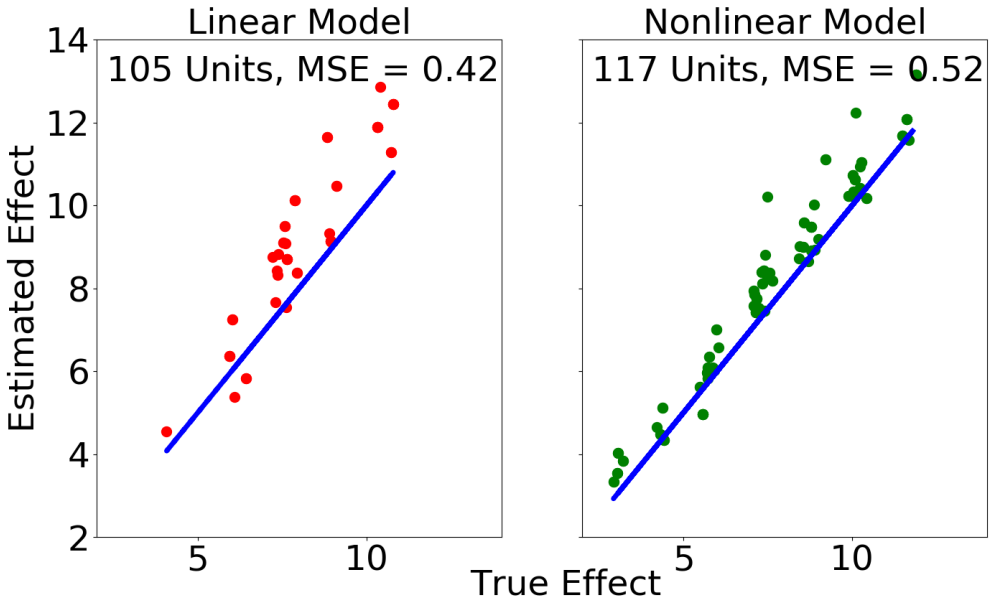

Figure 6.6: True Individual Causal Effect vs. Estimated Individual Causal Effect. *Notes:* The numbers on each plot represent the total number of instrumented units for calculating unit-level LATE, and MSE of our predictions. The concentration parameter is the same for the whole dataset, set to 288.84 for the linear outcome model, and 272.92 for the nonlinear outcome model.

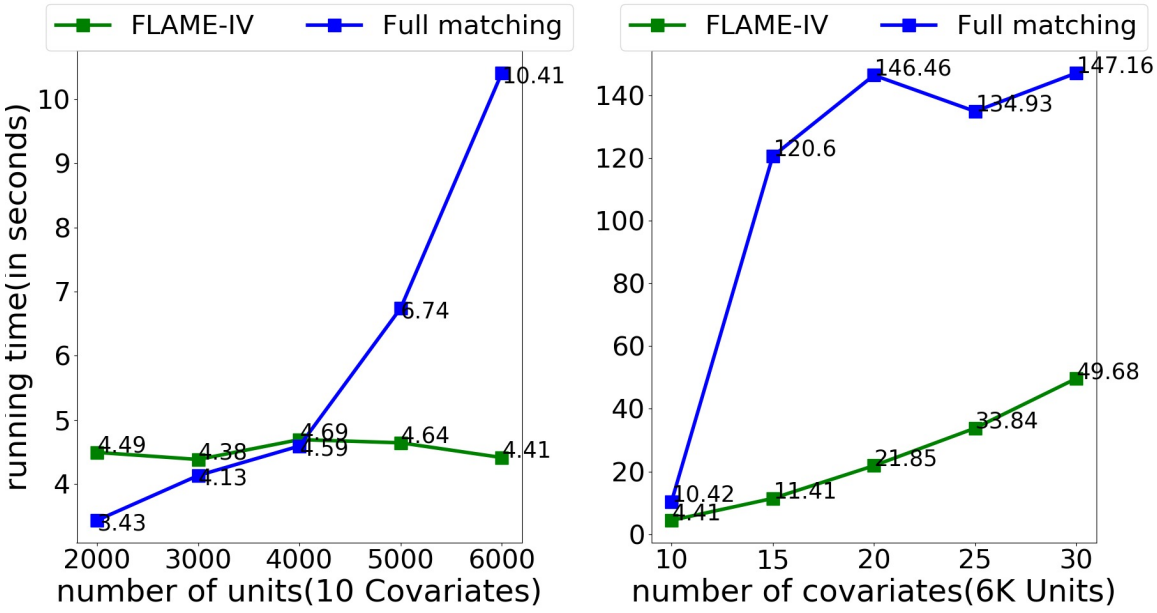

Figure 6.7: Running Time for FLAME-IV and Full Matching.
Notes: Left panel presents run time by varying number of units, whereas the right panel presents run time by varying number of covariates.

Chapter 7

Real Data Experiment

In this section, we apply the DAME and FLAME-IV in two interesting real-world applications.

7.1 Breaking the Cycle of Drugs and Crime in the United States

Breaking The Cycle (BTC) [HMM06] is a social program conducted in several U.S. states designed to reduce criminal involvement and substance abuse among current offenders. This dataset includes basic personal information for 380 individuals who were chosen to participate(treatment group) and not participate(control group) in this program. More details about how survey was conducted and which covariates we are choosing are in Appendix F. In this experiment, we study the effects of participating in the program on reducing non-drug offense rates.

The details of the data and our results are in Appendix. We compared CATE predictions of DAME and FLAME to double check the performance of a black box support vector machine (SVM) approach that predict positive, neutral, or negative

treatment effect for each individual. The result is that `DAME` and the SVM approach agreed on most of the matched units. All of the units for which `DAME` predicted approximately zero treatment effect all have a "neutral" treatment effect predicted label from the SVM approach. Most other predictions were similar between the two methods. There were only few disagreements between the methods. Upon further investigation, we found that the differences are due to the fact that `DAME` is a matching method and not a modeling method; the estimates could be smoothed afterwards if desired to create a model. The smoothed model is likely to agree with the predictions from the SVM.

7.2 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 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

Table 7.1: Effect of Door-to-Door Canvassing on Electoral Outcomes.

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

Notes: The table reports causal effect of door-to-door canvassing on electoral outcomes for the two rounds of 2012 French general election. Two electoral outcomes are of interest, (1) vote share for Parti Socialiste (PS), and (2) voter turnout. 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.

election [Pon18]. 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 7.1. 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 [Pon18], 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. 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 7.1. 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 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 employ. 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.

Table 8.2 in the supplement shows 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.

Chapter 8

Appendix

8.1 Naïve AME solutions

AME Solution 1 (quadratic in n, linear in p): For all treatment units t, we (i) iterate over all control units c, (ii) find the vector $\boldsymbol{\theta}_{tc} \in \{0,1\}^p$ with value 1 if there is a match on the values of the corresponding covariates, and 0 otherwise, (iii) find the control unit(s) with the highest value of $\boldsymbol{\theta}_{tc}^T \mathbf{w}$, and (iv) return them as the main matched group for the treatment unit t (and compute the auxiliary group). Whenever a previously matched unit α is matched to a previously unmatched unit η, record the η's main matched group as an auxiliary group for the previously matched unit α. When all units are 'done' (all units are either matched already or cannot be matched) then stop, and compute the CATE for each treatment and control unit using its main matched group. If a unit belongs to auxiliary matched groups then its outcome is used for computing both its own CATE (in its own main matched group) and the CATEs of units for whom it is in an auxiliary group (e.g., α will be used to compute η's estimated CATE). This algorithm is polynomial in both n and p, however, the quadratic time complexity in n also makes this approach impractical for large datasets

(for instance, when we have more than a million units with half being treatment units).

AME Solution 2 (order $n \log n$, exponential in p:) This approach solves the AMER problem simultaneously for all treatment and control units for a fixed weight vector $\mathbf{w}$. First, (i) enumerate every $\boldsymbol{\theta} \in \{0,1\}^p$ (which serves as an indicator for a subset of covariates), (ii) order the $\boldsymbol{\theta}$'s according to $\boldsymbol{\theta}^T \mathbf{w}$, (iii) call `GroupedMR` for every $\boldsymbol{\theta}$ in the predetermined order, (iv) the first time each unit is matched during a `GroupedMR` procedure, mark that unit with a 'done' flag, and record its corresponding main matched group and, to facilitate matching with replacement, (v) whenever a previously matched unit is matched to a previously unmatched unit, record this main matched group as an auxiliary group. When all units are 'done' (all units are either matched already or cannot be matched) then stop, and compute the CATE for each treatment and control unit using its main matched group. Each unit's outcome will be used to estimate CATEs for every auxiliary group that it is a member of, as before. Although this approach exploits the efficient 'group by' function (e.g., provided in database (SQL) queries), which can be implemented in $O(n \log n)$ time by sorting the units, iterating over all possible vectors $\boldsymbol{\theta} \in \{0,1\}^p$ makes this approach unsuitable for practical purposes (exponential in p).

8.2 Proof of Proposition 4.0.1

Proposition 4.0.1 *If for a superset r of a newly processed set s where $|s| = k$ and $|r| = k + 1$, all subsets s' of r of size k have been processed (i.e. r is eligible to be active after s is processed), then r is included in the set Z returned by* `GenerateNewActiveSets`.

Proof. Suppose all subsets of r of size k are already processed and belong to Δ^k. Let f be the covariate in $r \setminus s$. Clearly, f would appear in Δ^k, since at least one subset $s' \neq s$ of r of size k would contain f, and $s' \in \Delta^k$. Further all covariates in r, including f and those in s will have support at least k in Δ^k. To see this, note that there are $k + 1$ subsets of r of size k, and each covariate in r appears in exactly k of them. Hence $f \in \Omega$, which the set of high support covariates. Further, the 'if' condition to check minimum support for all covariates in s is also satisfied. In addition, the final 'if' condition to eliminate false positives is satisfied too by assumption (that all subsets of r are already processed). Therefore r will be included in Z returned by the procedure. □

8.3 Proof of Theorem 4.0.2

Theorem 4.0.2 ***(Correctness)*** *The* `DAME` *algorithm solves the AME problem.*

Proof. Consider any treatment unit t. Let s be the set of covariates in its main matched group returned in `DAME` (the while loop in `DAME` runs as long as there is a treated unit and the stopping criteria have not been met, and the `GroupedMR` returns the main matched group for every unit when it is matched for the first time). Let $\boldsymbol{\theta}_s$ be the indicator vector of s (see Eq. 4.1). Since the `GroupedMR` procedure returns a main matched group only if it is a *valid* matched group containing at least one treated and one control unit (see Algorithm 2), and since all units in the matched group on s have the same value of covariates in $\mathcal{J} \setminus s$, there exists a unit ℓ with $T_\ell = 0$ and $\mathbf{x}_\ell \circ \boldsymbol{\theta}_s = \mathbf{x}_t \circ \boldsymbol{\theta}_s$.

Hence it remains to show that the covariate set s in the main matched group for t corresponds to the maximum weight $\boldsymbol{\theta}^T\mathbf{w}$. Assume that there exists another covariate-set r such that $\boldsymbol{\theta}_r^T\mathbf{w} > \boldsymbol{\theta}_s^T\mathbf{w}$, there exists a unit ℓ' with $T_{\ell'} = 0$ and $\mathbf{x}_{\ell'} \circ \boldsymbol{\theta}_r = \mathbf{x}_t \circ \boldsymbol{\theta}_r$, and gives the maximum weight $\boldsymbol{\theta}_r^T\mathbf{w}$ over all such r.

(i) r cannot be a (strict) subset of s, since `DAME` ensures that all subsets are processed before a superset is processed to satisfy the downward closure property in Proposition 3.0.1.

(ii) r cannot be a (strict) superset of s, since it would violate the assumption that $\boldsymbol{\theta}_r^T\mathbf{w} > \boldsymbol{\theta}_s^T\mathbf{w}$ for non-negative weights.

(iii) Assume that r and s are incomparable (there exist covariates in both $r \setminus s$ and

$s \setminus r$). Suppose the active set s was chosen in iteration h. If r was processed in an earlier iteration $h' < h$, since r forms a valid matched group for t, it would give the main matched group for t violating the assumption.

Given (i)–(iii) we argue that r must be active at the start of iteration h, and will be chosen as the best covariate set in iteration h, leading to a contradiction.

Note that we start with all singleton sets as active sets in $\Delta_{(0)} = \{\{1\}, \cdots, \{p\}\}$ in the `DAME` algorithm. Consider any singleton subset $r_0 \subseteq r$ (comprising a single covariate in r). Due to the downward closure property in Proposition 3.0.1, $\boldsymbol{\theta}_{r_0}^T \mathbf{w} \geq \boldsymbol{\theta}_r^T \mathbf{w} > \boldsymbol{\theta}_s^T \mathbf{w}$. Hence all of the singleton subsets of r will be processed in earlier iterations $h' < h$, and will belong to the set of processed covariate sets $\Lambda_{(h-1)}$.

Repeating the above argument, consider any subset $r' \subseteq r$. It holds that $\boldsymbol{\theta}_{r'}^T \mathbf{w} \geq \boldsymbol{\theta}_r^T \mathbf{w} > \boldsymbol{\theta}_s^T \mathbf{w}$. All subsets r' of r will be processed in earlier iterations $h' < h$ starting with the singleton subsets of r. In particular, all subsets of size $|r| - 1$ will belong to $\Lambda_{(h-1)}$. As soon as the last of those subsets is processed, the procedure `GenerateNewActiveSets` will include r in the set of active sets in a previous iteration $h' < h$. Hence if r is not processed in an earlier iteration, it must be active at the start of iteration h, leading to a contradiction.

Hence for all treatment units t, the covariate-set r giving the maximum value of $\boldsymbol{\theta}_r^T \mathbf{w}$ will be used to form the main matched group of t, showing the correctness of

the DAME algorithm. □

8.4 Proof of Lemma 5.2.1

Since the result is exactly symmetric when non-instrumented units are matched we prove it only for the case when instrumented units are matched. Assume $\mathbf{w} \in \mathbb{R}^p$. For a given unit i with $z_i = 1$, suppose we could find a $\boldsymbol{\theta}^*$ as defined in the AME-IV problem. Let us define another unit k with $z_k = 0$, and $\mathbf{x}_k \circ \boldsymbol{\theta}^* = \mathbf{x} \circ \boldsymbol{\theta}^*$, by definition of $\mathcal{MG}(\boldsymbol{\theta}^*, \mathbf{x}_i)$ it must be that $\mathbf{x}_k \in \mathcal{MG}(\boldsymbol{\theta}^*, \mathbf{x}_i)$. So $\mathbb{1}_{[\mathbf{x}_i \neq \mathbf{x}_k]} = J - \boldsymbol{\theta}^*$, where J is a vector of length p that has all entries equals to 1.

Assume there is another unit j with $z_j = 0$, and $j \neq k$.

If $j \in \mathcal{MG}(\boldsymbol{\theta}^*, \mathbf{x}_i)$, then $\mathbb{1}_{[\mathbf{x}_i \neq \mathbf{x}_j]} = J - \boldsymbol{\theta}^*$. So

$$\mathbf{w}^T \mathbb{1}_{[\mathbf{x}_i \neq \mathbf{x}_k]} = \mathbf{w}^T (J - \boldsymbol{\theta}^*) = \mathbf{w}^T \mathbb{1}_{[\mathbf{x}_i \neq \mathbf{x}_j]}$$

If $j \notin \mathcal{MG}(\boldsymbol{\theta}^*, \mathbf{x}_i)$, let us define $\boldsymbol{\theta}^j = J - \mathbb{1}_{[\mathbf{x}_i \neq \mathbf{x}_j]}$, obviously $\boldsymbol{\theta}^j \neq \boldsymbol{\theta}^*$. Since

$\boldsymbol{\theta}^* \in \arg\max_{\boldsymbol{\theta}\in\{0,1\}^p} \boldsymbol{\theta}^T \mathbf{w}$, we have:

$$\begin{aligned}
\mathbf{w}^T \mathbb{1}_{[\mathbf{x}_i \neq \mathbf{x}_k]} &= \mathbf{w}^T (J - \boldsymbol{\theta}^*) \\
&= \mathbf{w}^T - \mathbf{w}^T \boldsymbol{\theta}^* \\
&< \mathbf{w}^T - \mathbf{w}^T \boldsymbol{\theta}^j \\
&= \mathbf{w}^T (J - \boldsymbol{\theta}^j) \\
&= \mathbf{w}^T \mathbb{1}_{[\mathbf{x}_i \neq \mathbf{x}_j]}.
\end{aligned}$$

Therefore,

$$k \in \arg\min_{\substack{j=1,\ldots,n \\ Z_j=0}} \mathbf{w}^T \mathbb{1}_{[\mathbf{x}_i \neq \mathbf{x}_j]}.$$

This concludes the proof.

8.5 Asymptotic Variance and Confidence Intervals for LATE Estimates

To construct estimators for the variance of $\hat{\lambda}$ we use an asymptotic approximation, that is, we will try to estimate the asymptotic variance of $\hat{\lambda}$, rather than its small sample variance. The strategy we use to do this is the same as [IR15], with the difference that our data is grouped: we adapt their estimators to grouped data using

canonical methods for stratified sampling. In order to define asymptotic quantities for our estimators, we must marginally expand the definitions of potential outcomes introduced in our paper. In practice, while our framework has been presented under the assumption that the potential outcomes and treatments are fixed, we now relax that assumption and instead treat $y_i(1), y_i(0), t_i(1), t_i(0)$ as realizations of random variables $Y_i(1), Y_i(0), T_i(1), T_i(0)$, which are drawn from some unknown distribution $f(Y_i(1), Y_i(0), T_i(1), T_i(0))$. In this case the SUTVA assumption requires that each set of potential outcomes and treatments is independently drawn from the same distribution for all units. As usual, lowercase versions of the symbols above denote observed realizations of the respective random variables.

Recall as well that in this scenario we have a set of m matched groups $\mathcal{MG}_1, \ldots \mathcal{MG}_m$ indexed by ℓ, such that each unit is only in one matched group. We denote the number of units in matched group ℓ that have $z_i = 1$ with n_ℓ^1 and the number of units in matched group ℓ with $z_i = 0$ with n_ℓ^0. Finally the total number of units in matched group ℓ is $n_\ell = n_\ell^0 + n_\ell^1$.

We make all the assumptions listed in Section 5.1 but we must require a variant of (A3), to be used instead of it. This assumption is:

(A3') $\Pr(Z_i = 1 | i \in \mathcal{MG}_\ell) = \Pr(Z_k = 1 | k \in \mathcal{MG}_\ell) = \frac{n_\ell^1}{n_\ell}, \forall i, k.$

That is, if two units are in the same matched group, then they have the same

probability of receiving the instrument. This probability will be equal to the ratio of instrument 1 units to all units in the matched group because we hold these quantities fixed. Note that this more stringent assumption holds when matches are made exactly, and is common in variance computation for matching estimators (see, for example, [KKM+16]).

We keep our exposition concise and we do not give explicit definitions for our variance estimands. These are all standard and can be found in [IR15].

We have to start from estimating variances of observed potential outcomes and treatments within each matched group. We do so with the canonical approach:

$$\begin{aligned}
\hat{s}^2_{\ell 0} &= \frac{1}{n^0_\ell - 1} \sum_{i \in \mathcal{MG}_\ell} \left(y_i(1 - z_i) - \frac{1}{n^0_\ell} \sum_{i \in \mathcal{MG}_\ell} y_i(1 - z_i) \right)^2 \\
\hat{s}^2_{\ell 1} &= \frac{1}{n^1_\ell - 1} \sum_{i \in \mathcal{MG}_\ell} \left(y_i z_i - \frac{1}{n^1_\ell} \sum_{i \in \mathcal{MG}_\ell} y_i z_i \right)^2 \\
\hat{r}^2_{\ell 0} &= \frac{1}{n^0_\ell - 1} \sum_{i \in \mathcal{MG}_\ell} \left(t_i(1 - z_i) - \frac{1}{n^0_\ell} \sum_{i \in \mathcal{MG}_\ell} t_i(1 - z_i) \right)^2 \\
&= 0 \\
\hat{r}^2_{\ell 1} &= \frac{1}{n^1_\ell - 1} \sum_{i \in \mathcal{MG}_\ell} \left(t_i z_i - \frac{1}{n^1_\ell} \sum_{i \in \mathcal{MG}_\ell} t_i z_i \right)^2,
\end{aligned}$$

where: $\hat{s}^2_{\ell 0}$ is an estimator for the variance of potential responses for the units with instrument value 0 in matched group ℓ, $\hat{s}^2_{\ell 1}$ for the variance of potential responses for the units with instrument value 1 in matched group ℓ, $\hat{r}^2_{\ell 0}$ for the variance of

potential treatments the units with instrument value 0 in matched group ℓ, and $\hat{r}^2_{\ell 1}$ is an estimator for the variance of potential treatments for the units with instrument value 1 in matched group ℓ. The fact that $\hat{r}^2_{\ell 0} = 0$ follows from Assumption A4.

We now move to variance estimation for the two ITTs. Conservatively biased estimators for these quantities are given in [IR15]. These estimators are commonly used in practice and simple to compute, hence why they are often preferred to unbiased but more complex alternative. We repeat them below:

$$\widehat{Var}(\widehat{ITT}_y) = \sum_{\ell=1}^{m} \left(\frac{n_\ell}{n}\right)^2 \left(\frac{\hat{s}^2_{\ell 1}}{n^1_\ell} + \frac{\hat{s}^2_{\ell 0}}{n^0_\ell}\right)$$

$$\widehat{Var}(\widehat{ITT}_t) = \sum_{\ell=1}^{m} \left(\frac{n_\ell}{n}\right)^2 \frac{\hat{r}^2_{\ell 1}}{n^1_\ell}.$$

To estimate the asymptotic variance of $\hat{\lambda}$ we also need estimators for the covariance of the two ITTs both within each matched group, and in the whole sample. Starting with the former, we can use the standard sample covariance estimator for $Cov(\widehat{ITT}_{y\ell}, \widehat{ITT}_{t\ell})$:

$$\begin{aligned}\widehat{Cov}(\widehat{ITT}_{y\ell}, \widehat{ITT}_{t\ell}) = &\frac{1}{n^1_\ell(n^1_\ell - 1)} \\ &\times \sum_{i \in \mathcal{MG}_\ell} \left(y_i z_i - \frac{1}{n^1_\ell} \sum_{i \in \mathcal{MG}_\ell} y_i z_i\right) \\ &\times \left(t_i z_i - \frac{1}{n^1_\ell} \sum_{i \in \mathcal{MG}_\ell} t_i z_i\right).\end{aligned}$$

The reasoning behind why we use only units with instrument value 1 to estimate this covariance is given in [IR15], and follows from A4. We can use standard techniques for covariance estimation in grouped data to combine the estimators above into an overall estimator for $Cov(\widehat{ITT}_y, \widehat{ITT}_t)$ as follows:

$$\widehat{Cov}(\widehat{ITT}_y, \widehat{ITT}_t) = \sum_{\ell=1}^{m} \left(\frac{n_\ell}{n}\right)^2 \widehat{Cov}(\widehat{ITT}_{y\ell}, \widehat{ITT}_{t\ell}).$$

Once all these estimators are defined, we can use them to get an estimate of the asymptotic variance of $\hat{\lambda}$. This quantity is obtained in [IR15] with an application of the delta method to convergence of the two ITTs. The final estimator for the asymptotic variance of $\hat{\lambda}$, which we denote by σ^2, is given by:

$$\begin{aligned}\hat{\sigma}^2 = &\frac{1}{\widehat{ITT}_t^2}\widehat{Var}(\widehat{ITT}_y) + \frac{\widehat{ITT}_y^2}{\widehat{ITT}_t^4}\widehat{Var}(\widehat{ITT}_t) \\ &- 2\frac{\widehat{ITT}_y}{\widehat{ITT}_t^3}\widehat{Cov}(\widehat{ITT}_y, \widehat{ITT}_t).\end{aligned}$$

Using this variance, $1-\alpha\%$ asymptotic confidence intervals can be computed in the standard way.

Algorithm 4: FLAME-IV Algorithm

Input : (i) Input data $D = (X, Y, T, Z)$. (ii) holdout training set $D^H = (X^H, Y^H, T^H, Z^H)$.

Output : A set of matched groups $\{\mathcal{MG}_h\}_{h\geq 1}$ and ordering of covariates $j_1^\star, j_2^\star, ..,$ eliminated.

Initialize $D_0 = D = (X, Y, T, Z)$, $J_0 = \{1, ..., p\}$, $h = 1$, $run = True$, $\mathcal{MG} = \varnothing$. `(h is the index for iterations, j is the index for covariates)`

$(D_0^m, D_0 \setminus D_0^m, \mathcal{MG}_1) = GroupedMR(D_0, J_0)$.

while $run = True$ and $D_{h-1} \setminus D_{h-1}^m \neq \varnothing$ `(we still have data to match)` **do**

  $D_h = D_{h-1} \setminus D_{h-1}^m$ `(remove matches)`

  for $j \in J_{h-1}$ (temporarily remove one feature at a time and compute match quality) **do**

    $(D_h^{mj}, D_h \setminus D_h^{mj}, \mathcal{MG}_{temp}^j) = GroupedMR(D_h, J_{h-1} \setminus j)$.

    $D^{Hj} = [X^H(:, J_{h-1} \setminus j), Y^H, T^H, Z^H]$

    $q_{hj} = MQ(D_h^{mj}, D^{Hj})$

  if other stopping conditions are met, **then**

    $run = False$ `(break from the` **while** `loop)`

  $j_h^\star \in \arg\min_{j \in J_{h-1}} q_{hj}$: `(choose feature to remove)`

  $J_h = J_{h-1} \setminus j_h^\star$ (`remove feature` $j_h^\star$)

  $D_h^m = D_h^{mj^\star}$ and $\mathcal{MG}_h = \mathcal{MG}_{temp}^{j_h^\star}$ `(newly matched data and groups)`

  $h = h + 1$

return $\{\mathcal{MG}_h, D_h^m, J_h\}_{h\geq 1}$ *`(return all the matched groups and covariates used)`*

8.6 The FLAME-IV Algorithm

Basic Matching Requirement (R1): There should be at least one instrumented and one noninstrumented unit in each matched group.

Algorithm 4 presents the matching algorithm for FLAME-IV. Initially, the input with n units is given as $D = (X, Y, T, Z)$, where X (and $n \times p$ matrix) denotes the covariates, Y (an $n \times 1$ vector) is the outcome, T (an $n \times 1$ vector) is the treatment, and Z is the instrument. The covariates are indexed with $J = 1, \cdots, p$.

Algorithm 5: `GroupedMR` procedure

Input : Unmatched Data $D^{um} = (X, Y, T, Z)$, subset of indexes of covariates $J^s \subseteq \{1, ..., p\}$.

Output : Newly matched units D^m using covariates indexed by J^s where groups obey (R1),the remaining data as $D^{um} \setminus D^m$ and matched groups for D^m.

M_{raw}=`group-by` (D^{um}, J^s) (`form groups by exact matching on` J^s)

M=`prune(`M_{raw}`) (remove groups not satisfying (R1))`

D^m=`Get subset of` D^{um} `where the covariates match with` M `(recover newly matched units)`

return $\{D^m, D^{um} \setminus D^m, M\}$.

Let $\mathcal{MG}_l$ represent a set of all matched groups at iteration h of the FLAME-IValgorithm. At iteration h of the algorithm, FLAME-IVcomputes a subset of the matched groups $\mathcal{MG}_h$ such that, for each matched group $mg \in \mathcal{MG}_l$, there is at least one treated and one control unit. Note that it is possible for $\mathcal{MG}_h = \varnothing$, in which case no matched groups are returned in that iteration. M_u denotes the iteration when a unit u is matched. Overloading notation, let M_{mg} denote the iteration when a matched group mg is formed. Hence if a unit u belongs to a matched group mg, $M_u = M_{mg}$ (although not every u with $M_u = M_{mg}$ is in mg).

We use $D_h \subseteq D$ to denote the unmatched units and $J_h \subseteq J$ to denote the remaining variables when iteration $h + 1$ of the while loop starts (*i.e.*, after iteration h ends). Initially $J_0 = J$. While the algorithm proceeds, the algorithm drops one covariate $\pi(h)$ in each iteration (whether or not there are any valid non-empty matched groups), and therefore, $J_h = J \setminus \{\pi(j)_{j=1}^h\}$, $|J_h| = p - h$. All matched groups $mg \in \mathcal{MG}_h$ in iteration

h use J_{h-1} as the subset of covariates on which to match.

The first call to `GroupedMR`: First we initialize the variables D_0, J_0, h, and run. The variable run is true as long as the algorithm is running, while $h \geq 1$ denotes an iteration. After the initialization step, the subroutine `GroupedMR` (see Algorithm 5) finds all of the exact matches in the data $D = D_0$ using *all* features $J = J_0$, such that each of the matched groups $mg \in \mathcal{MG}_1$ contains at least one instrumented and one uninstrumented observation (*i.e.*, satisfies constraint (R1)). The rest of the iterations in the algorithm aim to find the best possible matches for the rest of the data by selectively dropping covariates as discussed in the previous section.

The while loop and subsequent calls to `GroupedMR`: At each iteration of the **while** loop, each feature is temporarily removed (in the **for** loop over j) and evaluated to determine if it is the best one to remove by running `GroupedMR` and computing the matched quality MQ. Since `GroupedMR` does not consider feature j (one less feature from the immediately previous iteration), there are fewer constraints on the matches, and it is likely that there will be new matches returned from this subroutine.

We then need to determine whether a model that excludes feature j provides sufficiently high quality matches and predictions. We would not want to remove j if doing so would lead to poor predictions or if it led to few new matches. Thus, `MQ`

is evaluated by temporarily removing each j, and the j^* that is chosen for removal creates the most new matches and also does not significantly reduce the prediction quality. The algorithm always chooses the feature with largest `MQ` to remove, and remove it. After the algorithm chooses the feature to remove, the new matches and matched groups are stored. The remaining unmatched data are used for the next iteration $h+1$.

Stopping Conditions: If we run out of unmatched data, the algorithm stops. There are also a set of **early-stop conditions** we use to stop algorithm in advance.

Early-Stop Conditions:

(1) There are no more covariates to drop.

(2) Unmatched units are either all instrumented or uninstrumented.

(3) The matching quality drops by 5% or more than the matching quality of exact matching.

Finally, the matched groups are returned along with the units and the features used for each set of matched groups formed in different iterations.

The key component in the Basic FLAME-IV algorithm (Algorithm 4) is the `GroupedMR` procedure (Algorithm 5). The steps of `GroupedMR` can be easily implemented in Java, Python, or R. In the next two subsections we give two efficient implementations of `GroupedMR`, using database queries and bit vector techniques.

8.7 Implementation of `GroupedMR` using Database (SQL) Queries

In this implementation, we keep track of matched units globally by keeping an extra column in the input database D called `is_matched`. For every unit, the value of `is_matched` = ℓ if the unit is matched in a valid group with at least one instrumented and one uninstrumented unit in iteration ℓ of Algorithm 4, and `is_matched = 0` if the unit is still unmatched. Therefore instead of querying the set of unmatched data D^{um} at each iteration (as in the input of Algorithm 5), at each iteration we query the full database D, and consider only the unmatched units for matching by checking the predicate `is_matched = 0` in the query. Let $A_1, \cdots, A_p$ be the covariates in J_s. The SQL query is described below:

```
WITH tempgroups AS
  (SELECT A_1, A_2, ..., A_p
   /*matched groups identified by covariate values*/
   FROM D
   WHERE is_matched = 0
    /*use data that are not yet matched*/
   GROUP BY A_1, A_2, ..., A_p
```

```
    /*create matched groups with identical covariates*/
   HAVING SUM(Z) >= 1 AND
       SUM(Z) <= COUNT(*)-1
   /*groups have >=1 instrumented, but not all instrumented*/
   ),
UPDATE D
SET is_matched = ℓ
WHERE  EXISTS
   (SELECT D.A_1, D.A_2, ···, D.A_p
    FROM tempgroups S
    /*set of covariate values for valid groups*/
    WHERE  S.A_1 = D.A_1
    AND S.A_2 = D.A_2
    AND ··· AND S.A_p = D.A_p)
   AND is_matched = 0
```

The *WITH clause* computes a temporary relation *tempgroups* that computes the combination of values of the covariates forming 'valid groups' (*i.e.*, groups with at least one instrumented and one noninstrumented unit) on unmatched units. The *HAVING clause* of the SQL query discards groups that are invalid – since instruments

Z takes binary values $0, 1$, for any valid group the sum of Z values will be strictly > 0 and $<$ total number of units in the group. Then we update the population table D, where the values of the covariates of the existing units match with those of a valid group in *tempgroups*. Several optimizations of this basic query are possible and are used in our implementation. Setting the `is_matched` value to level ℓ (instead of a constant value like 1) helps us compute the conditional LATE for each matched group efficiently.

8.8 Implementation of `GroupedMR` using Bit Vectors

In this section we discuss an bit-vector implementation to the `GroupedMR` procedure discussed above. We will assign unit u's covariates to a single integer b_u. Unit u's covariates, appended with the instrumental variable indicator, will be assigned an integer b_u^+. Let us discuss how to compute b_u and b_u^+. Suppose $|J_s| = q$, and the covariates in J_s are indexed (by renumbering from J) as 0 to $q-1$. If the j-th covariate is $k_{(j)}$-ary ($k_{(j)} \geq 2$), we first rearrange the q covariates such that $k_{(j)} \geq k_{(j+1)}$ for all $0 \leq j \leq q-2$. Thus the (reordered) covariate values of unit u, $(a_{q-1}, a_{q-2}, \ldots, a_0)$, is represented by the number $b_u = \sum_{j=0}^{q-1} a_j k_{(j)}^j$. Together with the instrument indicator value $Z = z$, the set $(a_{q-1}, a_{q-2}, \ldots, a_0, z)$ for unit u is represented by the number $b_u^+ = z + \sum_{j=0}^{p-1} a_j k_{(j)}^{j+1}$. Since the covariates are rearranged so that $k_{(j)} \leq k_{(j+1)}$ for all

Table 8.1: Example population table illustrating the *bit-vector* implementation.

1st variable	2nd variable	Z	b_u	b_u^+	c_u	c_u^+	matched?
0	2	0	6	18	1	1	No
1	1	0	4	11	2	1	Yes
1	0	1	1	3	1	1	No
1	1	1	4	12	2	1	Yes

Notes: Here the second unit and the fourth unit are matched to each other while the first and third units are left unmatched.

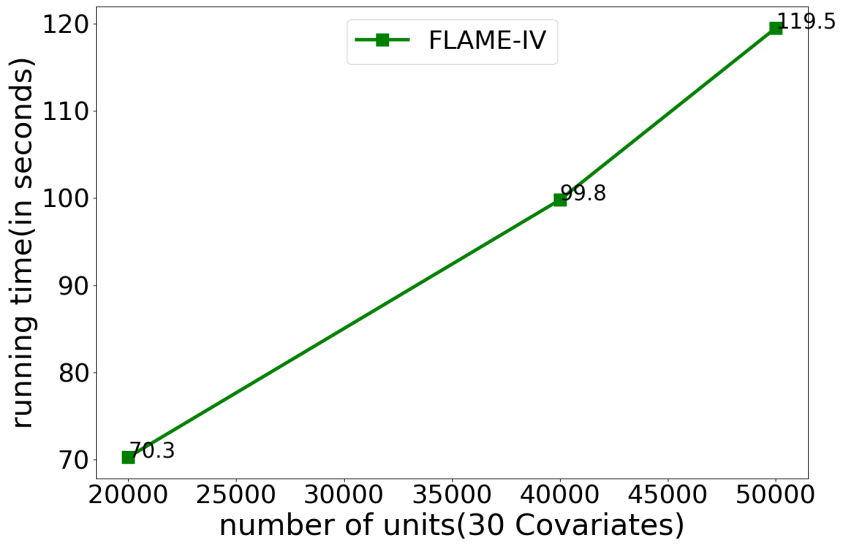

Figure 8.1: Running Time for FLAME-IV on large dataset.

$0 \le j \le q-2$, two units u and u' have the same covariate values if and only if $b_u = b_{u'}$. For each unit u, we count how many times b_u and b_u^+ appear, and denote them as c_u and c_u^+ respectively. (The counting is done by NumPy's `unique()` function.) To perform matching, we compute the b_u, b_u^+, c_u, c_u^+ values for all units and mark a unit as matched if its c_u value and c_u^+ value differ.

Proposition 8.8.1 guarantees the correctness of the bit-vector implementation.

Proposition 8.8.1. *A unit u is matched if and only if $c_u \neq c_u^+$, since the two counts*

b_u and b_u^+ differ iff the same combination of covariate values appear both as an instrumented unit and an uninstrumented unit.

An example of this procedure is illustrated in Table 8.1. We assume in this population the 0-th variable is binary and the next variable is ternary. In this example, the number b_1 for the first unit is $0 \times 2^0 + 2 \times 3^1 = 6$; the number b_1^+ including its treatment indicator is $0 + 0 \times 2^1 + 2 \times 3^2 = 18$. Similarly we can compute all the numbers b_u, b_u^+, c_u, c_u^+, and the matching results are listed in the last column in Table 8.1. subsectionMore Running Time Results on Large Dataset Figure 8.1 shows the results of running time for FLAME-IV on a larger dataset. The running time is still very short(< 2 min) on the large dataset for FLAME-IV. Full matching can not handle a dataset of this size.

8.9 Sample Matched Groups

Sample matched groups are given in Table 8.2. These groups were produced by FLAME-IVon the data from [Pon18], introduced in Section 8.9. The algorithm was ran on all of the covariates collected in the original study except for territory. Here we report some selected covariates for the groups. The first group is comprised of electoral districts in which previous turnout was relatively good but PS vote share was low. This suggest that existing partisan splits are being taken into account

Table 8.2: Two sample matched groups generated by FLAME on the application data described in Section 8.9.

Territory	Last Election PS Vote Share	Last Election Turnout	Population (in thousands)	Share Male	Share Unemployed	Treated	Instrumented
Matched Group 1							
Plouguenast et environs	(0.01, 0.05]	(0.77, 0.88]	(0, 450]	(0.47, 0.57]	(0, 0.1]	0	1
Lorrez-le-Bocage-Préaux et environs	(0.01, 0.05]	(0.77, 0.88]	(0, 450]	(0.47, 0.57]	(0, 0.1]	0	1
La Ferté-Macé et environs	(0.01, 0.05]	(0.77, 0.88]	(0, 450]	(0.47, 0.57]	(0, 0.1]	0	1
Mundolsheim et environs	(0.01, 0.05]	(0.77, 0.88]	(0, 450]	(0.47, 0.57]	(0, 0.1]	1	1
Paris, 7e arrondissement	(0.01, 0.05]	(0.77, 0.88]	(1,800, 2,250]	(0.47, 0.57]	(0.1, 0.2]	0	1
Sainte-Geneviève et environs	(0.01, 0.05]	(0.77, 0.88]	(0, 450]	(0.47, 0.57]	(0, 0.1]	0	0
Cranves-Sales et environs	(0.01, 0.05]	(0.77, 0.88]	(0, 450]	(0.47, 0.57]	(0, 0.1]	0	0
Hem et environs	(0.01, 0.05]	(0.77, 0.88]	(0, 450]	(0.47, 0.57]	(0, 0.1]	0	1
Legé et environs	(0.01, 0.05]	(0.77, 0.88]	(0, 450]	(0.47, 0.57]	(0, 0.1]	0	1
Moûtiers et environs	(0.01, 0.05]	(0.77, 0.88]	(0, 450]	(0.47, 0.57]	(0, 0.1]	0	0
Paris, 7e arrondissement	(0.01, 0.05]	(0.77, 0.88]	(1,800, 2,250]	(0.47, 0.57]	(0.1, 0.2]	0	1
Craponne-sur-Arzon et environs	(0.01, 0.05]	(0.77, 0.88]	(0, 450]	(0.47, 0.57]	(0, 0.1]	0	0
Matched Group 2							
Nantes	(0.19, 0.22]	(0.66, 0.77]	(0, 450]	(0.47, 0.57]	(0.1, 0.2]	1	1
Alès	(0.19, 0.22]	(0.66, 0.77]	(0, 450]	(0.37, 0.47]	(0.2, 0.3]	1	1
Sin-le-Noble	(0.19, 0.22]	(0.66, 0.77]	(0, 450]	(0.47, 0.57]	(0.2, 0.3]	1	1
Grand-Couronne et environs	(0.19, 0.22]	(0.66, 0.77]	(0, 450]	(0.47, 0.57]	(0.1, 0.2]	1	1
Dreux	(0.19, 0.22]	(0.66, 0.77]	(0, 450]	(0.47, 0.57]	(0.2, 0.3]	1	1
Vosges	(0.19, 0.22]	(0.77, 0.88]	(0, 450]	(0.47, 0.57]	(0.1, 0.2]	0	0
Arras et environs	(0.19, 0.22]	(0.66, 0.77]	(0, 450]	(0.37, 0.47]	(0.1, 0.2]	1	1
Montargis et environs	(0.19, 0.22]	(0.66, 0.77]	(0, 450]	(0.37, 0.47]	(0.2, 0.3]	1	1
Marseille, 3e arrondissement	(0.19, 0.22]	(0.66, 0.77]	(450, 900]	(0.47, 0.57]	(0.1, 0.2]	1	1
Nantes	(0.19, 0.22]	(0.66, 0.77]	(0, 450]	(0.47, 0.57]	(0.1, 0.2]	1	1
Mâcon et environs	(0.19, 0.22]	(0.66, 0.77]	(0, 450]	(0.37, 0.47]	(0.1, 0.2]	1	1

Notes: The columns are a subset of the covariates used for matching.
Territory was not used for matching. Original covariates are continuous and were coarsened into 5 bins. Last election PS vote share was coarsened into 10 bins. Labels in the cells represent lower and upper bounds of the covariate bin each unit belongs to.
The two groups have relatively good match quality overall.

by FLAME-IVfor matching. Municipalities in the second group have slightly lower turnout at the previous election but a much larger vote share for PS. Note also that treatment adoption is very high in the second group, while low in the first: this suggest that the instrument is weak in Group 1 and strong in Group 2.

Bibliography

[Aba03] Alberto Abadie. Semiparametric instrumental variable estimation of treatment response models. *Journal of econometrics*, 113(2):231–263, 2003.

[ADHI04] Alberto Abadie, David Drukker, Jane Leber Herr, and Guido W Imbens. Implementing matching estimators for average treatment effects in stata. *The Stata Journal*, 4(3):290–311, 2004.

[AEI15] Susan Athey, Dean Eckles, and Guido W Imbens. Exact p-values for network interference. Technical report, National Bureau of Economic Research, 2015.

[AHL+06] Ali Ahmed, Ahsan Husain, Thomas E Love, Giovanni Gambassi, Louis J Dell'Italia, Gary S Francis, Mihai Gheorghiade, Richard M Allman, Sreelatha Meleth, and Robert C Bourge. Heart failure, chronic diuretic use, and increase in mortality and hospitalization: an observational study using propensity score methods. *European heart journal*, 27(12):1431–1439, 2006.

[Aie15] Allison Aiello. Using social networks and smart phones to uncover the social determinants of respiratory infection transmission. In *143rd APHA Annual Meeting and Expo (Oct. 31-Nov. 4, 2015)*. APHA, 2015.

[AIR96] 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.

[AJR01] 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.

[AK91] 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.

[ALWB07] Julia E Aledort, Nicole Lurie, Jeffrey Wasserman, and Samuel A Bozzette. Non-pharmaceutical public health interventions for pandemic influenza: an evaluation of the evidence base. *BMC public health*, 7(1):208, 2007.

[AMP+10] Allison E Aiello, Genevra F Murray, Vanessa Perez, Rebecca M Coulborn, Brian M Davis, Monica Uddin, David K Shay, Stephen H Waterman, and Arnold S Monto. Mask use, hand hygiene, and seasonal influenza-like illness among young adults: a randomized intervention trial. *Journal of Infectious Diseases*, 201(4):491–498, 2010.

[AMS09] Sinan Aral, Lev Muchnik, and Arun Sundararajan. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences*, 106(51):21544–21549, 2009.

[AMS13] Sinan Aral, Lev Muchnik, and Arun Sundararajan. Engineering social contagions: Optimal network seeding in the presence of homophily. *Network Science*, 1(02):125–153, 2013.

[AOW14] Joshua Angrist, Philip Oreopoulos, and Tyler Williams. When opportunity knocks, who answers? new evidence on college achievement awards. *Journal of Human Resources*, 49(3):572–610, 2014.

[AS94] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB '94, pages 487–499, 1994.

[ASA15] Peter M. Aronow, Cyrus Samii, and Valentina A. Assenova. Cluster-robust variance estimation for dyadic data. *arXiv 1312.3398*, 2015.

[ASE+16] Allison E. Aiello, Amanda M. Simanek, Marisa C. Eisenberg, Alison R. Walsh, Brian Davis, Erik Volz, Caroline Cheng, Jeanette J. Rainey, Amra Uzicanin, Hongjiang Gao, Nathaniel Osgood, Dylan Knowles, Kevin Stanley, Kara Tarter, and Arnold S. Monto. Design and methods of a social network isolation study for reducing respiratory infection transmission: The ex-flu cluster randomized trial. *Epidemics*, 15:38 – 55, 2016.

[ATW19] Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized Random Forests. *The Annals of Statistics*, 47(2):1148–1178, 2019.

[BA15] Guillaume W Basse and Edoardo M Airoldi. Optimal design of experiments in the presence of network-correlated outcomes. *arXiv preprint arXiv:1507.00803*, 2015.

[BCH14] Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650, 2014.

[BCS14] Michael Baiocchi, Jing Cheng, and Dylan S Small. Instrumental variable methods for causal inference. *Statistics in medicine*, 33(13):2297–2340, 2014.

[BEYR12] Eytan Bakshy, Dean Eckles, Rong Yan, and Itamar Rosenn. Social influence in social advertising: Evidence from field experiments. In

Proceedings of the 13th ACM Conference on Electronic Commerce, pages 146–161. ACM, 2012.

[BJB95] John Bound, David A Jaeger, and Regina M Baker. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American statistical association*, 90(430):443–450, 1995.

[BN02] Frank Ball and Peter Neal. A general model for stochastic sir epidemics with two levels of mixing. *Mathematical biosciences*, 180(1):73–102, 2002.

[BNF+06] D Bell, A Nicoll, K Fukuda, P Horby, A Monto, F Hayden, C Wylks, L Sanders, and J Van Tam. Non-pharmaceutical interventions for pandemic influenza, international measures. *Emerging infectious diseases*, 12(1):81–87, 2006.

[BP02] 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.

[BPC+13] Léon Bottou, Jonas Peters, Joaquin Quinonero Candela, Denis Xavier Charles, Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Y Simard, and Ed Snelson. Counterfactual reasoning and learning systems: the example of computational advertising. *Journal of Machine Learning Research*, 14(1):3207–3260, 2013.

[BSLR10] 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.

[BSR+06] M Alan Brookhart, Sebastian Schneeweiss, Kenneth J Rothman, Robert J Glynn, Jerry Avorn, and Til Stürmer. Variable selection for propensity score models. *American journal of epidemiology*, 163(12):1149–1156, 2006.

[BVA16] Guillaume W Basse, Alexander Volfovsky, and Edoardo M Airoldi. Observational studies with unknown time of treatment. *arXiv preprint arXiv:1601.04083*, 2016.

[Car93] David Card. Using geographic variation in college proximity to estimate the return to schooling. Technical report, National Bureau of Economic Research, 1993.

[CBM+11] Simon Cauchemez, Achuyt Bhattarai, Tiffany L Marchbanks, Ryan P Fagan, Stephen Ostroff, Neil M Ferguson, David Swerdlow, , and the Pennsylvania H1N1 working group. Role of social networks in shaping disease transmission during a community outbreak of 2009 h1n1 pandemic influenza. *Proceedings of the National Academy of Sciences*, 108(7):2825–2830, 2011.

[cER11] Şeyda Ertekin and Cynthia Rudin. On equivalence relationships between classification and ranking algorithms. *Journal of Machine Learning Research*, 12:2905–2929, 2011.

[Cha47] F.S. Chapin. *Experimental Designs in Sociological Research*. Harper; New York, 1947.

[Chi13] A. Chiolero. Big data in epidemiology: too big to fail? *Epidemiology*, 24(6):938–939, Nov 2013.

[CPB16] 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.

[CR73] William G Cochran and Donald B Rubin. Controlling bias in observational studies: A review. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 417–446, 1973.

[CRC+12] Allison Chang, Cynthia Rudin, Michael Cavaretta, Robert Thomas, and Gloria Chou. How to reverse-engineer quality rankings. *Machine Learning*, 88:369–398, September 2012.

[CS10] Stephen R Cole and Elizabeth A Stuart. Generalizing evidence from randomized clinical trials to target populations the actg 320 trial. *American journal of epidemiology*, 172(1):107–115, 2010.

[CVB+08] Simon Cauchemez, Alain-Jacques Valleron, Pierre-Yves Boelle, Antoine Flahault, and Neil M Ferguson. Estimating the impact of school closure on influenza transmission from sentinel data. *Nature*, 452(7188):750–754, 2008.

[DDH13] H David, 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.

[DPR15] Tirthankar Dasgupta, Natesh S Pillai, and Donald B Rubin. Causal inference from 2k factorial designs by using potential outcomes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(4):727–753, 2015.

[DS13a] Alexis Diamond and Jasjeet S. Sekhon. Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics*, 95(3):932–945, 2013.

[DS13b] Alexis Diamond and Jasjeet S. Sekhon. Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *The Review of Economics and Statistics*, 95(3):932–945, 2013.

[EKB16] Dean Eckles, René F Kizilcec, and Eytan Bakshy. Estimating peer effects in networks with peer encouragement designs. *Proceedings of the National Academy of Sciences*, 113(27):7316–7322, 2016.

[EKU14] Dean Eckles, Brian Karrer, and Johan Ugander. Design and analysis of experiments in networks: Reducing bias from interference. *arXiv preprint arXiv:1404.7530*, 2014.

[EKWM06] WJ Edmunds, G Kafatos, J Wallinga, and JR Mossong. Mixing patterns and the spread of close-contact infectious diseases. *Emerging themes in epidemiology*, 3(1):1, 2006.

[ETW+10] KT Eames, NL Tilston, PJ White, E Adams, and WJ Edmunds. The impact of illness and the impact of school closure on social contact patterns. *Health Technol Assess*, 14(34):267–312, 2010.

[FAH15] Kai Fan, Allison E Aiello, and Katherine A Heller. Bayesian models for heterogeneous personalized health data. *arXiv preprint arXiv:1509.00110*, 2015.

[Far15] Max H Farrell. Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics*, 189(1):1–23, 2015.

[FCF+06] Neil M Ferguson, Derek AT Cummings, Christophe Fraser, James C Cajka, Philip C Cooley, and Donald S Burke. Strategies for mitigating an influenza pandemic. *Nature*, 442(7101):448–452, 2006.

[FEW+15] Kai Fan, Marisa Eisenberg, Alison Walsh, Allison Aiello, and Katherine Heller. Hierarchical graph-coupled hmms for heterogeneous personalized health data. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 239–248. ACM, 2015.

[FH14] Bailey K Fosdick and Peter D Hoff. Separable factor analysis with applications to mortality data. *The annals of applied statistics*, 8(1):120, 2014.

[FMGS15] Colin B Fogarty, Mark E Mikkelsen, David F Gaieski, and Dylan S Small. Discrete optimization for interpretable study populations and randomization inference in an observational study of severe sepsis mortality. *Journal of the American Statistical Association*, (just-accepted):1–38, 2015.

[Frö02] Markus Frölich. Nonparametric iv estimation of local average treatment effects with covariates. 2002.

[Frö07] Markus Frölich. Nonparametric iv estimation of local average treatment effects with covariates. *Journal of Econometrics*, 139(1):35–75, 2007.

[FSZ15] Jianqing Fan, Qi-Man Shao, and Wen-Xin Zhou. Are discoveries spurious? distributions of maximum spurious correlations and their applications. *arXiv preprint arXiv:1502.04237*, 2015.

[GG00] 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.

[GLM03] Steven Glazerman, Dan M Levy, and David Myers. Nonexperimental versus experimental estimates of earnings impacts. *The Annals of the American Academy of Political and Social Science*, 589(1):63–93, 2003.

[GR15] Siong Thye Goh and Cynthia Rudin. Cascaded high dimensional histograms: An approach to interpretable density estimation for categorical data. *ArXiv e-prints: arXiv:1510.06779*, 2015.

[GR18] Siong Thye Goh and Cynthia Rudin. A minimax surrogate loss approach to conditional difference estimation. *ArXiv e-prints: arXiv:1803.03769*, March 2018.

[Gre45] Ernest Greenwood. *Experimental sociology: A study in method*. King's crown Press, 1945.

[HFW14] Zonghui Hu, Dean A Follmann, and Naisyin Wang. Estimation of mean response via the effective balancing score. *Biometrika*, page asu022, 2014.

[HH08] Michael G Hudgens and M Elizabeth Halloran. Toward causal inference with interference. *Journal of the American Statistical Association*, 103(482):832–842, 2008.

[HIKS11a] Daniel Ho, Kosuke Imai, Gary King, and Elizabeth Stuart. Matchit: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software, Articles*, 42(8):1–28, 2011.

[HIKS11b] Daniel E. Ho, Kosuke Imai, Gary King, and Elizabeth A. Stuart. MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, 42(8):1–28, 2011.

[HIT98] James J Heckman, Hidehiko Ichimura, and Petra Todd. Matching as an econometric evaluation estimator. *The Review of Economic Studies*, 65(2):261–294, 1998.

[HLLBT17] 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. JMLR. org, 2017.

[HMM06] Adele V. Harrell, Douglas Marlowe, and Jeffrey Merrill. Breaking the cycle of drugs and crime in Birmingham, Alabama, Jacksonville, Florida, and Tacoma, Washington, 1997-2001. Ann Arbor, MI: Inter-university Consortium for Political and Social Research, 2006-03-30, 2006.

[Hol86] Paul W Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960, 1986.

[HR14] Jonathan H. Huggins and Cynthia Rudin. A statistical learning theory framework for supervised pattern discovery. In *In Proceedings of SIAM Conference on Data Mining (SDM)*, 2014.

[HRZ04] Jennifer L Hill, Jerome P Reiter, and Elaine L Zanutto. A comparison of experimental and observational data analyses. *Applied Bayesian modeling and causal inference from incomplete-data perspectives: An essential journey with Donald Rubin's statistical family*, pages 49–60, 2004.

[IA94] Guido W Imbens and Joshua D Angrist. Identification and estimation of local average. 1994.

[IKP11] Stefano M Iacus, Gary King, and Giuseppe Porro. Multivariate matching methods that are monotonic imbalance bounding. *Journal of the American Statistical Association*, 106(493):345–361, 2011.

[IKP12a] Stefano M Iacus, Gary King, and Giuseppe Porro. Causal inference without balance checking: Coarsened exact matching. *Political Analysis*, 20:1–24, 2012.

[IKP12b] Stefano M Iacus, Gary King, and Giuseppe Porro. Causal inference without balance checking: Coarsened exact matching. *Political analysis*, 20(1):1–24, 2012.

[IKS08] Kosuke Imai, Gary King, and Elizabeth Stuart. Misunderstandings among experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society, Series A*, 171, part 2:481–502, 2008.

[Imb00] Guido W Imbens. The role of the propensity score in estimating dose-response functions. *Biometrika*, 87(3):706–710, 2000.

[IR97] 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.

[IR13] Kosuke Imai and Marc Ratkovic. Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1):443–470, 2013.

[IR15] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

[IT01] Hidehiko Ichimura and Christopher Taber. Propensity-score matching with instrumental variables. *American economic review*, 91(2):119–124, 2001.

[JDMD+11] Tom Jefferson, Chris B Del Mar, Liz Dooley, Eliana Ferroni, Lubna A Al-Ansary, Ghada A Bawazeer, Mieke L van Driel, Sreekumaran Nair, Mark A Jones, Sarah Thorning, and John M Conly. Physical interventions to interrupt or reduce the spread of respiratory viruses. *The Cochrane Library*, 2011.

[Jos87] Paul L Joskow. Contract duration and relationship-specific investments: Empirical evidence from coal markets. *The American Economic Review*, pages 168–185, 1987.

[JPV16] Ravi Jagadeesan, Natesh Pillai, and Alexander Volfovsky. Design and analysis of randomized experiments in networks with interference. *unpublished manuscript*, 2016.

[JR99] Marshall M Joffe and Paul R Rosenbaum. Invited commentary: propensity scores. *American journal of epidemiology*, 150(4):327–333, 1999.

[JSN90] T. P. Speed Jerzy Splawa-Neyman, D. M. Dabrowska. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 5(4):465–472, 1990.

[KKA+13] 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.

[KKM+16] 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.

[KNC+11] Gary King, Richard Nielsen, Carter Coberley, James E Pope, and Aaron Wells. Comparative effectiveness of matching methods for causal inference. *Unpublished manuscript*, 15, 2011.

[KZ14] Luke Keele and Jose R Zubizarreta. Optimal multilevel matching in clustered observational studies: A case study of the school voucher system in chile. *arXiv preprint arXiv:1409.8597*, 2014.

[LDR+18] Yameng Liu, Awa Dieng, Sudeepa Roy, Cynthia Rudin, and Alexander Volfovsky. Interpretable almost-exact matching for causal inference. *arXiv e-prints: arXiv:1806.06802*, Jun 2018.

[Lic13] M. Lichman. UCI machine learning repository, 2013.

[LRM13] Benjamin Letham, Cynthia Rudin, and David Madigan. Sequential event prediction. *Machine Learning*, 93:357–380, 2013.

[LRMM15] Benjamin Letham, Cynthia Rudin, Tyler H. McCormick, and David Madigan. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *Annals of Applied Statistics*, 9(3):1350–1371, 2015.

[Man13] Charles F Manski. Identification of treatment response with social interactions. *The Econometrics Journal*, 16(1):S1–S23, 2013.

[MGMS10] Alexandra Meliou, Wolfgang Gatterbauer, Katherine F. Moore, and Dan Suciu. The complexity of causality and responsibility for query answers and non-answers. *PVLDB*, 4(1):34–45, 2010.

[MHJ+08] Joël Mossong, Niel Hens, Mark Jit, Philippe Beutels, Kari Auranen, Rafael Mikolajczyk, Marco Massari, Stefania Salmaso, Gianpaolo Scalia Tomba, Jacco Wallinga, Janneke Heijne, Malgorzata Sadkowska-Todys, Magdalena Rosinska, and W. John Edmunds. Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Med*, 5(3):e74, 2008.

[Mic16a] Microsoft SQL Server. https://www.microsoft.com/en-us/sql-server/sql-server-2016. 2016.

[Mic16b] Microsoft SQL Server. *Microsoft SQL Server*, 2016.

[MJB07] Paul W Mielke Jr and Kenneth J Berry. *Permutation methods: a distance function approach*. Springer Science & Business Media, 2007.

[MJG+11] Alessia Melegaro, Mark Jit, Nigel Gay, Emilio Zagheni, and W John Edmunds. What types of contacts are important for the spread of infections? using contact survey data to explore european mixing patterns. *Epidemics*, 3(3):143–151, 2011.

[MNEAR18] Marco Morucci, Md. Noor-E-Alam, and Cynthia Rudin. Hypothesis tests that are robust to choice of matching method. *ArXiv e-prints, arXiv:1812.02227*, December 2018.

[MRM12] Tyler H. McCormick, Cynthia Rudin, and David Madigan. Bayesian hierarchical modeling for predicting medical conditions. *The Annals of Applied Statistics*, 6(2):652–668, 2012.

[MWES15] S. J. Mooney, D. J. Westreich, and A. M. El-Sayed. Commentary: Epidemiology in the era of big data. *Epidemiology*, 26(3):390–394, May 2015.

[NCH] NCHS. User guide to the 2010 natality public use file.

[NEAR15a] M. Noor-E-Alam and C. Rudin. Robust nonparametric testing for causal inference in natural experiments. *Working paper*, 2015.

[NEAR15b] M. Noor-E-Alam and C. Rudin. Robust testing for causal inference in natural experiments. *Working paper*, 2015.

[NEAR15c] Md. Noor-E-Alam and Cythia Rudin. Robust nonparametric testing for causal inference in observational studies. *Optimization Online*, Dec, 2015.

[NP03] Whitney K Newey and James L Powell. Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578, 2003.

[ORR15] 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.

[OV16] Elizabeth L. Ogburn and Alexander Volfovsky. Statistics for networks and causal inference. In *Handbook of Big Data.*, pages 171–190. 2016.

[Pon18] 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.

[Pos16] PostgreSQL. *PostgreSQL*, 2016.

[PRV18] Harsh Parikh, Cynthia Rudin, and Alexander Volfovsky. MALTS: Matching After Learning to Stretch. *arXiv e-prints: arXiv:1811.07415*, Nov 2018.

[PS15] D. U. Pfeiffer and K. B. Stevens. Spatial and temporal epidemiological analysis in the Big Data era. *Prev. Vet. Med.*, 122(1-2):213–220, Nov 2015.

[RKH+15] Michelle E Ross, Amanda R Kreider, Yuan-Shung Huang, Meredith Matone, David M Rubin, and A Russell Localio. Propensity score methods for analyzing observational data like randomized experiments: challenges and solutions for rare outcomes and exposures. *American journal of epidemiology*, 181(12):989–995, 2015.

[RLM13] Cynthia Rudin, Benjamin Letham, and David Madigan. Learning theory analysis for association rules and sequential event prediction. *Journal of Machine Learning Research*, 14:3384–3436, 2013.

[Ros10] Paul R Rosenbaum. *Design of observational studies*, volume 10. Springer, 2010.

[Ros12] Paul R Rosenbaum. Interference between units in randomized experiments. *Journal of the American Statistical Association*, 2012.

[Ros16] Paul R Rosenbaum. Imposing minimax and quantile constraints on optimal matching in observational studies. *Journal of Computational and Graphical Statistics*, 26(1), 2016.

[RR83] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[RR84] Paul R Rosenbaum and Donald B Rubin. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American statistical Association*, 79(387):516–524, 1984.

[RR85] Paul R Rosenbaum and Donald B Rubin. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1):33–38, 1985.

[RRB+09] Carla V Rodriguez, Krista Rietberg, Atar Baer, Tao Kwan-Gett, and Jeffrey Duchin. Association between school closure and subsequent absenteeism during a seasonal influenza epidemic. *Epidemiology*, 20(6):787–792, 2009.

[RS12] Jeremy A Rassen and Sebastian Schneeweiss. Using high-dimensional propensity scores to automate confounding control in a distributed medical product safety surveillance system. *Pharmacoepidemiology and Drug Safety*, 21(S1):41–49, 2012.

[RS14] Sudeepa Roy and Dan Suciu. A formal approach to finding explanations for database queries. In *International Conference on Management of Data, SIGMOD 2014, Snowbird, UT, USA, June 22-27, 2014*, pages 1579–1590, 2014.

[RT92] Donald B Rubin and Neal Thomas. Characterizing the effect of matching using linear propensity score methods with normal distributions. *Biometrika*, 79(4):797–809, 1992.

[RT96] Donald B Rubin and Neal Thomas. Matching using estimated propensity scores: relating theory to practice. *Biometrics*, pages 249–264, 1996.

[RT00] Donald B Rubin and Neal Thomas. Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association*, 95(450):573–585, 2000.

[Rub73a] Donald B Rubin. Matching to remove bias in observational studies. *Biometrics*, 29(1):159–183, March 1973.

[Rub73b] Donald B Rubin. The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics*, 29(1):185–203, March 1973.

[Rub74] Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

[Rub76] Donald B Rubin. Multivariate matching methods that are equal percent bias reducing, I: Some examples. *Biometrics*, 32(1):109–120, March 1976.

[Rub78] Donald B Rubin. Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, pages 34–58, 1978.

[Rub80a] Donald B. Rubin. Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591–593, 1980.

[Rub80b] Donald B Rubin. Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591–593, 1980.

[Rub90] Donald B Rubin. Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science*, 5(4):472–480, 1990.

[Rub01] Donald B Rubin. Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services and Outcomes Research Methodology*, 2(3-4):169–188, 2001.

[Rub05] Donald B. Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100:322–331, 2005.

[Rub06] Donald B Rubin. *Matched sampling for causal effects*. Cambridge University Press, 2006.

[Rub08] Donald B Rubin. For objective causal inference, design trumps analysis. *The Annals of Applied Statistics*, pages 808–840, 2008.

[Rub11] Donald B Rubin. Causal inference using potential outcomes. *Journal of the American Statistical Association*, 2011.

[RZ16] María Resa and José R Zubizarreta. Evaluation of subset matching methods and forms of covariate balance. *Statistics in Medicine*, 2016.

[SCPS17] Babak Salimi, Corey Cole, Dan R. K. Ports, and Dan Suciu. Zaliql: Causal inference from observational data at scale. *PVLDB*, 10(12):1957–1960, 2017.

[SCS08] William R Shadish, Margaret H Clark, and Peter M Steiner. Can nonrandomized experiments yield accurate answers? a randomized experiment comparing random and nonrandom assignments. *Journal of the American Statistical Association*, 103(484):1334–1344, 2008.

[SJS14] Jason J Sauppe, Sheldon H Jacobson, and Edward C Sewell. Complexity and approximation results for the balance optimization subset selection model for causal inference in observational studies. *INFORMS Journal on Computing*, 26(3):547–566, 2014.

[Sob06] Michael E Sobel. What do randomized studies of housing mobility demonstrate? causal inference in the face of interference. *Journal of the American Statistical Association*, 101(476):1398–1407, 2006.

[SR08] Dylan S Small and Paul R Rosenbaum. War and wages: the strength of instrumental variables and their sensitivity to unobserved biases. *Journal of the American Statistical Association*, 103(483):924–933, 2008.

[SRG+09] Sebastian Schneeweiss, Jeremy A Rassen, Robert J Glynn, Jerry Avorn, Helen Mogun, and M Alan Brookhart. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology (Cambridge, Mass.)*, 20(4):512, 2009.

[SS17] Ricardo Silva and Shohei Shimizu. Learning instrumental variables with structural and non-gaussianity assumptions. *Journal of Machine Learning Research*, 18(120):1–49, 2017.

[Stu10] Elizabeth A Stuart. Matching methods for causal inference: A review and a look forward. *Statistical Science: a Review Journal of the Institute of Mathematical Statistics*, 25(1):1, 2010.

[SVA15] Daniel L Sussman, Alexander Volfovsky, and Edoardo M Airoldi. Analyzing statistical and computational tradeoffs of estimation procedures. *arXiv preprint arXiv:1506.07925*, 2015.

[SWY02] 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.

[TAGT14] Lu Tian, Ash A Alizadeh, Andrew J Gentles, and Robert Tibshirani. A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association*, 109(508):1517–1532, 2014.

[Tan06] Zhiqiang Tan. Regression and weighting methods for causal inference using instrumental variables. *Journal of the American Statistical Association*, 101(476):1607–1618, 2006.

[TR13] Theja Tulabandhula and Cynthia Rudin. Machine learning with operational costs. *Journal of Machine Learning Research*, 14:1989–2028, 2013.

[TR14a] Theja Tulabandhula and Cynthia Rudin. Generalization bounds for learning with linear, polygonal, quadratic and conic side knowledge. *Machine Learning*, pages 1–34, 2014.

[TR14b] Theja Tulabandhula and Cynthia Rudin. On combining machine learning with decision making. *Machine Learning (ECML-PKDD journal track)*, 97(1-2):33–64, 2014.

[TR14c] Theja Tulabandhula and Cynthia Rudin. Robust optimization using machine learning for uncertainty sets. In *Proceedings of the International Symposium on Artificial Intelligence and Mathematics (ISAIM)*, 2014.

[TV12] Eric J Tchetgen Tchetgen and Tyler J VanderWeele. On causal inference in the presence of interference. *Statistical Methods in Medical Research*, 21(1):55–75, 2012.

[TVA15] Panos Toulis, Alexander Volfovsky, and Edoardo Airoldi. Causal effects of professional networking on labor mobility. *unpublished manuscript*, 2015.

[UKBK13] Johan Ugander, Brian Karrer, Lars Backstrom, and Jon Kleinberg. Graph cluster randomization: Network exposure to multiple universes. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 329–337. ACM, 2013.

[USC90] UCI Census Data. https://archive.ics.uci.edu/ml/datasets/US+Census+Data+(1990), 1990.

[VA16] Alexander Volfovsky and Edoardo M Airoldi. Sharp total variation bounds for finitely exchangeable arrays. *Statistics & Probability Letters*, 114:54–59, 2016.

[VAR15] Alexander Volfovsky, Edoardo M Airoldi, and Donald B Rubin. Causal inference for ordinal outcomes. *arXiv preprint arXiv:1501.01234*, 2015.

[VH15] Alexander Volfovsky and Peter D Hoff. Testing for nodal dependence in relational data matrices. *Journal of the American Statistical Association*, (just-accepted):00–00, 2015.

[VKHEE13] Kim Van Kerckhove, Niel Hens, W John Edmunds, and Ken TD Eames. The impact of illness on social networks: implications for transmission and control of influenza. *American journal of epidemiology*, 178(11):1655–1662, 2013.

[WA17] Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, (just-accepted), 2017.

[WF15] Hadley Wickham and Romain Francois. *dplyr: A Grammar of Data Manipulation*, 2015. R package version 0.4.3.

[Woo10] Jeffrey M Wooldridge. *Econometric analysis of cross section and panel data*. MIT press, 2010.

[WRDV+15] Tong Wang, Cynthia Rudin, Finale Doshi-Velez, Yimin Liu, Erica Klampfl, and Perry MacNeille. Or's of and's for interpretable classification, with application to context-aware recommender systems. *arXiv preprint arXiv:1504.07614*, 2015.

[WRRV17] Tianyu Wang, Sudeepa Roy, Cynthia Rudin, and Alexander Volfovsky. FLAME: A fast large-scale almost matching exactly approach to causal inference. *arXiv e-prints: arXiv:1707.06315*, July 2017.

[WRVR17] Tianyu Wang, Cynthia Rudin, Alexander Volfovsky, and Sudeepa Roy. Flame: A fast large-scale almost matching exactly approach to causal inference. *arXiv preprint arXiv:1707.06315*, 2017.

[ZPR14] José R Zubizarreta, Ricardo D Paredes, and Paul R Rosenbaum. Matching for balance, pairing for heterogeneity in an observational study of the effectiveness of for-profit and not-for-profit high schools in chile. *The Annals of Applied Statistics*, 8(1):204–231, 2014.

[Zub12] José R Zubizarreta. Using mixed integer programming for matching in an observational study of kidney failure after surgery. *Journal of the American Statistical Association*, 107(500):1360–1371, 2012.

[Zub15] José R Zubizarreta. Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110(511):910–922, 2015.