Interpretable and scalable causal analysis for data with discrete covariates

-- Paper Review of FLAME and DAME

Yiyang Sun, Zhehan Qu

Department of Electrical and Computer Engineering

Department of Computer Science

Duke

Presentation Outline

- Casual Treatment Effect Estimation, Matching and Previous Works
- Almost Matching Exactly (AME)
- Dynamic Almost Matching Exactly (DAME) Algorithm
- Fast Large-Scale Almost Matching Exactly Algorithm (FLAME)
- Simulation and Performance Comparison
- Conclusion and Limitations

Big idea

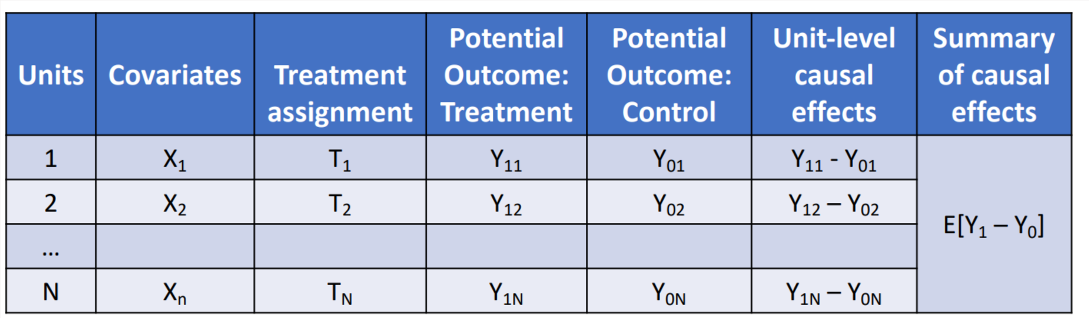
- **Goal:** Match the treatment and control units "almost exactly" based on categorical covariates
- **Methods:**
  - **DAME:** Considering all "needed" combinations of covariates for matching
  - **FLAME:** Considering the covariates based on their feature importance
- **Key ideas:**
  - Match exactly on the selected largest set of good covariates that together predict the outcome well.
  - Use the ML methods to determines the prediction quality of a set of covariates.
  - FLAME can handle large dataset

(Conditional) Average Treatment Effect Estimation (ATE/CATE), Matching and Previous Works

General Treatment Effect Calculation

Average Treatment Effect (ATE)

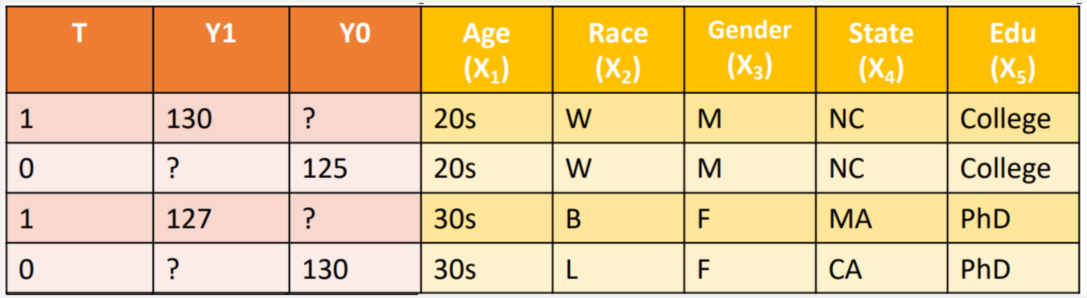
- N units, each unit could be exposed or not to a treatment T
- Casual Effect on unit i of treatment T defined as the difference between the outcome
- $ATE = \tau = E\left[Y_i(1) - Y_{i(0)}\right]$



Units	Covariates	Treatment assignment	Potential Outcome: Treatment	Potential Outcome: Control	Unit-level causal effects	Summary of causal effects
1	X_1	T_1	Y_{11}	Y_{01}	Y_{11} - Y_{01}	$E[Y_1 – Y_0]$
2	X_2	T_2	Y_{12}	Y_{02}	Y_{12} – Y_{02}	
…						
N	X_n	T_N	Y_{1N}	Y_{0N}	Y_{1N} – Y_{0N}	

Conditional Average Treatment Effect (CATE)

- Measure the average treatment effect on a **subpopulation**
- $CATE = \tau(X) = E_{i:x_i \in X}[\tau_i]$



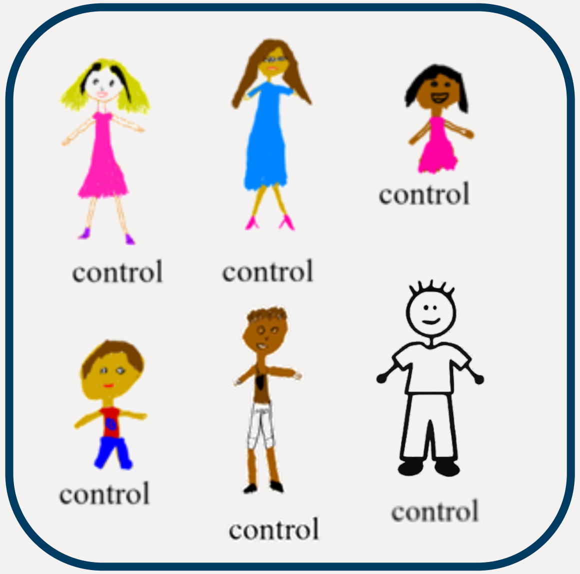
T	Y1	Y0	Age (X_1)	Race (X_2)	Gender (X_3)	State (X_4)	Edu (X_5)
1	130	?	20s	W	M	NC	College
0	?	125	20s	W	M	NC	College
1	127	?	30s	B	F	MA	PhD
0	?	130	30s	L	F	CA	PhD

Duke

To calculate ATE and CATE, the most straightforward way is to use the exact matching methods.

In observational study, we tend to find “identical twins” that share exactly the same covariates.

$$ATE = \tau = E\left[Y_i(1) - Y_{i(0)}\right]$$



Control Group

treated
treated
treated
treated
treated
treated

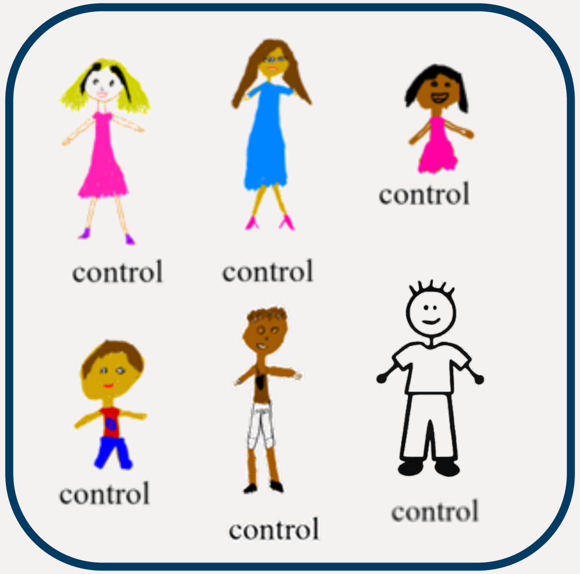
Treated Group

Image from the talk given by Cynthia, Alexander and Sudeepa; https://www.youtube.com/watch?v=-So_cL-eMFQ

To calculate ATE and CATE, the most straightforward way is to use the exact matching methods.

In observational study, we tend to find “identical twins” that share exactly the same covariates.

$$ATE = \tau = E\left[Y_i(1) - Y_{i(0)}\right]$$

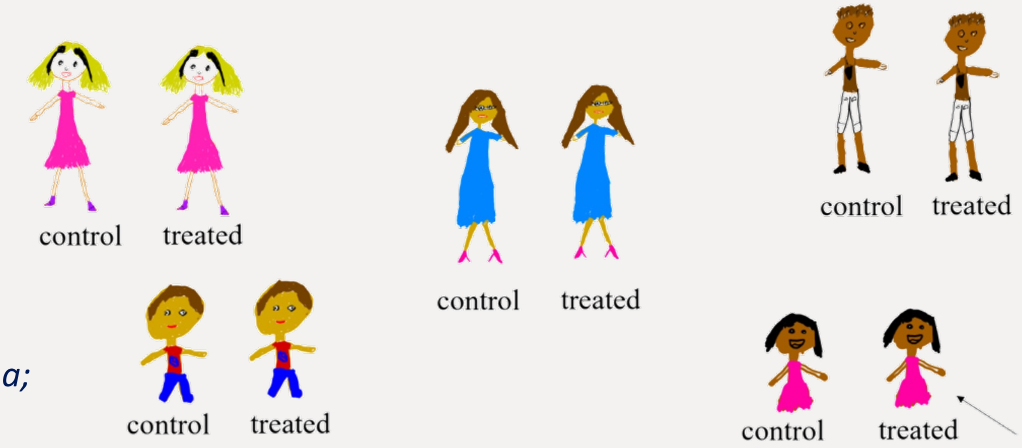


Control Group

treated
treated
treated
treated
treated
treated

Treated Group

Match the twins!



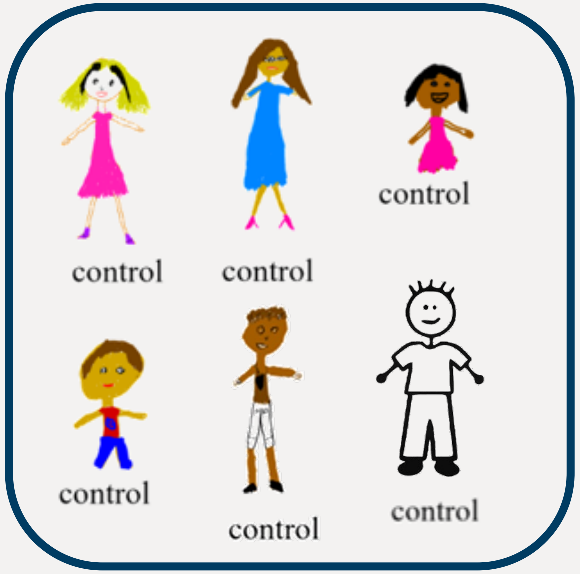
Duke

Image from the talk given by Cynthia, Alexander and Sudeepa; https://www.youtube.com/watch?v=-So_cL-eMFQ

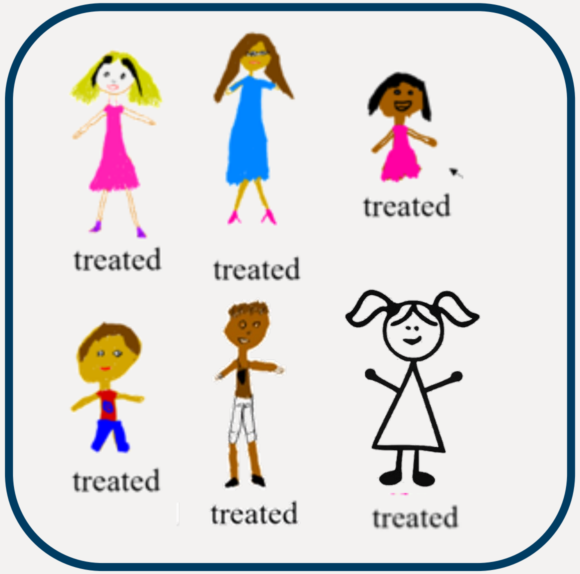
To calculate ATE and CATE, the most straightforward way is to use the exact matching methods.

We can also calculate the conditional average treatment effect for subgroups.

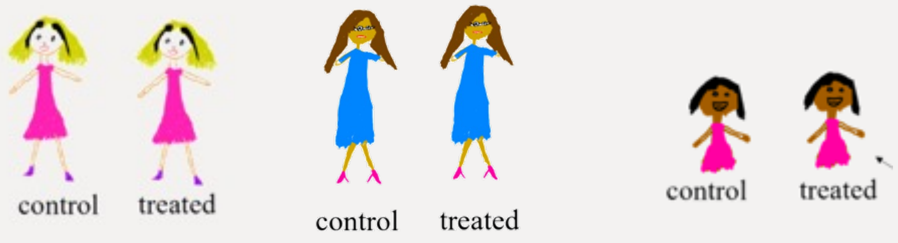
$$CATE = \tau(X) = E_{i:x_i \in X}[\tau_i]$$



Control Group



Treated Group



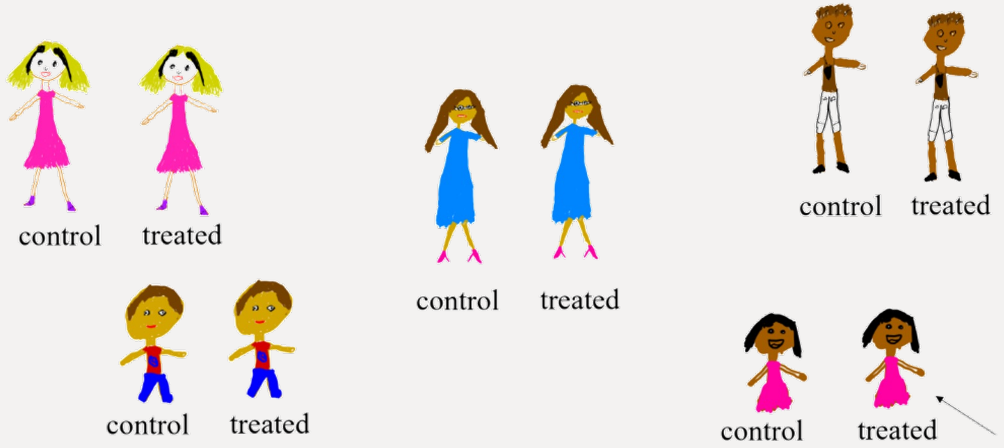
Conditional Average Treatment Effect on Female

The advantages of exactly matching methods

- **Unit-wise**
  - provide crucial information on who benefits from treatment most
- **Explanability**
  - provide explanations for treatment effects estimates
- **Explainable Feature Selection**
  - determine what type of additional data must be collected
  - It is especially important for the calculation of CATE

The disadvantage of Exact Matching Methods

- In observational study, hard to find “identical twins” that share exactly the same covariates
- Yet without exact matching, have to search for other ways like dimension reduction and optimization.



Current research on matching methods

Dimension Reduction

- Propensity Score Matching
- Doubly Robust Model
- Neural Network-based

Extension to exact matching

- Coarsened exact matching

Distance Metric Related

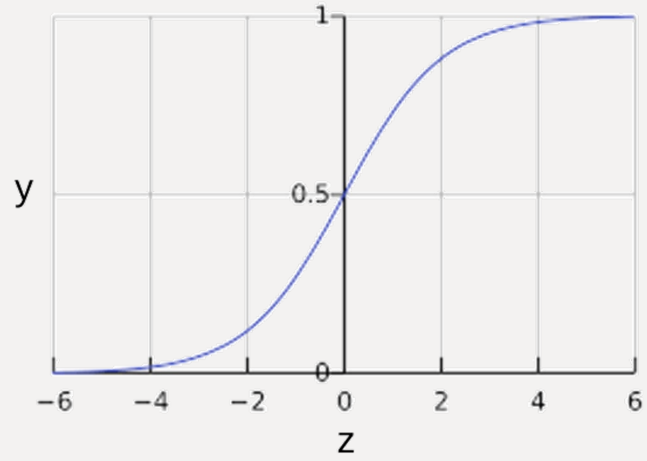
- Optimal matching and network flow optimization
- Mixed Integer Programming (MIP)

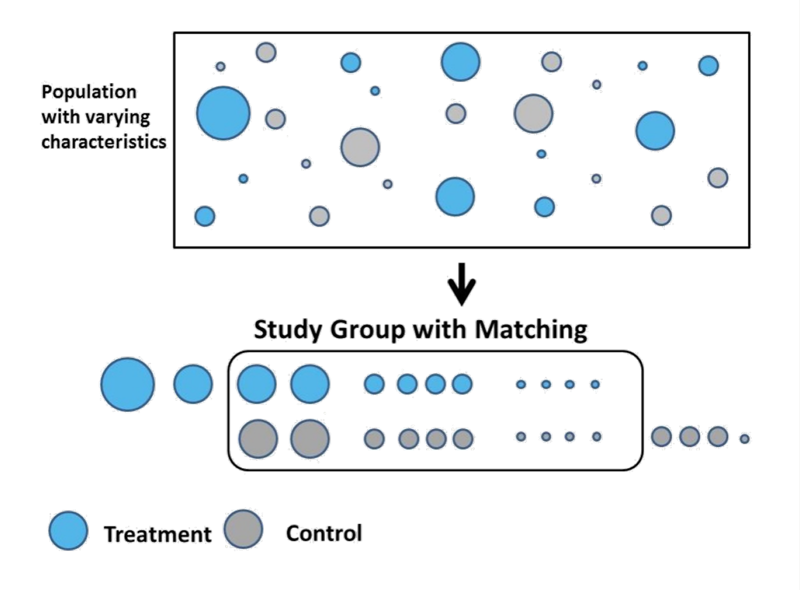
The current research on matching methods

- Propensity Score Matching (Rosenbaum and Rubin, 1983)
  - Using the probability of treatment trained by the known covariates using logistics regression models and finding the nearest samples in the treatment (make sure it is 1 to 1 or 1 to N).
  - Problem: The matching methods only depends on the treatment and ignores the outcome.

$$e(X) = \Pr(Z = 1|X) = E\,(Z|X)$$

$$\Pr(Z_i = 1|X) = \frac{exp(\beta X_i)}{1 + \exp(\beta X_i)}$$

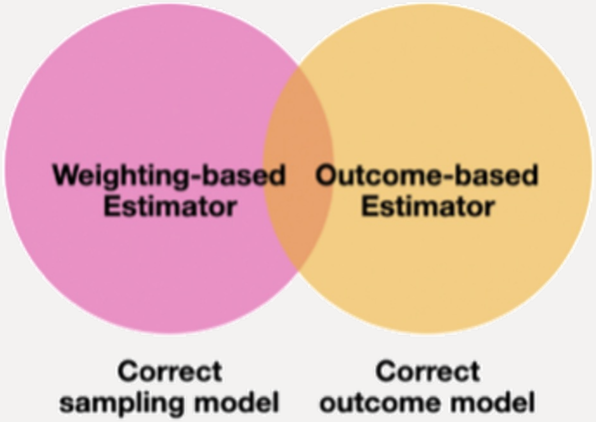




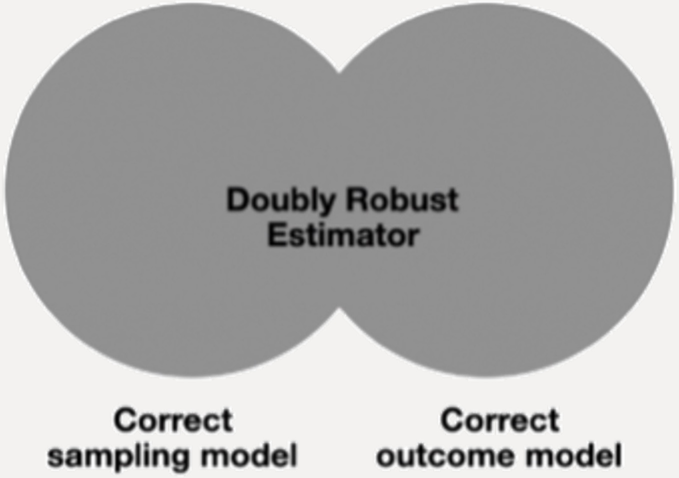
Population with varying characteristics
Study Group with Matching
Treatment
Control

[1] https://developers.google.com/machine-learning/crash-course/logistic-regression/calculating-a-probability
[2] https://www.researchgate.net/figure/An-illustration-explaining-the-Propensity-Score-Matching-model-Note-figure-does-not_fig1_361733978

The current research on matching methods



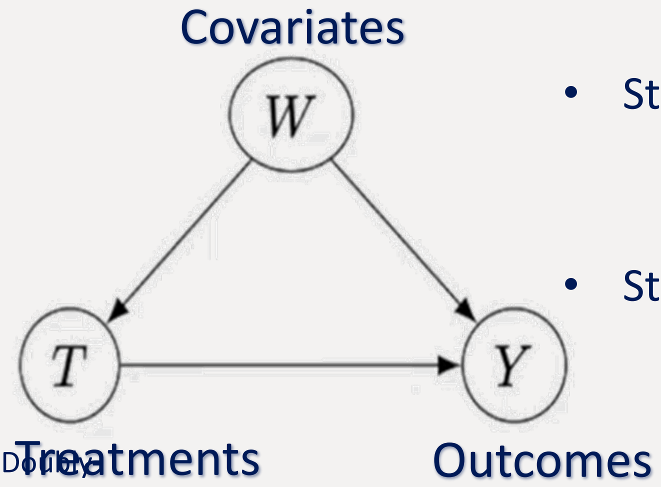
(a) Weighting- and Outcome-based Estimators:



(b) Doubly Robust Estimator:

https://www.researchgate.net/figure/Properties-of-Doubly-Robust-Estimator_fig3_364303676

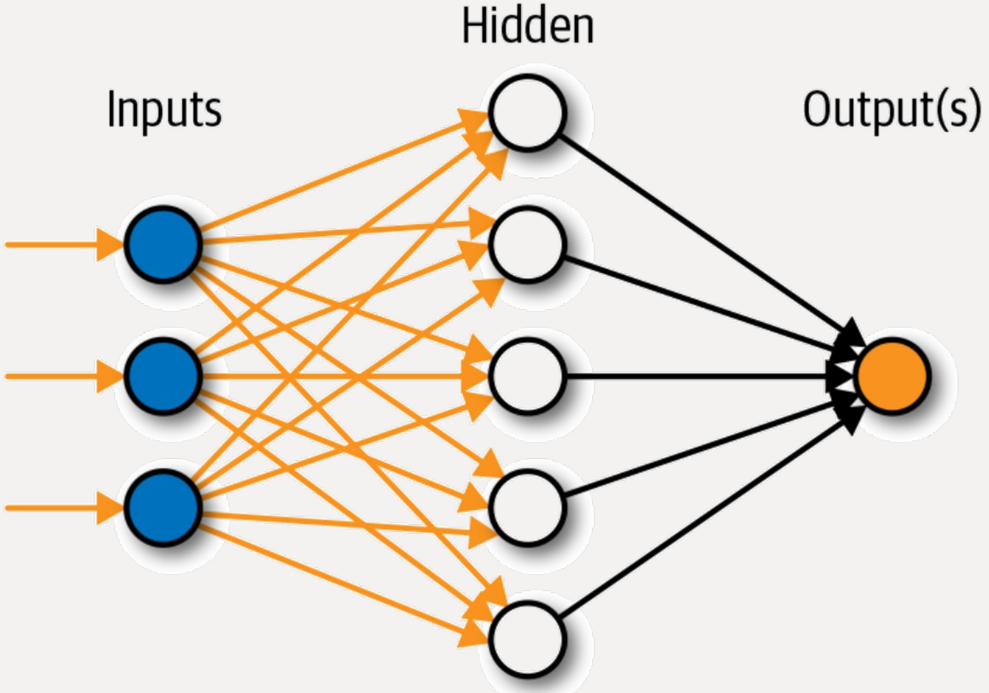
- Doubly Robust Model (Hahn,2004)
  - The correct label is based on both models on outcomes and propensity score, as one of the models has to be correct to be an unbiased estimate.
  - Problem: Cannot be used for estimating CATE (units within the matched groups often differ on important covariates)



- Stage 1:
  - Fit a model to predict Y from W get $\hat{Y}$
  - Fit a model to predict T from W get $\hat{T}$
- Stage 2:
  - Partial out W by fitting a model to predict $Y - \hat{Y}$ from $T - \hat{T}$

The current research on matching methods

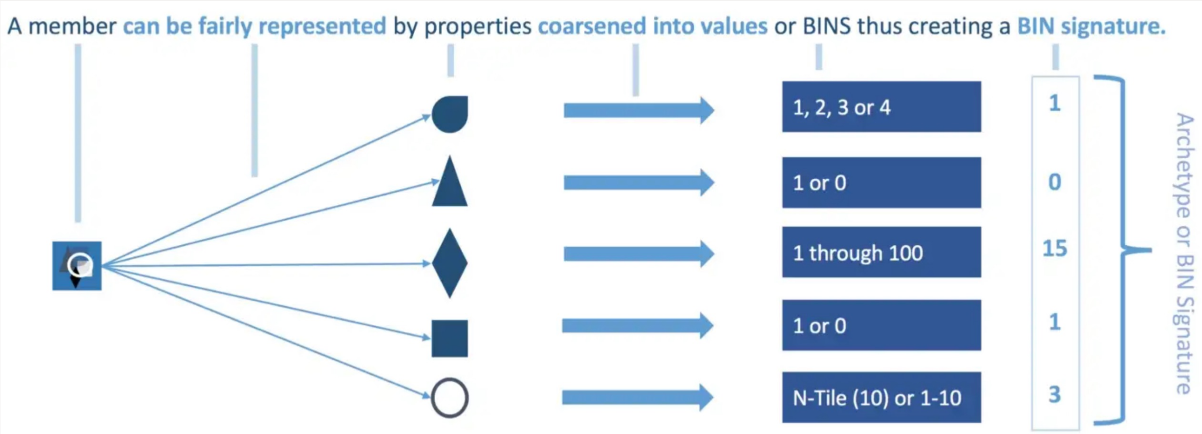
- Neural Network-based Model (Schneeweiss, 2012)
  - Use the neural network perspective to do the dimension reduction.
  - Problem: Neural networks cease to be interpretable, matches are no longer meaningful.



https://www.whyofai.com/blog/ai-explained

Duke

The current research on matching methods

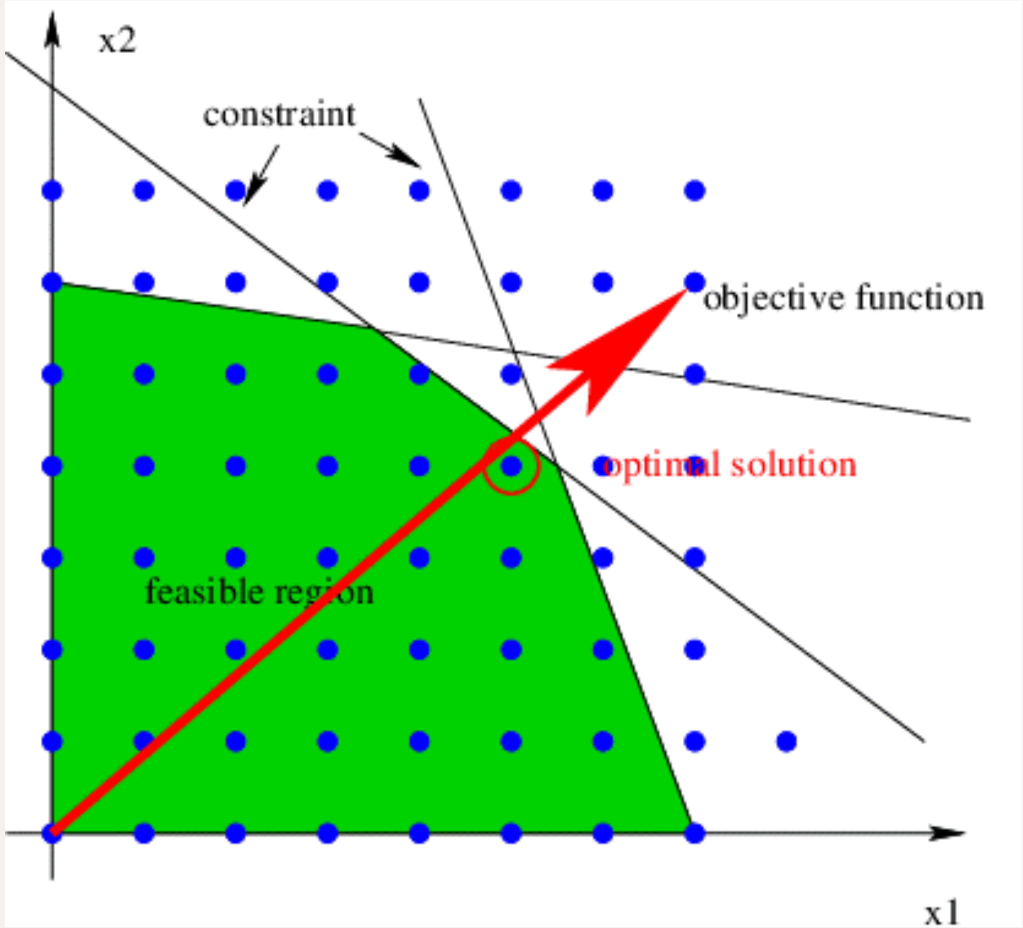


A member can be fairly represented by properties coarsened into values or BINS thus creating a BIN signature.

- Propensity Score Matching
- Doubly Robust Model
- Neural Network-based
- Coarsened exact matching(SM Iacus, 2012)
  - coarsening or discretizing covariates in such a way that the newly constructed covariates allow for exact matching
  - Problem: For discrete, or binary variables, coarsening is equivalent to variable selection, and variable selection is hard.
- Optimal matching and network flow optimization highlight

The current research on matching methods

- Propensity Score Matching
- Doubly Robust Model
- Neural Network-based
- Coarsened exact matching
- Optimal matching (Rosenbaum, 1989)
  - A distance metric over variables is defined manually, and used as input to a network flow problem which optimizes match quality
  - Problem: cannot handle constraints
- Mixed Integer Programming : consider all possible reasonable distance metrics (Zubizarreta,2012)
  - Slow and irrelativity variables included (toenail problem, will discuss later)



Duke

Almost Matching Exactly (AME)

Matching with a Distance Metric

- We want to find a match for each treatment unit t that matches at least one control unit on as many **relevant** covariates as possible.
  - Consider finding matches as finding the closest unit using a certain distance metric
  - Irrelevant covariates: might lead to the “toenail problem”

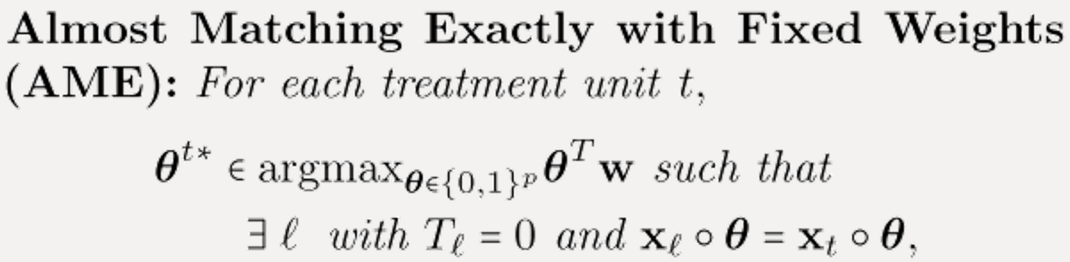
Covariates	Age	Heart Conditions	Toenail Length	Eyeball width
Treated P1	50	1011	1.5cm	2cm
Controlled P2	50	1011	14cm	1cm

- Naïve or pre-defined distance metrics do not have domain knowledge and may let irrelevant covariates dominate the distance.
- Learned distance metrics could learn to lay close to zero weight on toenail covariates.

Basic Assumptions

- Stable Unit Treatment Value (SUTVA): The treatment of one unit does not affect the outcome of another unit.
- Overlap of Support: It requires that there is sufficient overlap in the covariates of the treated and untreated units so that they are comparable.
- Strong Ignorability: $Y^{(1)}, Y^{(0)} \perp T \mid X^{REL}$ The treatment assignment is independent of the potential outcomes given the observed covariates.
- For irrelevant covariates: $Y^{(1)}, Y^{(0)} \perp T \mid X^{IRR}\ and\ T \perp X^{IRR} \mid X^{REL}$
    - The estimated ATE is $E[Y^{(1)} - Y^{(0)} | X^{REL}] = E[Y^{(1)} - Y^{(0)} | X]$

Almost Exact Matching (AME) with Fixed Weights



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

$$\boldsymbol{\theta}^{t*} \in \text{argmax}_{\boldsymbol{\theta}\in\{0,1\}^p} \boldsymbol{\theta}^T \mathbf{w} \ \textit{such that}$$
$$\exists\, \ell \ \ \textit{with}\ T_\ell = 0 \ \textit{and}\ \mathbf{x}_\ell \circ \boldsymbol{\theta} = \mathbf{x}_t \circ \boldsymbol{\theta},$$

θ	x_i	x'_i
1	Male	Male
1	Blue	Blue
0	20	*
1	Yes	Yes
0	No	*

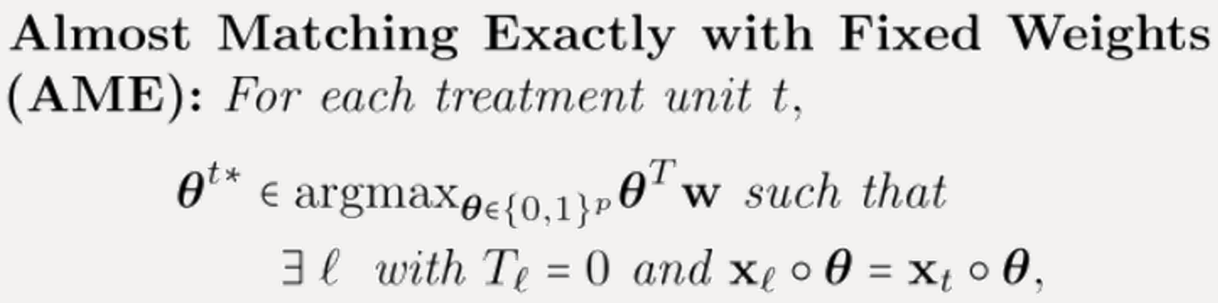
- *Denote p: number of covariates*
- *Let θ* $\in \{0,1\}^p$: a subset of covariates to match on
  - Relevance of covariate j is denoted by $w_j \geq 0$. For now let's just say it's known beforehand.
  - Valid matched groups contain at least one control unit

Duke

Dynamic Almost Matching Exactly (DAME)

Consider Only the Case of DISCRETE Covariates From Now on

Monotonicity of θ^* in AME solutions



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

$$\boldsymbol{\theta}^{t*} \in \text{argmax}_{\boldsymbol{\theta}\in\{0,1\}^p} \boldsymbol{\theta}^T \mathbf{w} \text{ such that}$$
$$\exists\, \ell \text{ with } T_\ell = 0 \text{ and } \mathbf{x}_\ell \circ \boldsymbol{\theta} = \mathbf{x}_t \circ \boldsymbol{\theta},$$

- Any feasible vectors θ' such that $\theta' < \theta$ elementwise will have $\theta'^T w \leq \theta^T w$
    - Start from θ being all 1's, and drop one element to zero at a time, then two, then three
- Consequently, consider feasible vectors θ and θ'. Define $\tilde{\theta}$ as the elementwise $\min(\theta,\ \theta')$. Then $\tilde{\theta}w < \theta^T w$, and $\tilde{\theta}^T w < \theta'^T w$
    - Must evaluate both θ and θ' as possible AME solutions before evaluating $\tilde{\theta}$
- *For covariates set {1,2,3}, we should drop {1} and {2} before we drop {1,2}*

The DAME Algorithm (with Fixed Weights)

- Designed on monotonicity property and ideas from *apriori algorithm* (Agrawal and Srikant, 1994)
  - *apriori algorithm:* Bottom-up search for frequent set mining. Frequent subsets are extended one item at a time
  - Consider “frequent” as appearing in at least 3 transactions

Itemsets
{1,2,3,4}
{1,2,4}
{1,2}
{2,3,4}
{2,3}
{3,4}
{3,4}

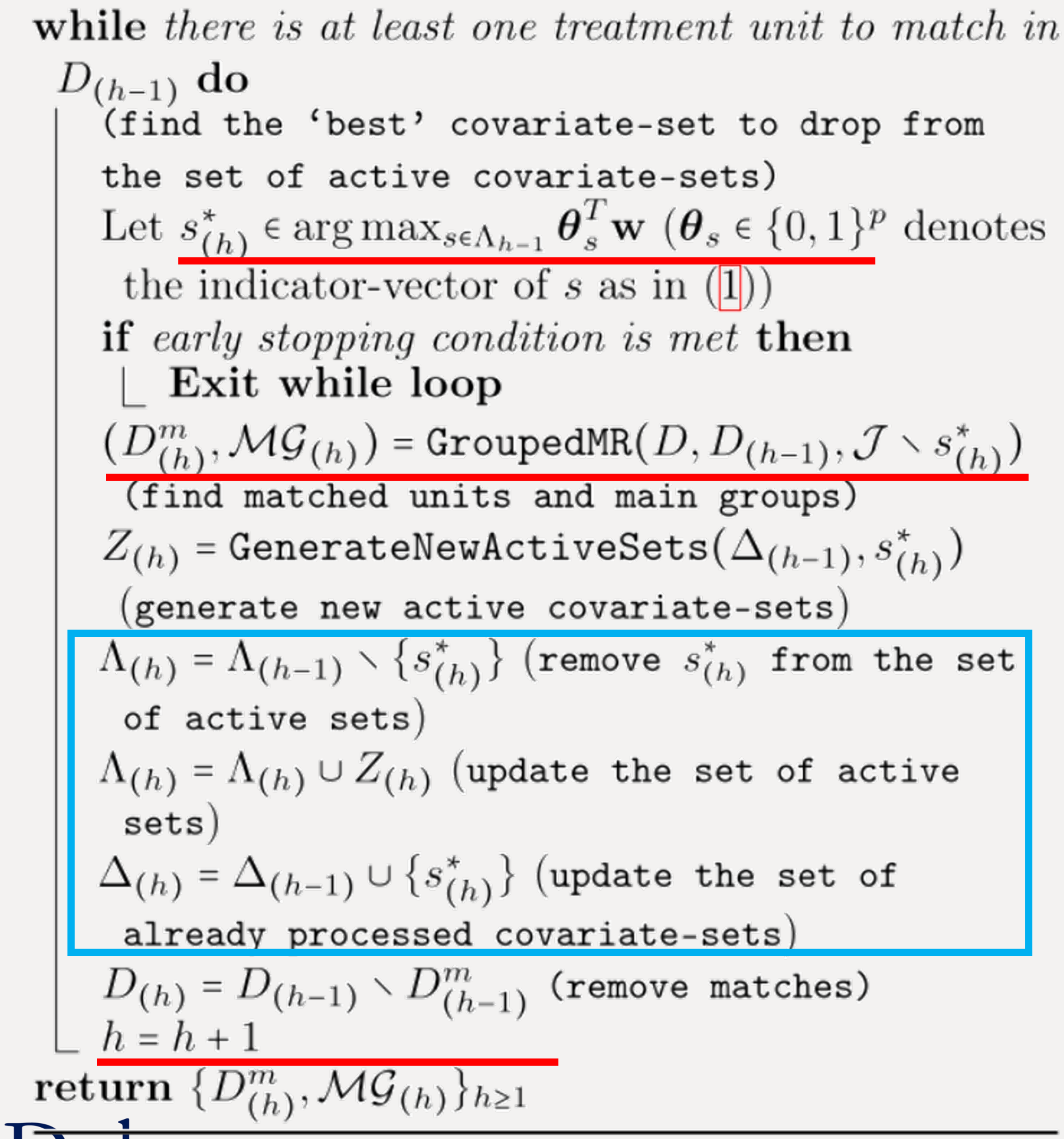
Item	Support
{1}	3
{2}	6
{3}	4
{4}	5

Item	Support
{1,2}	3
{1,3}	1
{1,4}	2
{2,3}	3
{2,4}	4
{3,4}	3

Item	Support
{2,3,4}	2

The DAME Algorithm (with Fixed Weights)

- Designed on monotonicity property and ideas from *apriori algorithm* (Agrawal and Srikant, 1994)
  - Consider bottom-up search on the set of covariates we drop
- $\mathcal{J}$: the original set of all covariates; $p = |\mathcal{J}|$
- s: the set of covariates we drop, meaning we are matching on $\mathcal{J} \setminus s$
- $\theta_s \in \{0, 1\}^p$: indicator-vector, $\theta_{s,j} = \mathbb{1}_{\{j \notin s\}, \forall j \in \{1,2,\ldots,p\}}$
  - the value is 1 if the covariate is not in s, implying that it is being used for matching
- $\mathcal{MG}_{(h)}$: matched groups at the end of iteration h
- $\Lambda_{(h)}$: active covariate-sets that are eligible to be dropped at iteration h
- $\Delta_{(h)}$: processed covariates at iteration h
- Full dataset D and the unmatched subset at iteration h: $D_{(h)}$



```
while there is at least one treatment unit to match in $D_{(h-1)}$ do
    (find the 'best' covariate-set to drop from the set of active covariate-sets)
    Let $s^*_{(h)} \in \arg\max_{s \in \Lambda_{h-1}} \boldsymbol{\theta}_s^T \mathbf{w}$ ($\boldsymbol{\theta}_s \in \{0,1\}^p$ denotes the indicator-vector of $s$ as in (1))
    if early stopping condition is met then
        Exit while loop
    $(D^m_{(h)}, \mathcal{MG}_{(h)})$ = GroupedMR($D, D_{(h-1)}, \mathcal{J} \setminus s^*_{(h)}$) (find matched units and main groups)
    $Z_{(h)}$ = GenerateNewActiveSets($\Delta_{(h-1)}, s^*_{(h)}$) (generate new active covariate-sets)
    $\Lambda_{(h)} = \Lambda_{(h-1)} \setminus \{s^*_{(h)}\}$ (remove $s^*_{(h)}$ from the set of active sets)
    $\Lambda_{(h)} = \Lambda_{(h)} \cup Z_{(h)}$ (update the set of active sets)
    $\Delta_{(h)} = \Delta_{(h-1)} \cup \{s^*_{(h)}\}$ (update the set of already processed covariate-sets)
    $D_{(h)} = D_{(h-1)} \setminus D^m_{(h-1)}$ (remove matches)
    $h = h + 1$
return $\{D^m_{(h)}, \mathcal{MG}_{(h)}\}_{h \geq 1}$
```

Find the optimal s within the current active set

Note that we need more than the unmatched units to find newly matched units/groups

Update the processed and active covariate-sets

The remaining unmatched units after iteration h

A DAME Matching Example

all covariates $\mathcal{J}$; $p = |\mathcal{J}| = 5$

T	Covariates
0	2, 3, 5, 6, 9
1	2, 3, 5, 6, 8
1	2, 3, 5, 6, 7
0	2, 5, 5, 2, 8
1	2, 3, 5, 3, 9
1	2, 5, 5, 6, 10
0	2, 5, 5, 3, 7
…	…

Start with:

$\Lambda_{(h)} = \{\{1\}, \dots, \{5\}\}$

$\Delta_{(h)} = \phi$

Full Data D

Let's assume for now that $w_5 < w_4 < w_5 + \mathrm{w}_4 < w_3 < \cdots$

Round 1: Drop #5 covariate

T	Covariates
0	2, 3, 5, 6, 9
1	2, 3, 5, 6, 8
1	2, 3, 5, 6, 7
0	2, 5, 5, 2, 8
1	2, 3, 5, 3, 9
1	2, 5, 5, 6, 10
0	2, 5, 5, 3, 7
…	…

D_1: unmatched units left

$s = \{5\}, \theta_{s,j} = 11110$

Update:

$\Lambda_{(1)} = \{\{1\}, \dots, \{4\}\}$

$\Delta_{(1)} = \{\{5\}\}$

Duke

Round 2: Drop #4 covariate

T	Covariates
0	2, 3, 5, 6, 9
1	2, 3, 5, 6, 8
1	2, 3, 5, 6, 7
0	2, 5, 5, 2, 8
1	2, 3, 5, 3, 9
1	2, 5, 5, 6, 10
0	2, 5, 5, 3, 7
…	…

D_2: unmatched units left

$s = \{4\}, \theta_{s,j} = 11101$

Update:

$\Lambda_{(2)} = \{\{1\}, \dots, \{3\}, \{4,5\}\}$

$\Delta_{(2)} = \{\{4\}, \{5\}\}$

Round 3: Drop #4 and #5 covariate

T	Covariates
0	2, 3, 5, 6, 9
1	2, 3, 5, 6, 8
1	2, 3, 5, 6, 7
0	2, 5, 5, 2, 8
1	2, 3, 5, 3, 9
1	2, 5, 5, 6, 10
0	2, 5, 5, 3, 7
…	…

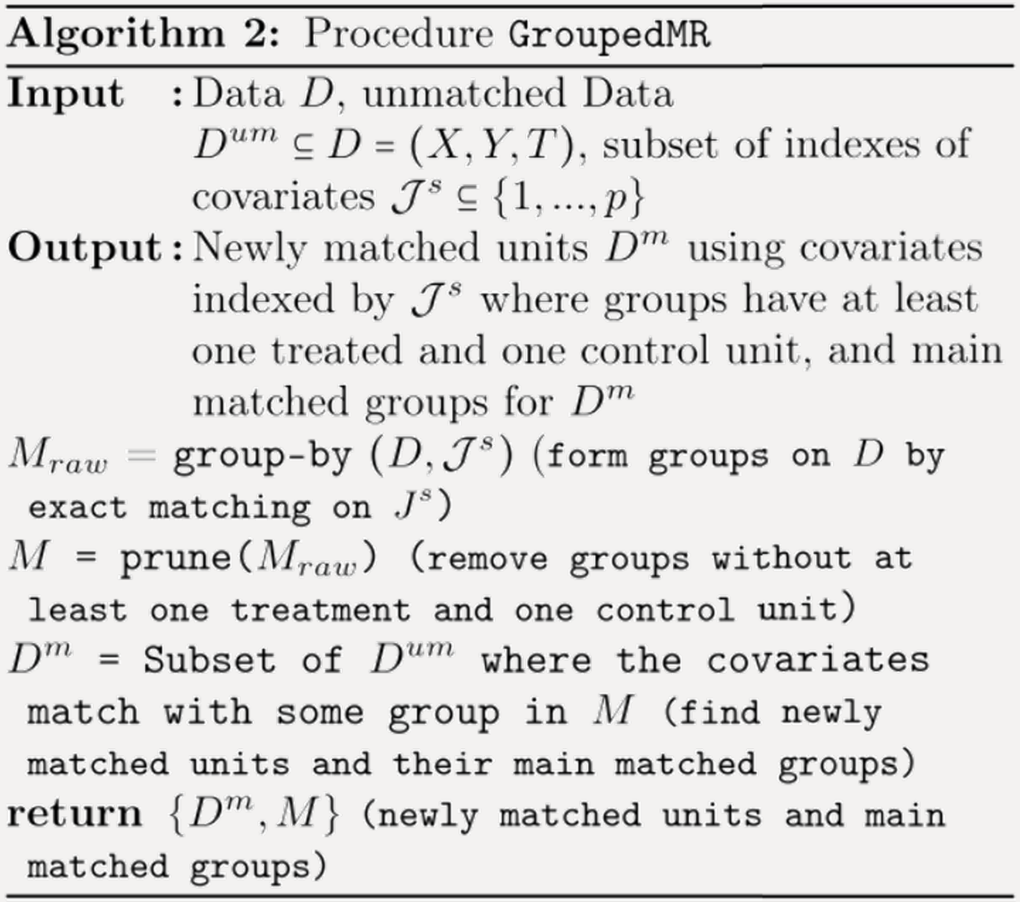
D_3: unmatched units left

$s = \{4, 5\}, \theta_{s,j} = 11100$

Update:
$\Lambda_{(3)} = \{\{1\}, \ldots, \{3\}\}$
$\Delta_{(3)} = \{\{4\}, \{5\}, \{4, 5\}\}$

Duke

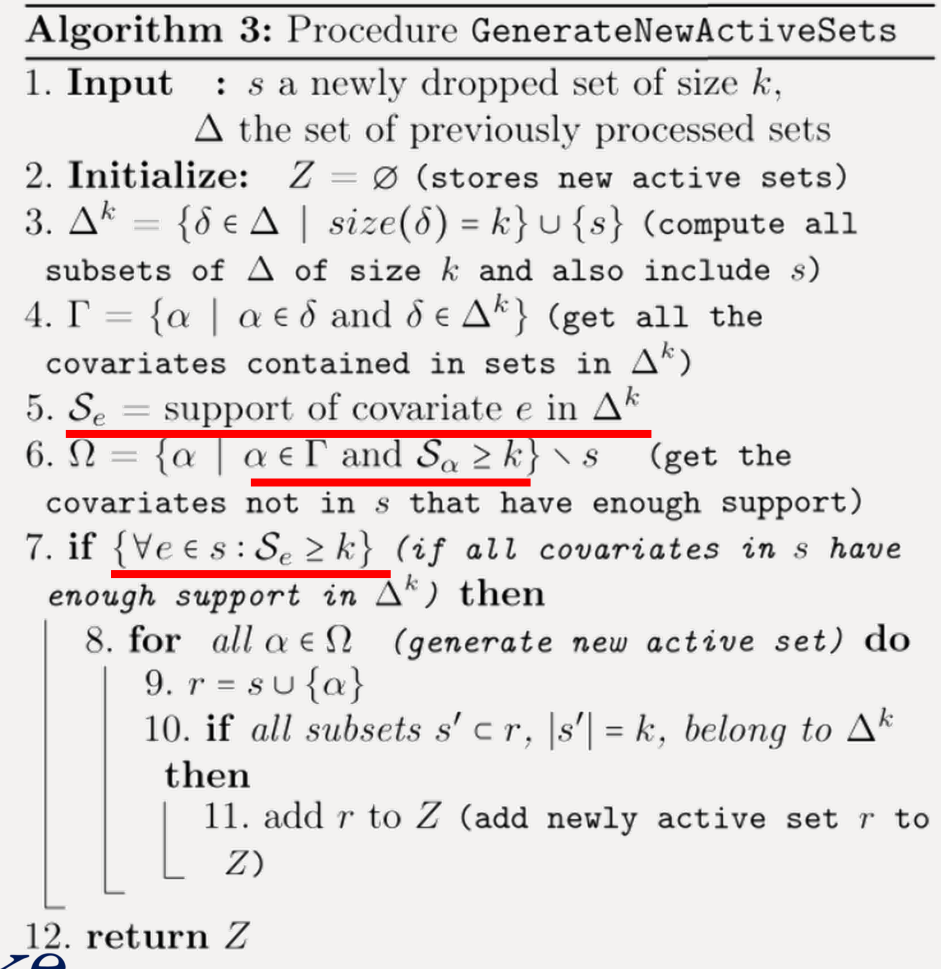
Grouping Procedure



```
Algorithm 2: Procedure GroupedMR
Input   : Data $D$, unmatched Data
          $D^{um} \subseteq D = (X, Y, T)$, subset of indexes of
          covariates $\mathcal{J}^s \subseteq \{1, ..., p\}$
Output  : Newly matched units $D^m$ using covariates
          indexed by $\mathcal{J}^s$ where groups have at least
          one treated and one control unit, and main
          matched groups for $D^m$
$M_{raw}$ = group-by $(D, \mathcal{J}^s)$ (form groups on $D$ by
 exact matching on $J^s$)
$M$ = prune($M_{raw}$) (remove groups without at
 least one treatment and one control unit)
$D^m$ = Subset of $D^{um}$ where the covariates
 match with some group in $M$ (find newly
 matched units and their main matched groups)
return $\{D^m, M\}$ (newly matched units and main
 matched groups)
```

- Matches all units in D to allow for matching with replacement!
- Meaning, if some unit is already matched in previous iterations, it would still appear in the matched groups of this iteration, but the unit itself won’t be in the subset that’s returned in this iteration

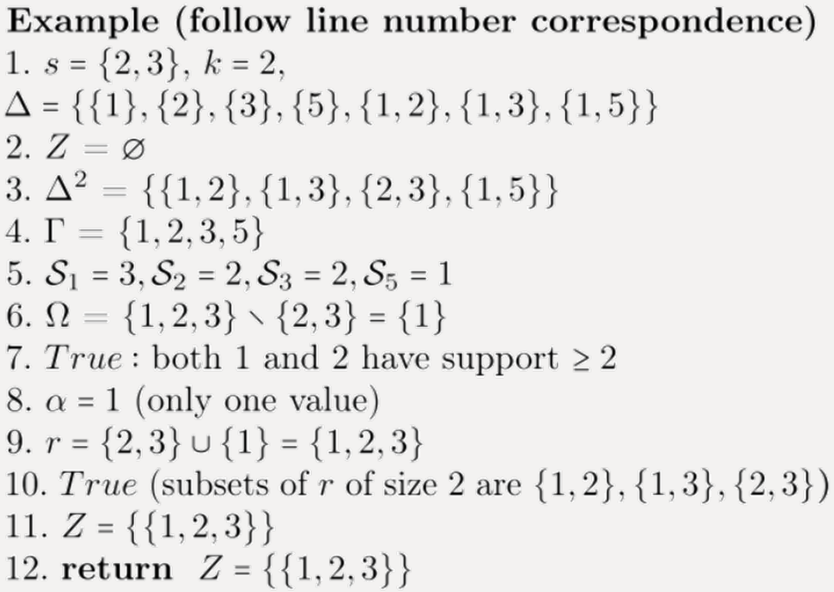
New Active Set Generation



Algorithm 3: Procedure `GenerateNewActiveSets`

```
1. Input  : $s$ a newly dropped set of size $k$,
            $\Delta$ the set of previously processed sets
2. Initialize:  $Z = \varnothing$ (stores new active sets)
3. $\Delta^k = \{\delta \in \Delta \mid size(\delta) = k\} \cup \{s\}$ (compute all
   subsets of $\Delta$ of size $k$ and also include $s$)
4. $\Gamma = \{\alpha \mid \alpha \in \delta$ and $\delta \in \Delta^k\}$ (get all the
   covariates contained in sets in $\Delta^k$)
5. $\underline{\mathcal{S}_e = \text{support of covariate } e \text{ in } \Delta^k}$
6. $\Omega = \{\alpha \mid \underline{\alpha \in \Gamma \text{ and } \mathcal{S}_\alpha \geq k}\} \setminus s$   (get the
   covariates not in $s$ that have enough support)
7. if $\underline{\{\forall e \in s : \mathcal{S}_e \geq k\}}$ (if all covariates in $s$ have
   enough support in $\Delta^k$) then
    8. for all $\alpha \in \Omega$ (generate new active set) do
        9. $r = s \cup \{\alpha\}$
        10. if all subsets $s' \subset r$, $|s'| = k$, belong to $\Delta^k$
            then
                11. add $r$ to $Z$ (add newly active set $r$ to
                    $Z$)
12. return $Z$
```

- Only generate a new active set r of size k+1 if all of r's subsets of size k have been processed.
- To prune candidates, do support check
  - Support of a covariate e defined as the number of sets in Δ^k containing e



Example (follow line number correspondence)

1. $s = \{2, 3\}$, $k = 2$,
$\Delta = \{\{1\}, \{2\}, \{3\}, \{5\}, \{1,2\}, \{1,3\}, \{1,5\}\}$
2. $Z = \varnothing$
3. $\Delta^2 = \{\{1,2\}, \{1,3\}, \{2,3\}, \{1,5\}\}$
4. $\Gamma = \{1, 2, 3, 5\}$
5. $\mathcal{S}_1 = 3, \mathcal{S}_2 = 2, \mathcal{S}_3 = 2, \mathcal{S}_5 = 1$
6. $\Omega = \{1, 2, 3\} \setminus \{2, 3\} = \{1\}$
7. $True$: both 1 and 2 have support ≥ 2
8. $\alpha = 1$ (only one value)
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. $Z = \{\{1, 2, 3\}\}$
12. **return** $Z = \{\{1, 2, 3\}\}$

Do this on whiteboard

Duke

If we don’t know the weights…

- Idea: find $\theta \in \{0,1\}^p$ that selects the covariates that could train a model to predict outcome with minimum prediction error $PE_{\mathcal{F}}(\theta) = \min_{f\in\mathcal{F}} \mathbb{E}(f(X \circ \theta, T) - Y)^2$
- Denote two dataset
  - $S^{matching} = \{X^M, Y^M, T^M\}$ the matching set
  - $S^{training} = \{X^T, Y^T, T^T\}$ the training set
- Denote the matched group for unit i on covariates θ from matching set $S^{matching}$:
  $\mathcal{MG}_{\mathrm{i}}\left(\theta, S^{matching}\right) = \{\mathrm{i}' \in S^{matching}, s.t. x_{i'} \circ \theta = x_{i'} \circ \theta\}$

Full Almost Exact Matching (Full-AME)

- For treatment unit i, match on a set of variables that minimizes empirical PE:

$$\theta^*_{i,S^{matching}} \in \arg\min_{\theta} \widehat{PE}_{\mathcal{F}}\left(\theta, S^{training}\right)\ s.t.\ (\exists \ell \in \mathcal{MG}_i\left(\theta, S^{matching}\right)\ s.t.\ T_\ell = 0)$$

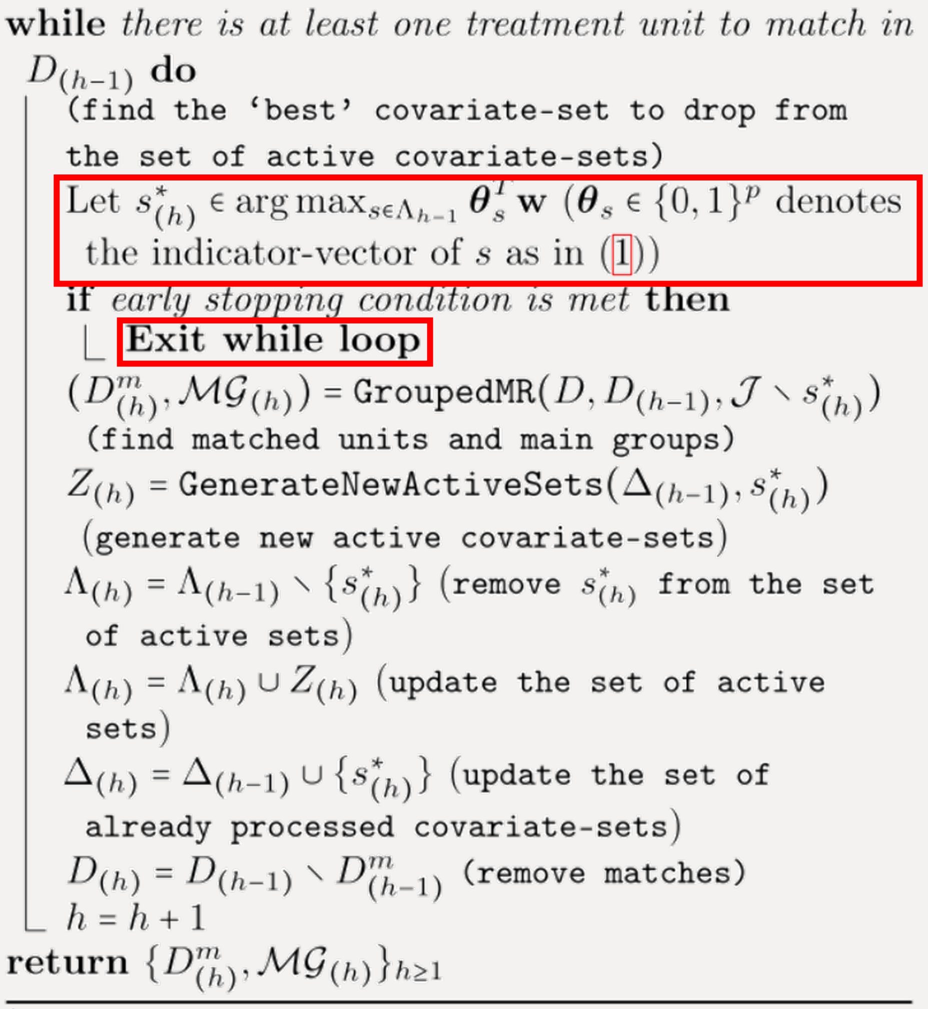
Find set of covariates

For each treatment point i

That minimizes the prediction error on the training set

Such that there’s at least one control unit in the matched group of i in the matching set

DAME with Adaptive Weights



```
while there is at least one treatment unit to match in $D_{(h-1)}$ do
    (find the 'best' covariate-set to drop from the set of active covariate-sets)
    Let $s^*_{(h)} \in \arg\max_{s \in \Lambda_{h-1}} \boldsymbol{\theta}_s^T \mathbf{w}$ ($\boldsymbol{\theta}_s \in \{0,1\}^p$ denotes the indicator-vector of $s$ as in (1))
    if early stopping condition is met then
        Exit while loop
    $(D^m_{(h)}, \mathcal{MG}_{(h)}) = \texttt{GroupedMR}(D, D_{(h-1)}, \mathcal{J} \setminus s^*_{(h)})$ (find matched units and main groups)
    $Z_{(h)} = \texttt{GenerateNewActiveSets}(\Delta_{(h-1)}, s^*_{(h)})$ (generate new active covariate-sets)
    $\Lambda_{(h)} = \Lambda_{(h-1)} \setminus \{s^*_{(h)}\}$ (remove $s^*_{(h)}$ from the set of active sets)
    $\Lambda_{(h)} = \Lambda_{(h)} \cup Z_{(h)}$ (update the set of active sets)
    $\Delta_{(h)} = \Delta_{(h-1)} \cup \{s^*_{(h)}\}$ (update the set of already processed covariate-sets)
    $D_{(h)} = D_{(h-1)} \setminus D^m_{(h-1)}$ (remove matches)
    $h = h + 1$
return $\{D^m_{(h)}, \mathcal{MG}_{(h)}\}_{h \geq 1}$
```

$$s^*_{(h)} \in \arg \min_{s \in \Lambda_{(h-1)}} \widehat{\mathrm{PE}}(\theta_s)$$

**Typically, train two separate models for treatment/control*

- During iterations where we drop covariates, the prediction error $\widehat{PE}_{\mathcal{F}}$ should never increase too far above the original value using all covariates
    - i.e. $\widehat{PE}_{\mathcal{F}}(\theta^*_u) < \min_{\theta} \widehat{PE}_{\mathcal{F}}(\theta) + \epsilon$

Duke

Advantages and Disadvantages of DAME

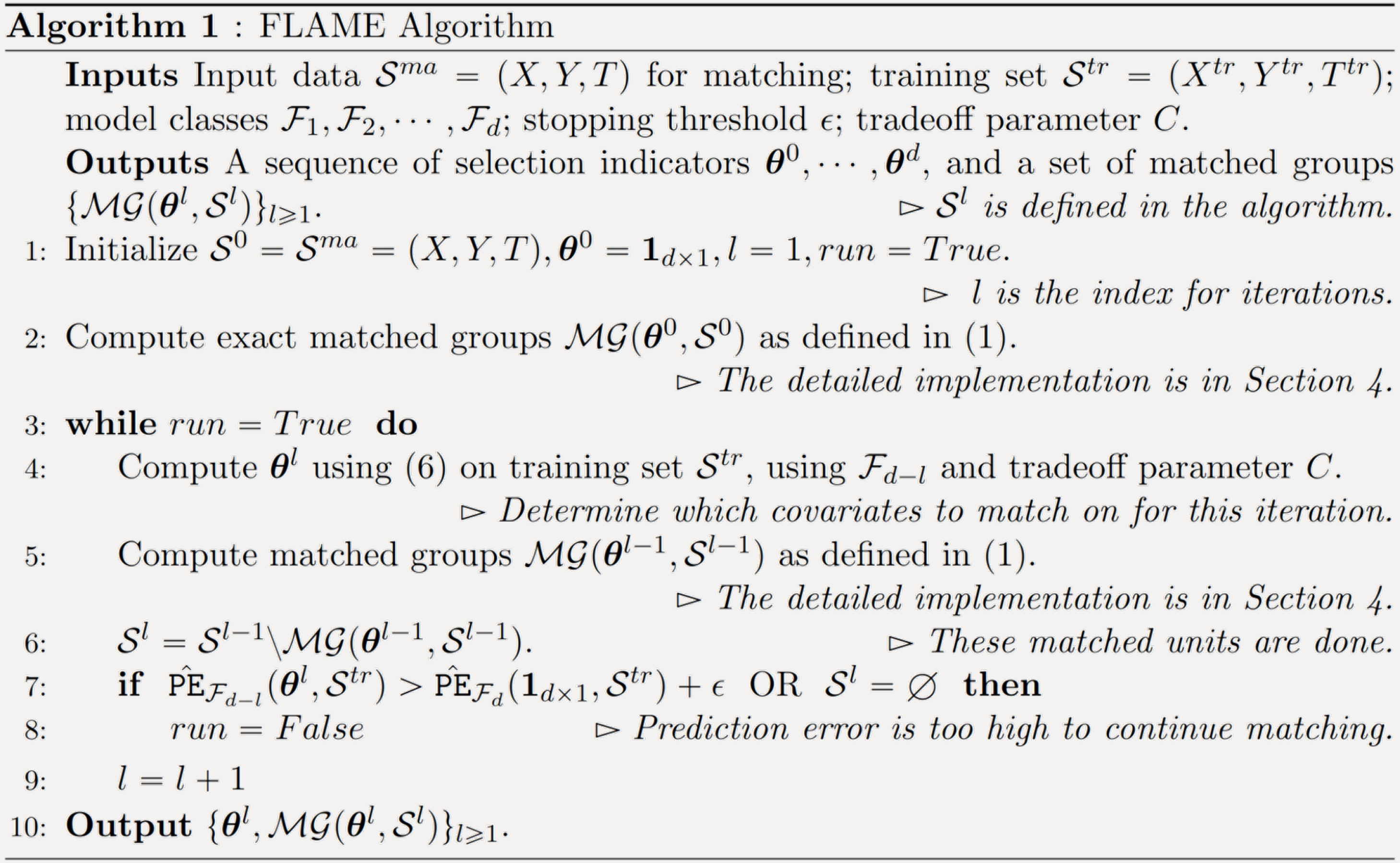
- Advantages:
  - DAME can be used to estimate CATE;
  - DAME produces interpretable matches that are guarantee to be high quality since it goes over all the possible feature combinations and matches the controlled and treatment group.
- Disadvantages:
  - Time Consuming: all the possible subsets are calculated
  - Not high-dimensional friendly

Fast Large-Scale Almost Matching Exactly Algorithm (FLAME)

An approach to match under the potential outcome framework with binary treatments and a possibly large number of discrete covariates

Duke

FLAME Algorithm



Algorithm 1 : FLAME Algorithm

```
**Inputs** Input data $\mathcal{S}^{ma} = (X, Y, T)$ for matching; training set $\mathcal{S}^{tr} = (X^{tr}, Y^{tr}, T^{tr})$; model classes $\mathcal{F}_1, \mathcal{F}_2, \cdots, \mathcal{F}_d$; stopping threshold $\epsilon$; tradeoff parameter $C$.
**Outputs** A sequence of selection indicators $\boldsymbol{\theta}^0, \cdots, \boldsymbol{\theta}^d$, and a set of matched groups $\{\mathcal{MG}(\boldsymbol{\theta}^l, \mathcal{S}^l)\}_{l \geqslant 1}$.    ▷ $\mathcal{S}^l$ is defined in the algorithm.
Initialize $\mathcal{S}^0 = \mathcal{S}^{ma} = (X, Y, T), \boldsymbol{\theta}^0 = \mathbf{1}_{d \times 1}, l = 1, run = True$.
                                                            ▷ $l$ is the index for iterations.
Compute exact matched groups $\mathcal{MG}(\boldsymbol{\theta}^0, \mathcal{S}^0)$ as defined in (1).
                                              ▷ The detailed implementation is in Section 4.
**while** $run = True$ **do**
    Compute $\boldsymbol{\theta}^l$ using (6) on training set $\mathcal{S}^{tr}$, using $\mathcal{F}_{d-l}$ and tradeoff parameter $C$.
                                  ▷ Determine which covariates to match on for this iteration.
    Compute matched groups $\mathcal{MG}(\boldsymbol{\theta}^{l-1}, \mathcal{S}^{l-1})$ as defined in (1).
                                              ▷ The detailed implementation is in Section 4.
    $\mathcal{S}^l = \mathcal{S}^{l-1} \backslash \mathcal{MG}(\boldsymbol{\theta}^{l-1}, \mathcal{S}^{l-1})$.    ▷ These matched units are done.
    **if** $\hat{\text{PE}}_{\mathcal{F}_{d-l}}(\boldsymbol{\theta}^l, \mathcal{S}^{tr}) > \hat{\text{PE}}_{\mathcal{F}_d}(\mathbf{1}_{d \times 1}, \mathcal{S}^{tr}) + \epsilon$ OR $\mathcal{S}^l = \emptyset$ **then**
        $run = False$    ▷ Prediction error is too high to continue matching.
    $l = l + 1$
**Output** $\{\boldsymbol{\theta}^l, \mathcal{MG}(\boldsymbol{\theta}^l, \mathcal{S}^l)\}_{l \geqslant 1}$.
```

Step 1: Find the exact matching group for all features

Step 2: Match the individuals for remained feature with exact matching

Step 3: Find the feature eliminated in this round

Step 4: Stop when the Prediction Error is too high

Duke

Example for FLAME

- Suppose the relevance of the features for x_i is $x_1 > x_2 > x_3 > x_4$

FLAME Algorithm (Backward Elimination)					DAME Algorithm (apriori)			
Step 1	x1	x2	x3	x4	x1	x2	x3	x4
Step 2	x1	x2	x3		x1	x2	x3	
Step 3	x1	x2			x1	x2		x4
Step 4	x1				x1	x2		
Step 5					x1		x3	x4
Step 6					x1		x3	
Step 7					x1			x4
Step n					…			

Balancing Factor (BF) Criteria

$$BF(MG(\theta, S^{ma})) = \frac{\#\text{ control in } MG(\theta, S^{ma})}{\#\text{ available control}} + \frac{\#\text{ treated in } MG(\theta, S^{ma})}{\#\text{ available treated}}$$

The objection function for FLAME is

$$\theta^i \in \arg\max_{\theta}[-PE_{F_{\|\theta\|_0}}(\theta, S^{tr}) + C \cdot BF(MG(\theta, S^{ma}))]$$

- Encourages a large fraction of both treatment and control units to be used for the matched groups and more units would be matched in earlier iterations.

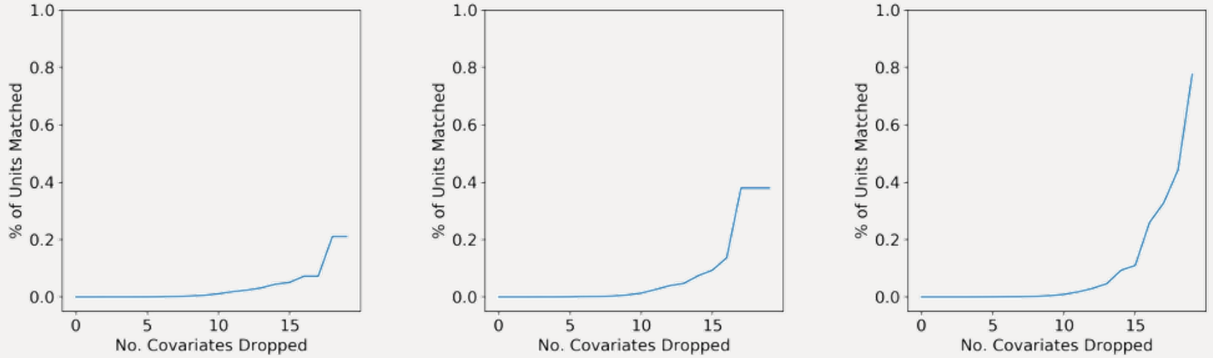
Balancing Factor Effect

$$\theta^i \in \arg\max_{\theta}[-PE_{F_{\|\theta\|_0}}(\theta, S^{tr}) + C \cdot BF(MG(\theta, S^{ma}))]$$

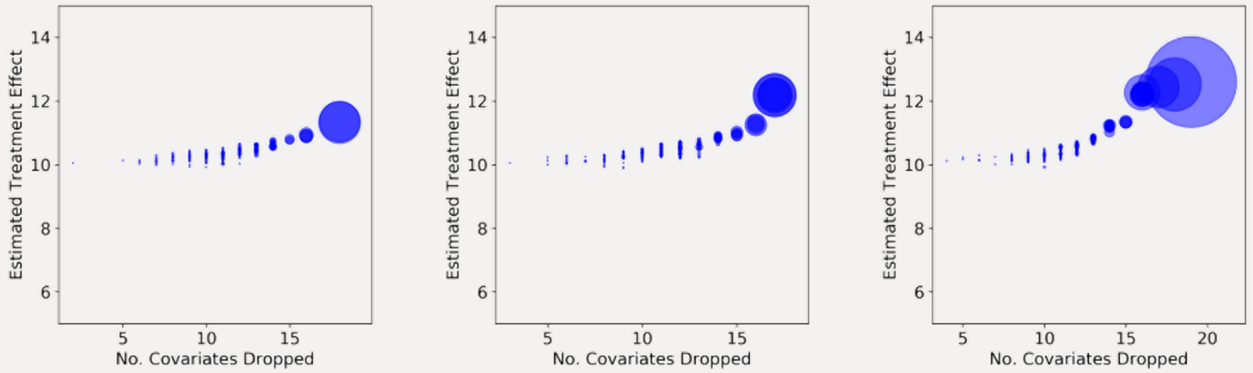
$$BF(MG(\theta, S^{ma})) = \frac{\#\text{ control in } MG(\theta, S^{ma})}{\#\text{ available control}} + \frac{\#\text{ treated in } MG(\theta, S^{ma})}{\#\text{ available treated}}$$

- Maximizing BF: encourage more units to be matched in earlier iterations (even if they have rather large PE)

- Simulation: $y = \sum_{i=1}^{20} \frac{1}{i} x_i + 10T + \epsilon$
  - $x_i \sim \text{Bernoulli}(0.1 + \frac{3(i-1)}{190})$ for control and $x_i \sim \text{Bernoulli}(0.9 - \frac{3(i-1)}{190})$ for treatment, $\epsilon \sim \mathcal{N}(0, 0.1)$



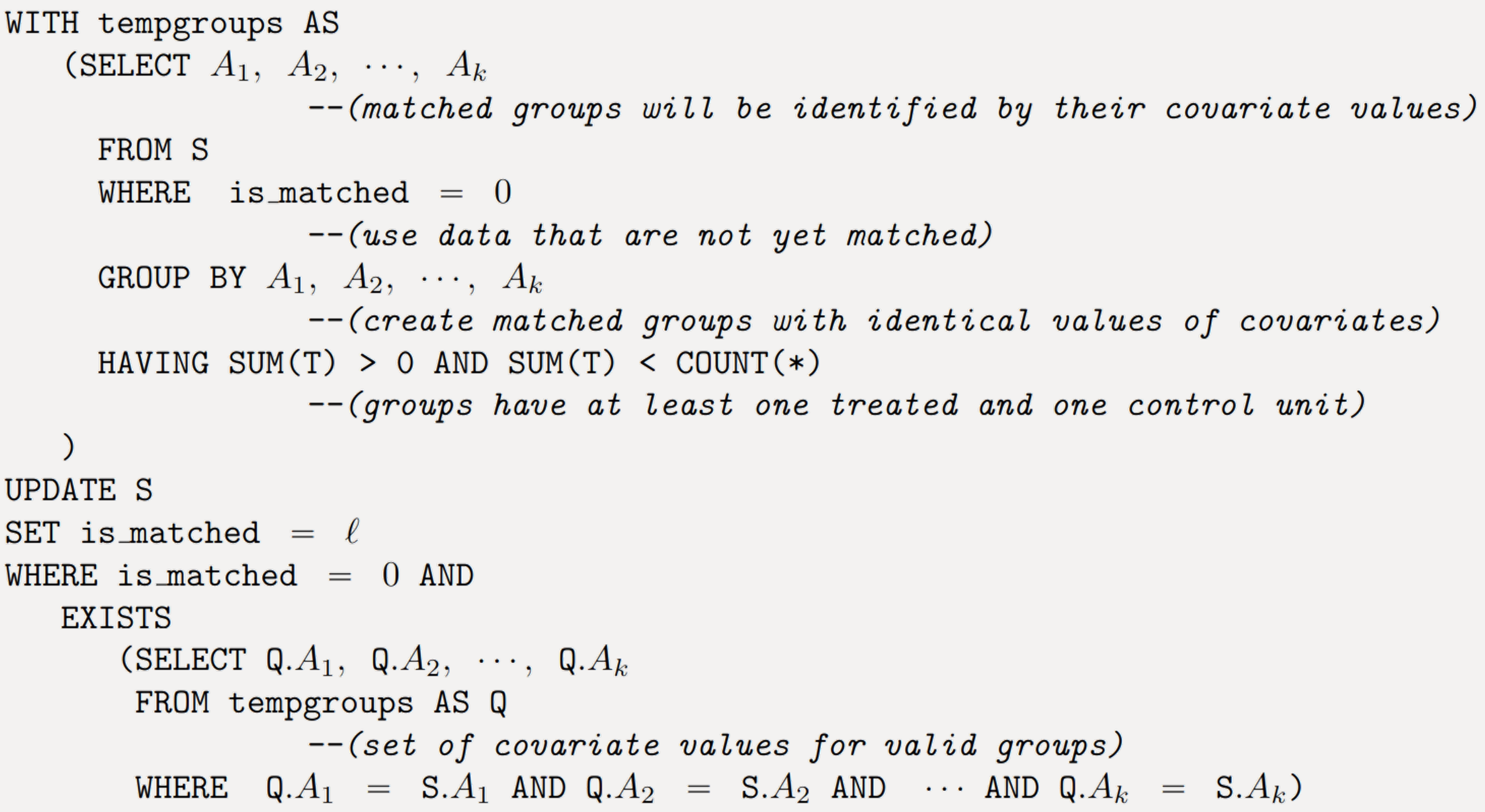
(a) $C = 0.1$ (b) $C = 0.5$ (c) $C = 1$



(a) $C = 0.1$ (b) $C = 0.5$ (c) $C = 1$

Duke

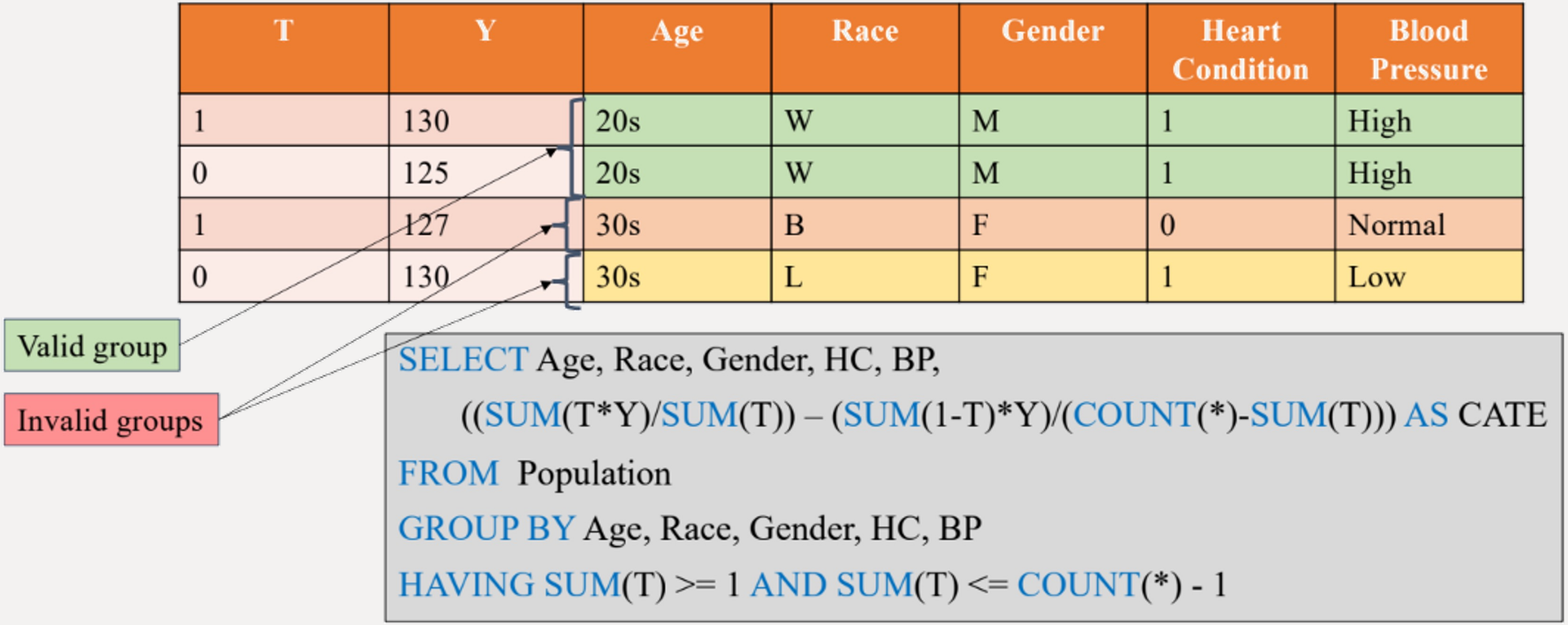
Implementation using Database (SQL) Queries



```
WITH tempgroups AS
   (SELECT $A_1$, $A_2$, $\cdots$, $A_k$
                --(matched groups will be identified by their covariate values)
     FROM S
     WHERE  is_matched $=$ $0$
                --(use data that are not yet matched)
     GROUP BY $A_1$, $A_2$, $\cdots$, $A_k$
                --(create matched groups with identical values of covariates)
     HAVING SUM(T) > 0 AND SUM(T) < COUNT(*)
                --(groups have at least one treated and one control unit)
   )
UPDATE S
SET is_matched $=$ $\ell$
WHERE is_matched $=$ $0$ AND
   EXISTS
      (SELECT Q.$A_1$, Q.$A_2$, $\cdots$, Q.$A_k$
       FROM tempgroups AS Q
                --(set of covariate values for valid groups)
       WHERE  Q.$A_1$ $=$ S.$A_1$ AND Q.$A_2$ $=$ S.$A_2$ AND  $\cdots$ AND Q.$A_k$ $=$ S.$A_k$)
```

Duke

Implementation using Database (SQL) Queries

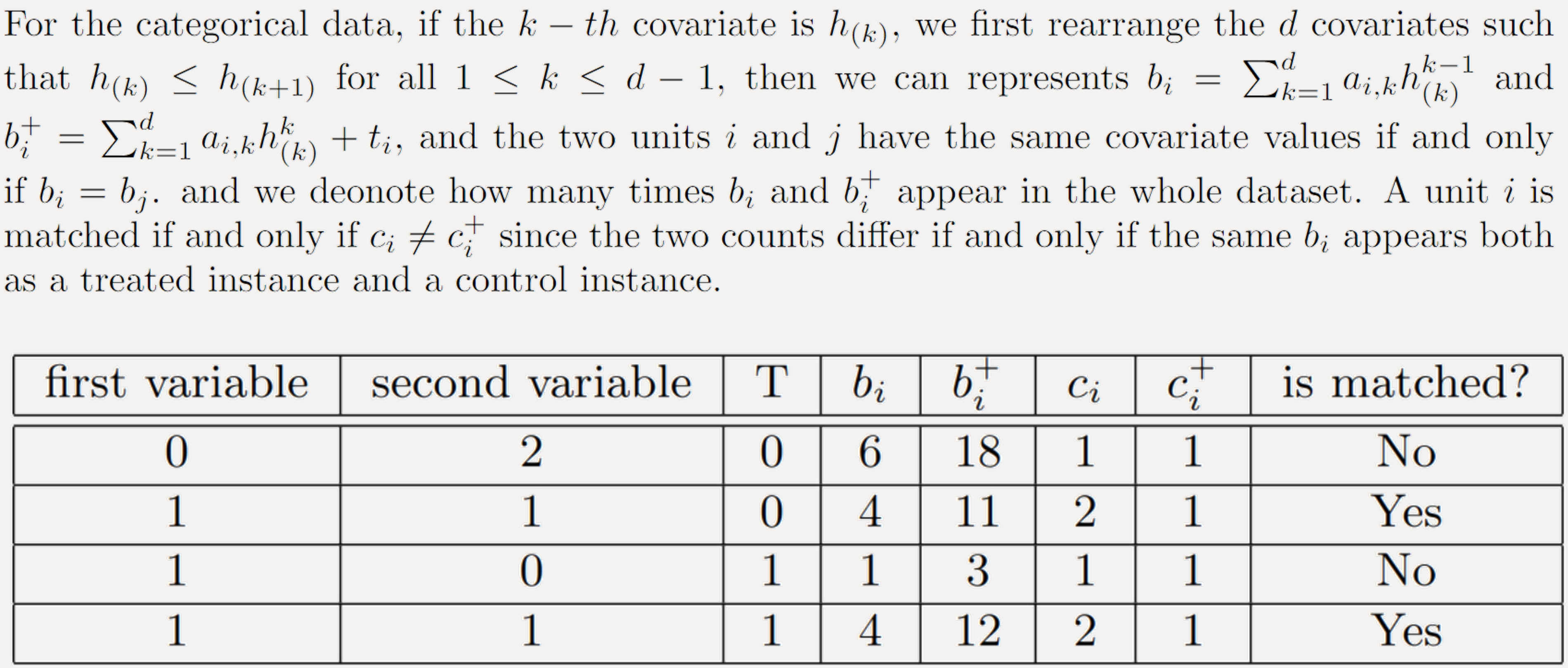


T	Y	Age	Race	Gender	Heart Condition	Blood Pressure
1	130	20s	W	M	1	High
0	125	20s	W	M	1	High
1	127	30s	B	F	0	Normal
0	130	30s	L	F	1	Low

```
SELECT Age, Race, Gender, HC, BP,
    ((SUM(T*Y)/SUM(T)) – (SUM(1-T)*Y)/(COUNT(*)-SUM(T))) AS CATE
FROM  Population
GROUP BY Age, Race, Gender, HC, BP
HAVING SUM(T) >= 1 AND SUM(T) <= COUNT(*) - 1
```

Duke

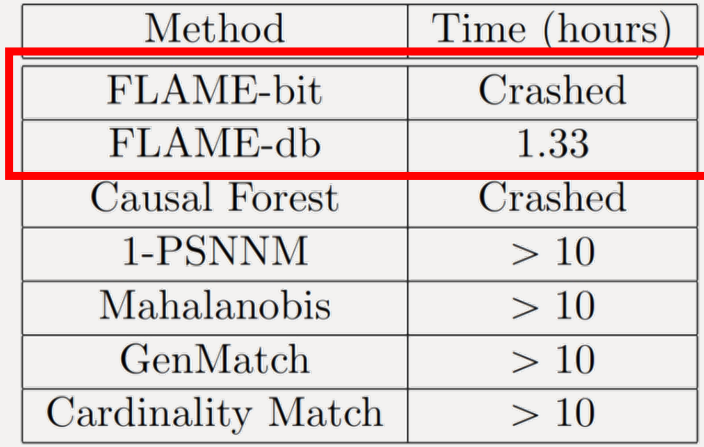
Implementation using Bit Vectors



For the categorical data, if the $k-th$ covariate is $h_{(k)}$, we first rearrange the d covariates such that $h_{(k)} \leq h_{(k+1)}$ for all $1 \leq k \leq d-1$, then we can represents $b_i = \sum_{k=1}^{d} a_{i,k} h_{(k)}^{k-1}$ and $b_i^+ = \sum_{k=1}^{d} a_{i,k} h_{(k)}^{k} + t_i$, and the two units i and j have the same covariate values if and only if $b_i = b_j$. and we deonote how many times b_i and b_i^+ appear in the whole dataset. A unit i is matched if and only if $c_i \neq c_i^+$ since the two counts differ if and only if the same b_i appears both as a treated instance and a control instance.

first variable	second variable	T	b_i	b_i^+	c_i	c_i^+	is 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

A time comparison between matching methods

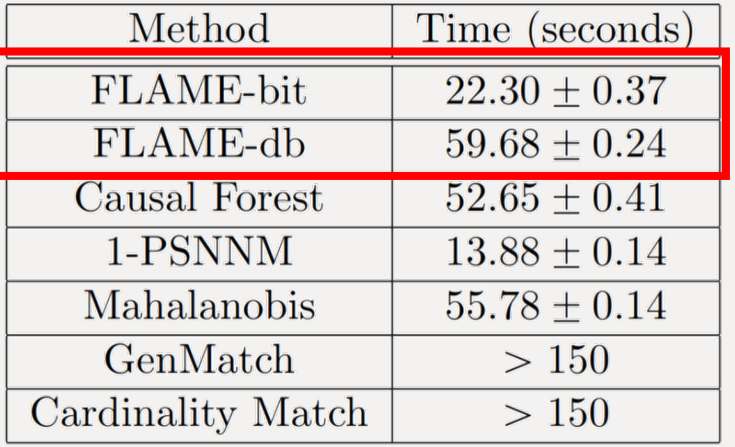


Method	Time (hours)
FLAME-bit	Crashed
FLAME-db	1.33
Causal Forest	Crashed
1-PSNNM	> 10
Mahalanobis	> 10
GenMatch	> 10
Cardinality Match	> 10

US Census 1990 dataset

Units = 1.2 million

Covariates = 59



Method	Time (seconds)
FLAME-bit	22.30 ± 0.37
FLAME-db	59.68 ± 0.24
Causal Forest	52.65 ± 0.41
1-PSNNM	13.88 ± 0.14
Mahalanobis	55.78 ± 0.14
GenMatch	> 150
Cardinality Match	> 150

Synthetic dataset

Units = 20k

Covariates = 30

(Averaged over three runs)

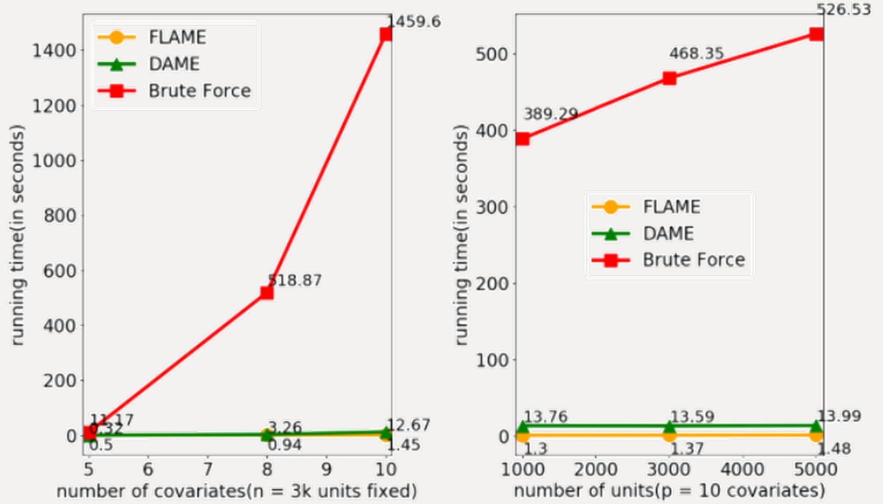


Figure 3: Run-time comparison between DAME FLAME, and brute force. *Left:* varying number of units. *Right:* varying number of covariates.

Brute Force AME solver: Brute force pairwise comparison of treatment points to control points.

Quadratic in number of units n, linear in number of covariates p

Duke

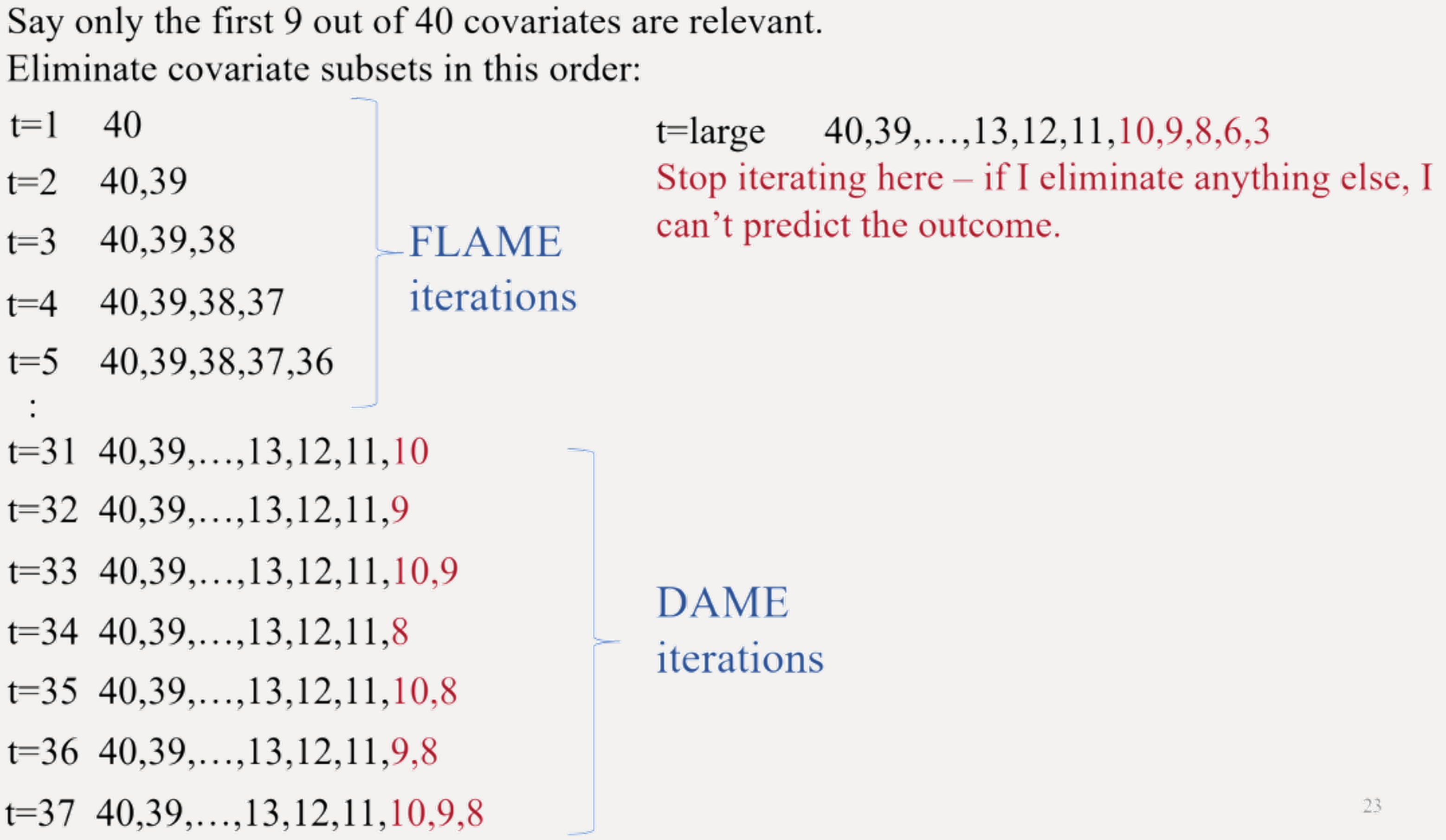
The Pros and Cons for two Implementations

- FLAME-Bit
  - Use bit-vectors to find valid groups
  - Uses main-memory
  - Faster in smaller dataset
  - Cannot handle large datasets(millions of units)
- FLAME-DB
  - Use SQL queries from database
  - Stores bulk of data on disk
  - Less efficient than Bit vector for small datasets
  - Can handle bigger data (millions of units)

Advantages and Disadvantages of FLAME

- Advantages:
  - FLAME is a greedy algorithm to find the matched group, so it is faster than DAME.
  - FLAME consider bias introduced by irrelevance covariates and introduces balance factor (BF) to match more units.
- Disadvantages:
  - The units of groups might not be perfectly matched since it is a greedy algorithm.

The combination of FLAME and DAME



Say only the first 9 out of 40 covariates are relevant.
Eliminate covariate subsets in this order:

t=1 40
t=2 40,39
t=3 40,39,38
t=4 40,39,38,37
t=5 40,39,38,37,36
:

FLAME iterations

t=31 40,39,…,13,12,11,10
t=32 40,39,…,13,12,11,9
t=33 40,39,…,13,12,11,10,9
t=34 40,39,…,13,12,11,8
t=35 40,39,…,13,12,11,10,8
t=36 40,39,…,13,12,11,9,8
t=37 40,39,…,13,12,11,10,9,8

DAME iterations

t=large 40,39,…,13,12,11,10,9,8,6,3
Stop iterating here – if I eliminate anything else, I can't predict the outcome.

23

Duke

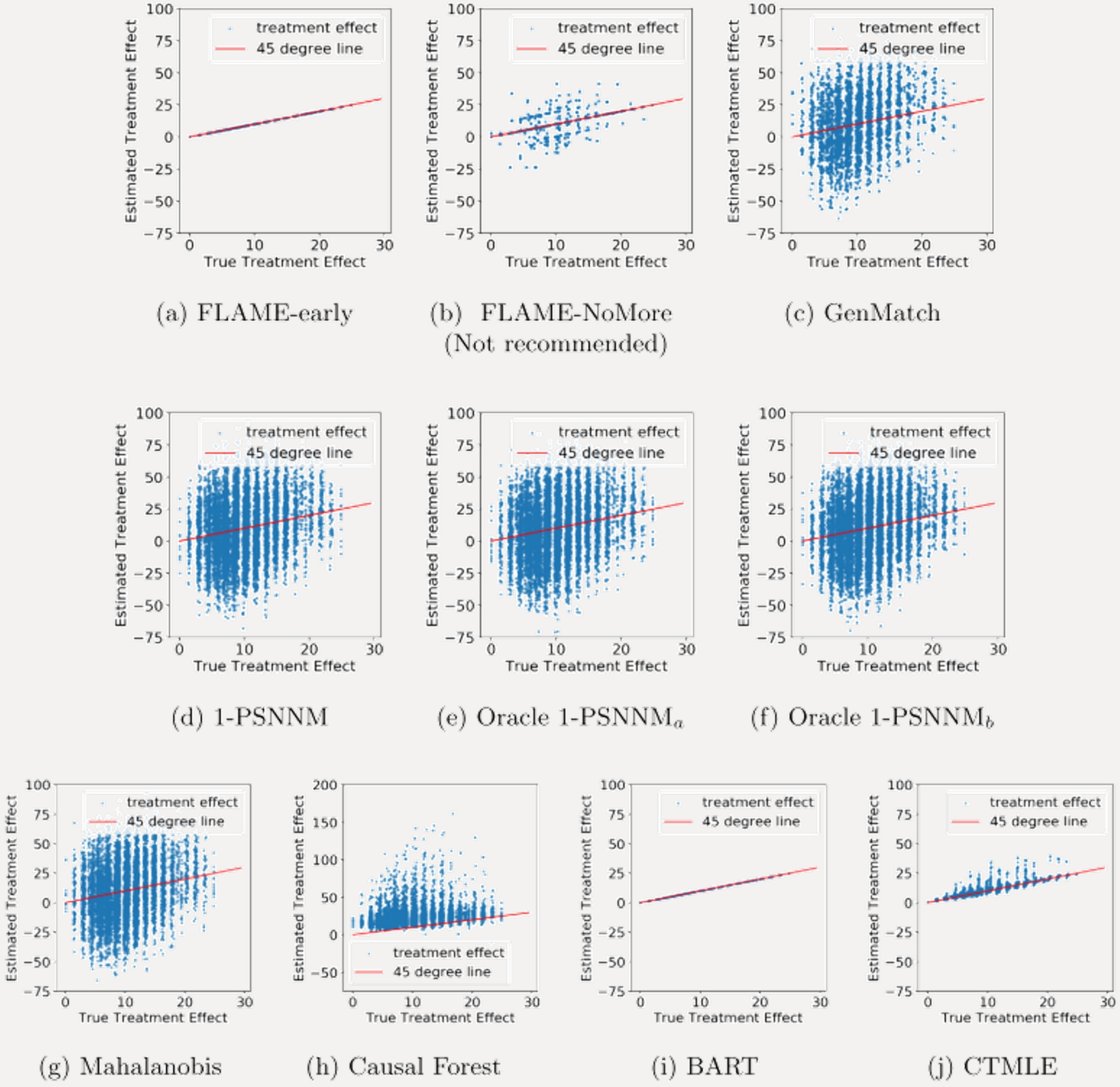
Simulation and Performance Comparison

Duke

FLAME Simulation

- $y = \underbrace{\boxed{\sum_{i=1}^{10} \alpha_i x_i}}_{\text{Baseline linear effect}} + T \underbrace{\boxed{\sum_{i=1}^{10} \beta_i x_i}}_{\text{linear treatment effect}} + T \cdot U \underbrace{\boxed{\sum_{1 \leq i < j \leq 5} x_i x_j}}_{\text{quadratic (nonlinear) treatment effect}} + \epsilon$
- Relevant covariates $1 \leq i \leq 10$: $\alpha_i \sim N(10s, 1)$, with $s \sim \text{Uniform}\{-1, 1\}$; $\beta_i \sim N(1.5, 0.15)$; $\epsilon \sim \mathcal{N}(0, 0.1)$; $x_i \sim \text{Bernoulli}(0.5)$
- Irrelevant covariates $10 < i \leq 30$: $\alpha_i = \beta_i = 0$, $x_i \sim \text{Bernoulli}(0.1)$ in the control group, $x_i \sim \text{Bernoulli}(0.9)$ in the treatment group.
- 10000 each for treatment/control units.
- Matching set generated identically with the training set

Duke



(a) FLAME-early (b) FLAME-NoMore (Not recommended) (c) GenMatch

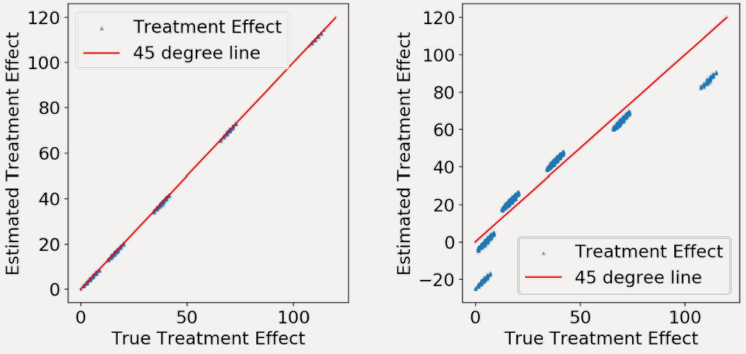
(d) 1-PSNNM (e) Oracle 1-PSNNM$_a$ (f) Oracle 1-PSNNM$_b$

(g) Mahalanobis (h) Causal Forest (i) BART (j) CTMLE

We should stop FLAME before eliminating important variables when the PE drops to an unacceptable value

propensity score matching (of any kind) projects the data to one dimension, and thus cannot be used for CATE estimation.

Regression and other modeling methods are subject to misspecification.



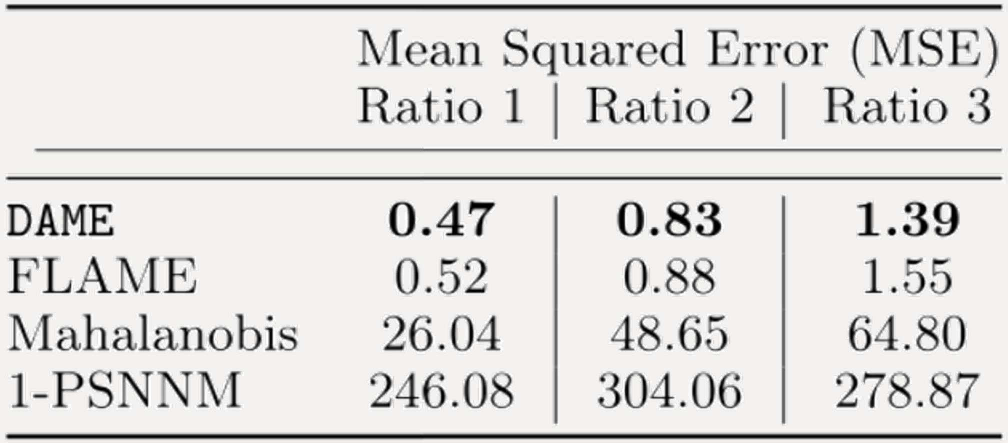
(a) FLAME (b) Double linear regressors

Figure 4: Scatter plots of true treatment effect versus estimated treatment effect on every unit in the synthetic data. The regression model is misspecified, and performs poorly.

DAME Simulation

- - $y = \underbrace{\boxed{\sum_i \alpha_i x_i}}_{\text{Baseline linear effect}} + \underbrace{\boxed{T \sum_{i=1} \beta_i x_i}}_{\text{linear treatment effect}} + \underbrace{\boxed{\mathrm{T} \cdot U \sum_{i,\gamma,\gamma>i} x_i x_\gamma}}_{\text{quadratic (nonlinear) treatment effect}}$

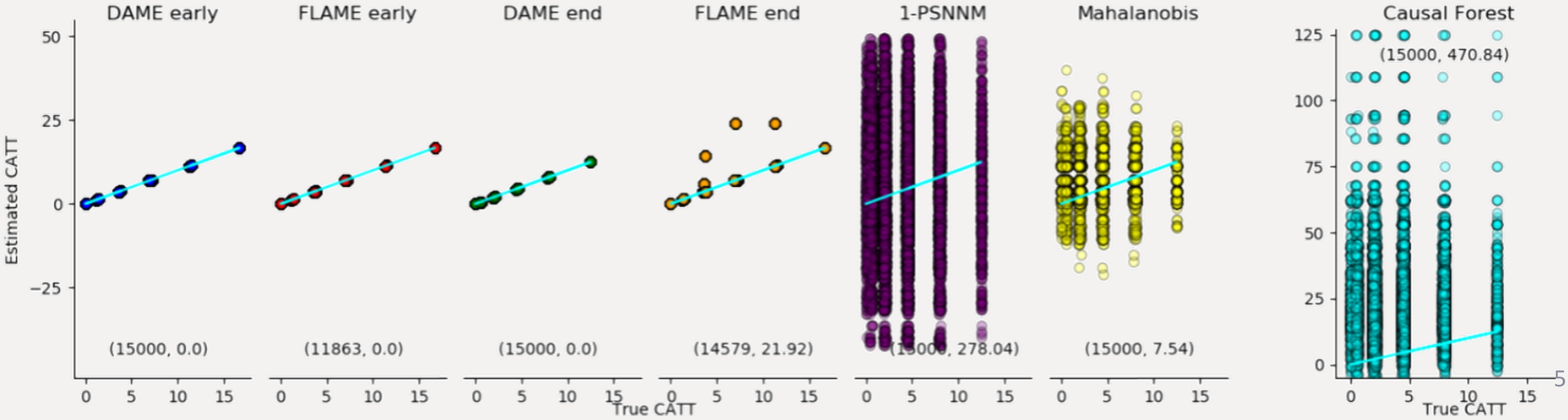
- - Experiments on imbalanced data
  - 2000 treatment, 40000/20000/10000 control units
  - DAME: 4 covariates not matched on average, 84% matched on all but 2 covariates;
  - FLAME: 7 covariates not matched on average, 25% matched on all but 2 covariates



	Mean Squared Error (MSE)		
	Ratio 1	Ratio 2	Ratio 3
DAME	**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

Irrelevant Covariates

- $y = \sum_i \alpha_i x_i + T \sum_{i=1} \beta_i x_i + \mathrm{T} \cdot U \sum_{i,\gamma,\gamma>i} x_i x_\gamma$
  - Important covariates $1 \leq i \leq 5$: $\alpha_i \sim N(10s, 1)$, with $s \sim \mathrm{Uniform}\{-1, 1\}$; $\beta_i \sim N(1.5, 0.15)$; $x_i \sim \mathrm{Bernoulli}(0.5)$
  - Unimportant covariates $5 < i \leq 15$: $x_i \sim \mathrm{Bernoulli}(0.1)$ in the control group, $x_i \sim \mathrm{Bernoulli}(0.9)$ in the treatment group.
  - 15000 control/treatment units.



Exponentially Decaying Covariates

$$y = \sum_{i} \alpha_i x_i + T \sum_{i=1} \beta_i x_i + \mathrm{T} \cdot U \sum_{i,\gamma,\gamma>i} x_i x_\gamma$$

- Let α decrease exponentially as

$$\alpha_i = 64 \times \left(\frac{1}{2}\right)^{\mathrm{i}}$$

- DAME produces more high-quality matches before resorting to lower quality matches

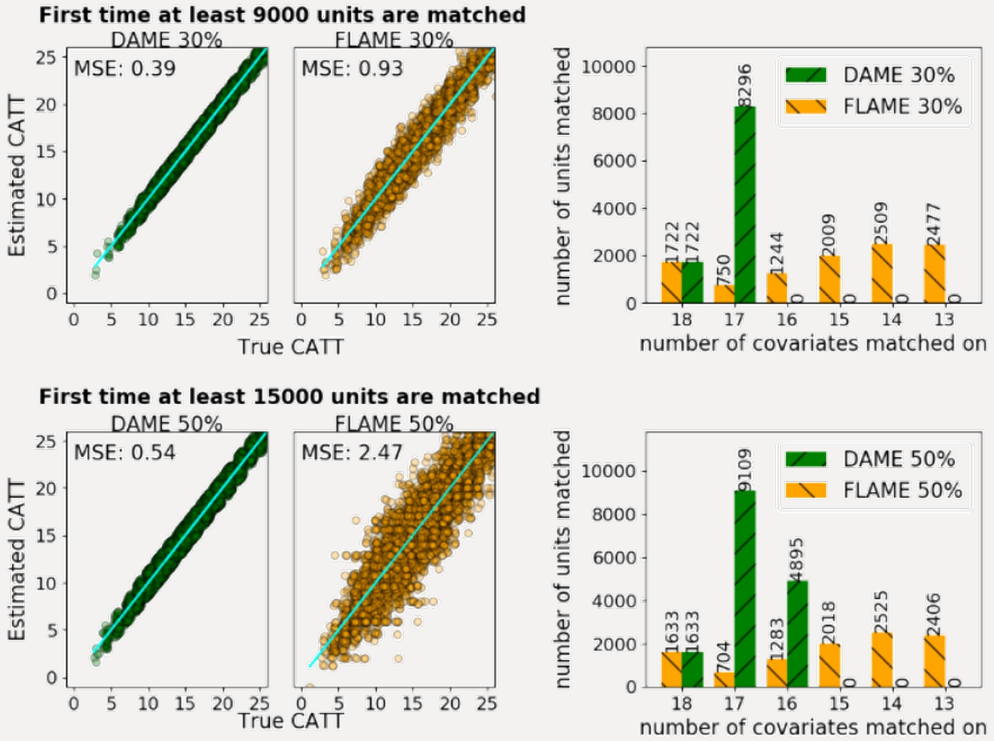
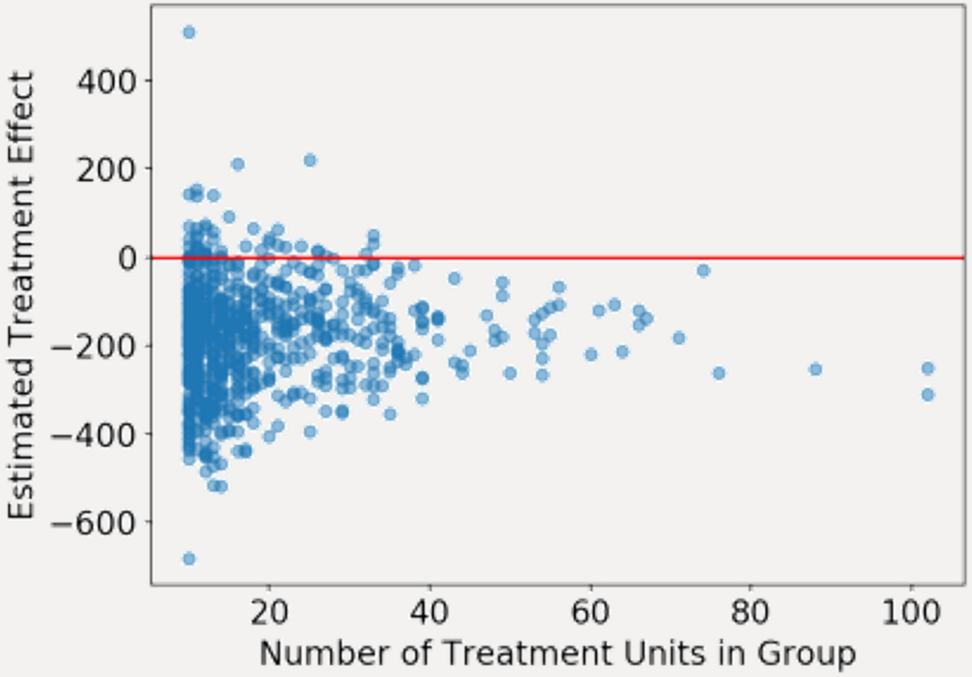


Figure 2: DAME makes higher quality matches early on. Rows correspond to stopping thresholds (top row 30%, bottom row 50%). DAME matches on more covariates than FLAME, yielding lower MSE from matched groups.

FLAME on Real Data: Extreme Smoking on birth weight

- Treatment: smoking at least 10 cigarettes per day for the duration of the pregnancy
- Control Group: women who did not smoke at all during pregnancy
- ATE on birth weight:
  - -248 grams of infant’s weight



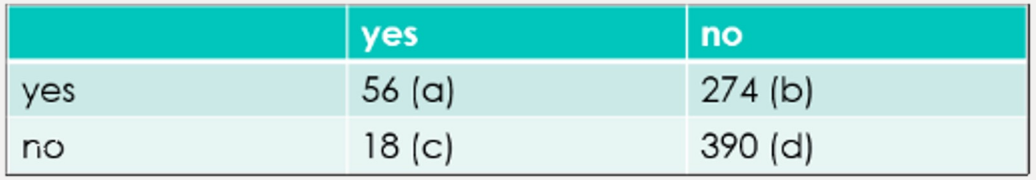
~2.1M units in total
~75K units are treated units

Duke

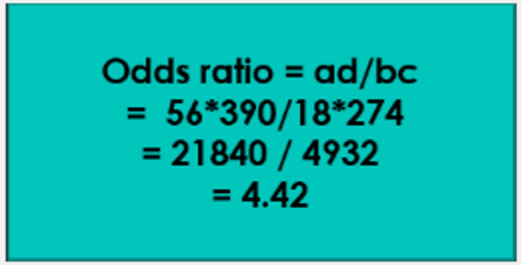
How to determine if an outcome is not influenced by extreme smoking?

Real Data: Extreme Smoking on NICU admission

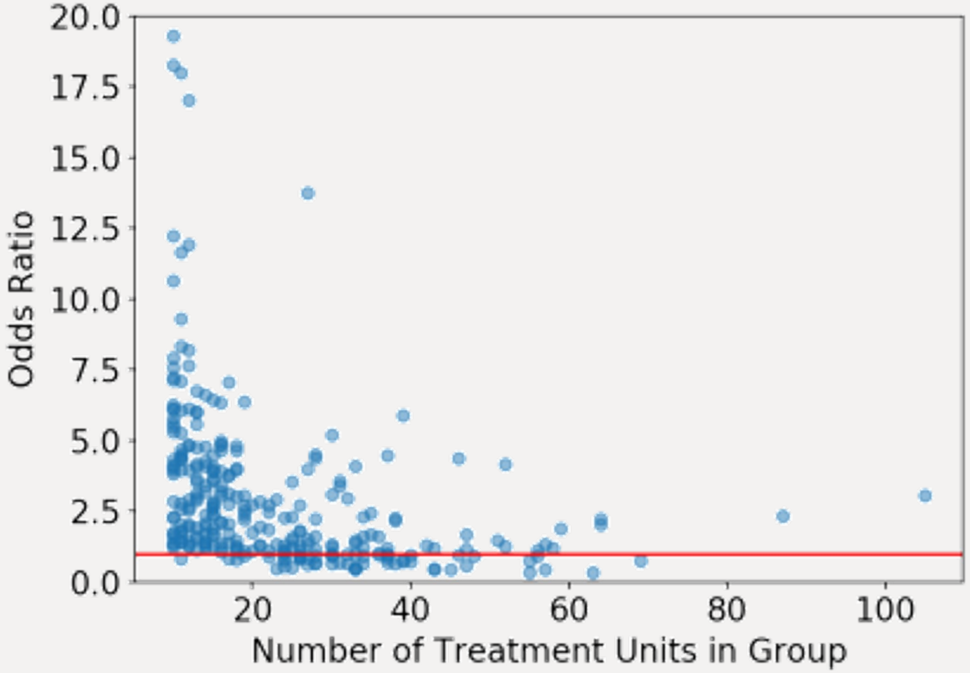
- **Odd ratio** to test the data quality
- **Conclusion:** the available data are not sufficient for granular analysis, or any strong conclusion on NICU admissions.



	yes	no
yes	56 (a)	274 (b)
no	18 (c)	390 (d)



Odds ratio = ad/bc
= 56*390/18*274
= 21840 / 4932
= 4.42

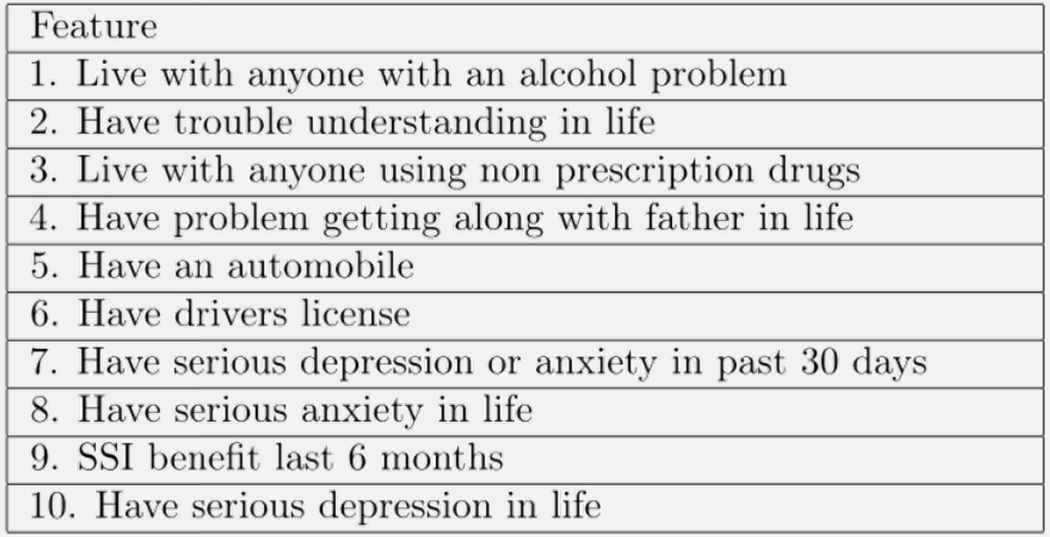


Scatter plot of odds ratios versus matched group size

Real Data: BTC

- Breaking The Cycle (BTC) (Harrell et al., 2006) is a social program conducted in several U.S. states designed to reduce criminal involvement and substance abuse among current offenders. The effect of participating in the program on reducing non-drug future arrest rates is studied.

Table 2: Features for BTC data.



Feature
1. Live with anyone with an alcohol problem
2. Have trouble understanding in life
3. Live with anyone using non prescription drugs
4. Have problem getting along with father in life
5. Have an automobile
6. Have drivers license
7. Have serious depression or anxiety in past 30 days
8. Have serious anxiety in life
9. SSI benefit last 6 months
10. Have serious depression in life

DAME and FLAME Dropping Order

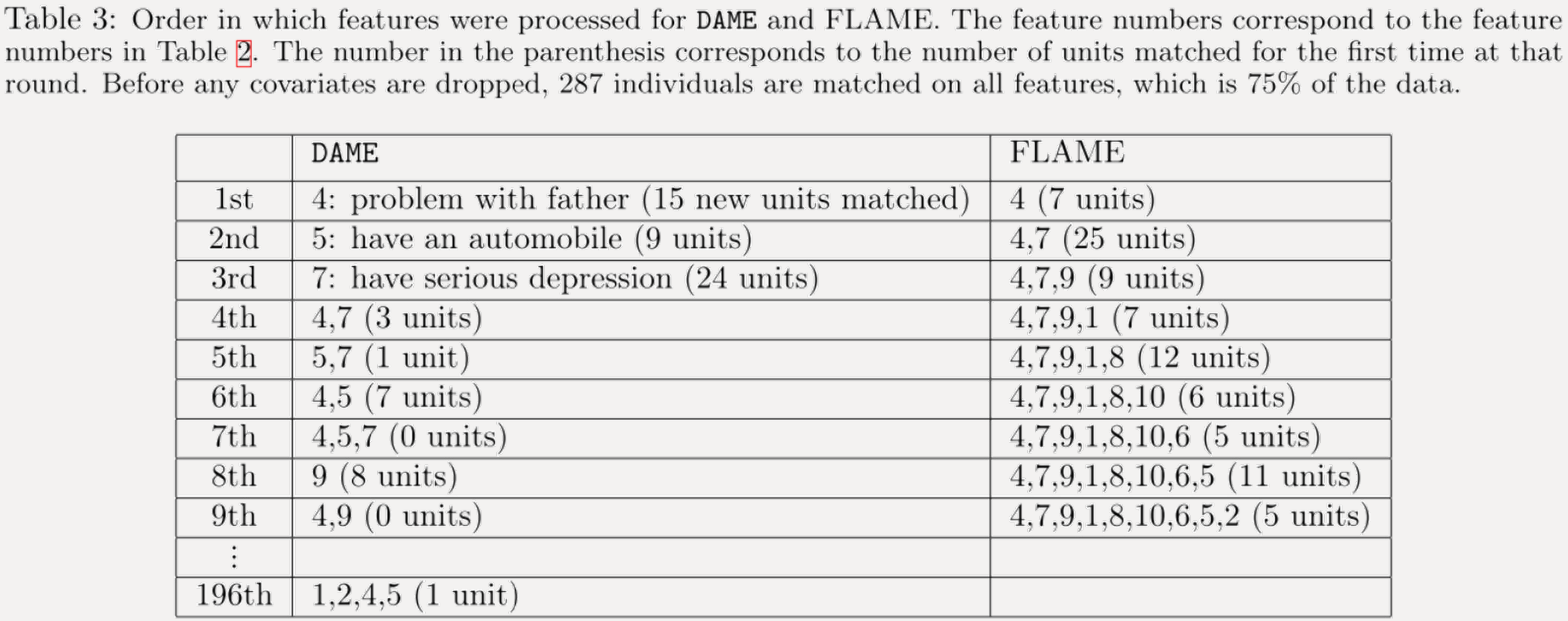


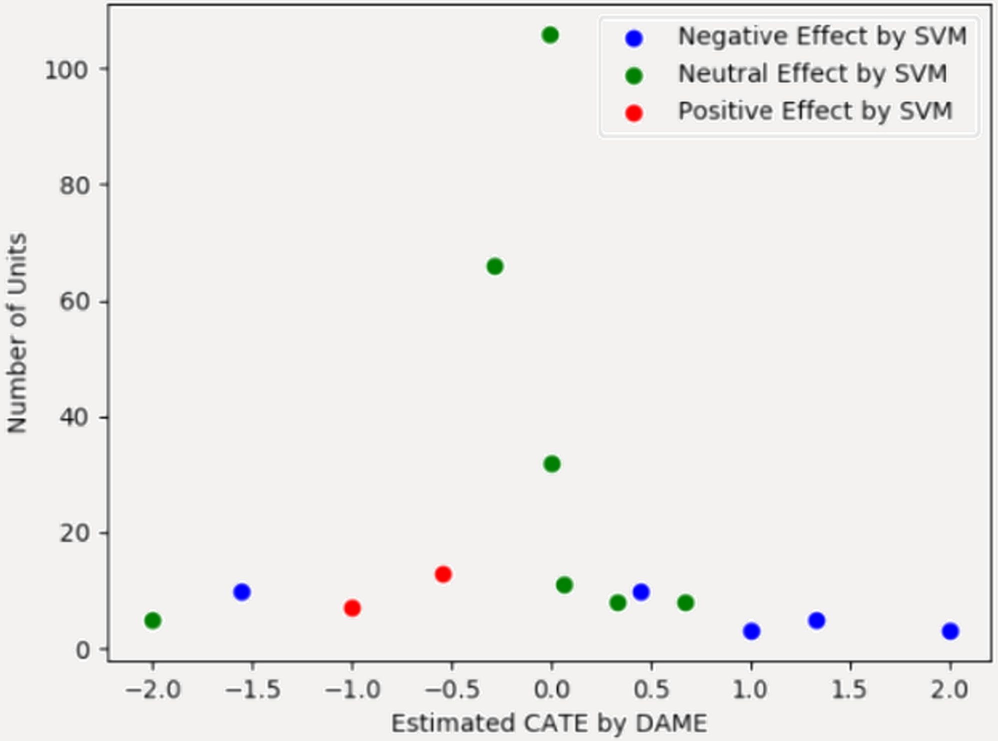
Table 3: Order in which features were processed for DAME and FLAME. The feature numbers correspond to the feature numbers in Table 2. The number in the parenthesis corresponds to the number of units matched for the first time at that round. Before any covariates are dropped, 287 individuals are matched on all features, which is 75% of the data.

	DAME	FLAME
1st	4: problem with father (15 new units matched)	4 (7 units)
2nd	5: have an automobile (9 units)	4,7 (25 units)
3rd	7: have serious depression (24 units)	4,7,9 (9 units)
4th	4,7 (3 units)	4,7,9,1 (7 units)
5th	5,7 (1 unit)	4,7,9,1,8 (12 units)
6th	4,5 (7 units)	4,7,9,1,8,10 (6 units)
7th	4,5,7 (0 units)	4,7,9,1,8,10,6 (5 units)
8th	9 (8 units)	4,7,9,1,8,10,6,5 (11 units)
9th	4,9 (0 units)	4,7,9,1,8,10,6,5,2 (5 units)
⋮		
196th	1,2,4,5 (1 unit)	

- DAME is able to construct matched groups by only dropping subsets of what FLAME drops as early as the second and third iteration of the algorithm.
- DAME matches on more covariates than FLAME.

Comparison With a Black Box Model

- CATE predictions of DAME compared with a black box SVM that predicts positive, neutral, or negative effect of each individual.
- The discrepancy between the two methods could be tackled by smoothing:
    - units within the leftmost blue (negative) labeled matched group were much closer (in hamming distance) to other blue (negative) labeled matched groups than to green (neutral) or red (positive) labeled groups

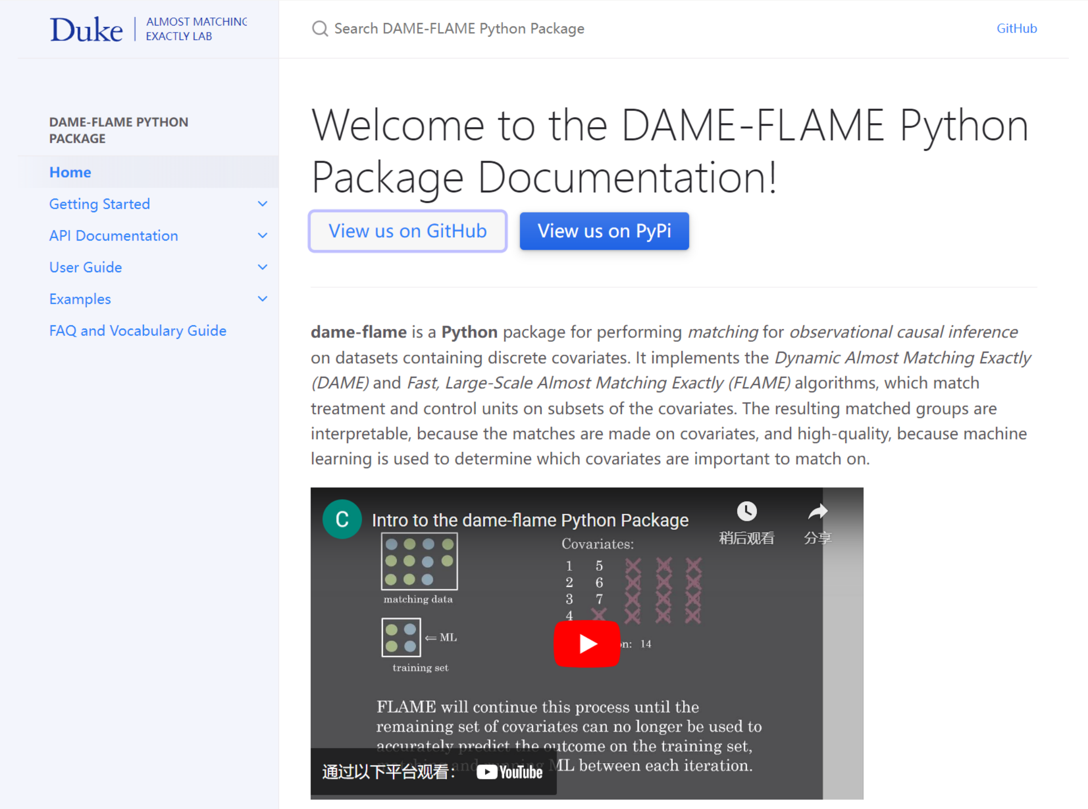


Duke

Conclusion and Limitations

- Conclusion:
  - DAME produces interpretable matches that are of high quality.
  - FLAME is a method for adaptive, interpretable, large-scale matching. The bias of FLAME can be calculated directly in specific settings.
- Limitations:
  - Continuous Variables Unfriendly (We can discretize the feature but the result is not that accurate)
    - Adaptive Hyperboxes
      - Creates an adaptive axis-parallel box for continuous and discrete datasets
    - MALTS – Matching After Learning to Stretch
      - Creates an interpretable stretch metric

Usage of DAME and FLAME



Duke | ALMOST MATCHING EXACTLY LAB
Search DAME-FLAME Python Package
GitHub
DAME-FLAME PYTHON PACKAGE
Home
Getting Started
API Documentation
User Guide
Examples
FAQ and Vocabulary Guide

Welcome to the DAME-FLAME Python Package Documentation!

dame-flame is a **Python** package for performing *matching* for *observational causal inference* on datasets containing discrete covariates. It implements the *Dynamic Almost Matching Exactly (DAME)* and *Fast, Large-Scale Almost Matching Exactly (FLAME)* algorithms, which match treatment and control units on subsets of the covariates. The resulting matched groups are interpretable, because the matches are made on covariates, and high-quality, because machine learning is used to determine which covariates are important to match on.

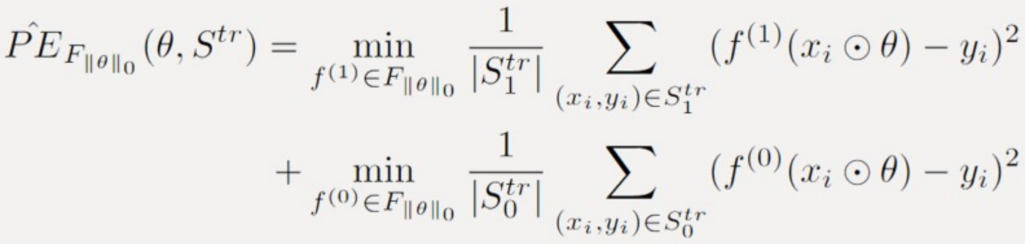
FLAME will continue this process until the remaining set of covariates can no longer be used to accurately predict the outcome on the training set, ML between each iteration.

- A Python Package created by Almost Matching Exactly Lab is presented and can be easily installed through pypl (https://almost-matching-exactly.github.io/DAME-FLAME-Python-Package/)

Reference

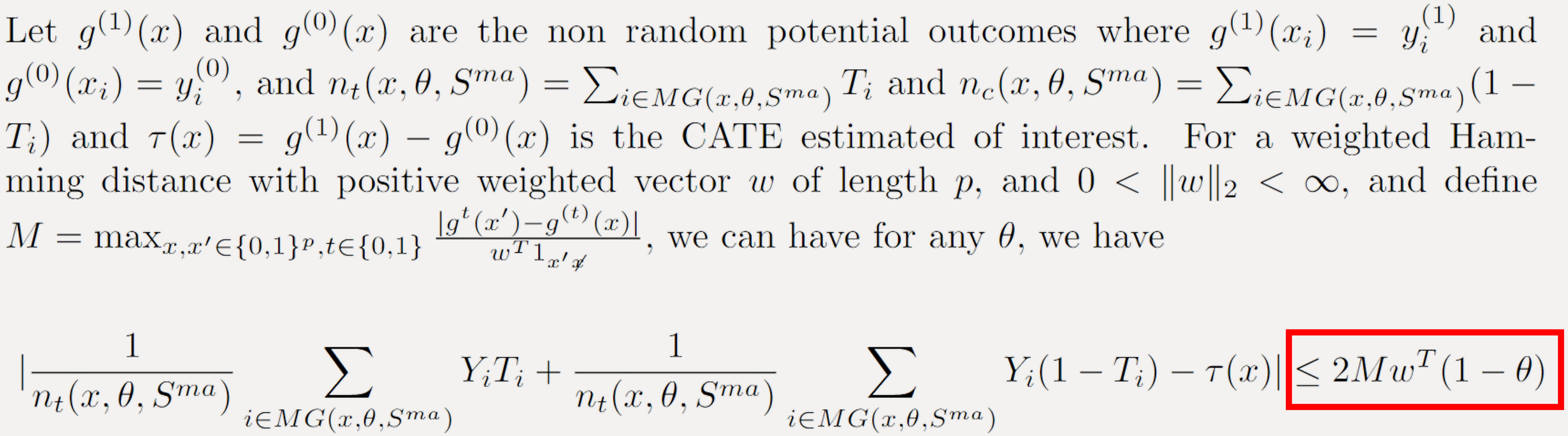
- Wang, T., Morucci, M., Awan, M. U., Liu, Y., Roy, S., Rudin, C., & Volfovsky, A. (2021). Flame: A fast large-scale almost matching exactly approach to causal inference. *The Journal of Machine Learning Research*, *22*(1), 1477-1517.
- Liu, Y., Dieng, A., Roy, S., Rudin, C., & Volfovsky, A. (2018). Interpretable almost matching exactly for causal inference. *arXiv preprint arXiv:1806.06802*.
- Morucci, M., Orlandi, V., Roy, S., Rudin, C., & Volfovsky, A. (2020, August). Adaptive hyper-box matching for interpretable individualized treatment effect estimation. In *Conference on Uncertainty in Artificial Intelligence* (pp. 1089-1098). PMLR.
- Parikh, H., Rudin, C., & Volfovsky, A. (2018). Malts: Matching after learning to stretch.
- Zubizarreta, J. R. (2012). 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.
- Rosenbaum, P. R. (1989). Optimal matching for observational studies. Journal of the American Statistical Association, 84(408), 1024-1032.
- Schneeweiss, S., Rassen, J. A., Glynn, R. J., Avorn, J., Mogun, H., & Brookhart, M. A. (2009). High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. Epidemiology (Cambridge, Mass.), 20(4), 512.
- Hahn, J. (2004). Functional restriction and efficiency in causal inference. The Review of Economics and Statistics, 86(1), 73-76.
- We have also referenced the presentation presented by Professors in AME Lab (https://www.youtube.com/watch?v=-So_cL-eMFQ)

Appendix: Bias in CATE



$$\hat{PE}_{F_{\|\theta\|_0}}(\theta, S^{tr}) = \min_{f^{(1)} \in F_{\|\theta\|_0}} \frac{1}{|S_1^{tr}|} \sum_{(x_i, y_i) \in S_1^{tr}} (f^{(1)}(x_i \odot \theta) - y_i)^2 + \min_{f^{(0)} \in F_{\|\theta\|_0}} \frac{1}{|S_0^{tr}|} \sum_{(x_i, y_i) \in S_0^{tr}} (f^{(0)}(x_i \odot \theta) - y_i)^2$$

- If we do not match on all relevant covariates, a bias is induced on the treatment effect estimates. Here we present a simple worst-case bound on the in-sample estimation bias when a CATE is estimated with units matched according to a chosen subset of covariates.



Let $g^{(1)}(x)$ and $g^{(0)}(x)$ are the non random potential outcomes where $g^{(1)}(x_i) = y_i^{(1)}$ and $g^{(0)}(x_i) = y_i^{(0)}$, and $n_t(x, \theta, S^{ma}) = \sum_{i \in MG(x,\theta,S^{ma})} T_i$ and $n_c(x, \theta, S^{ma}) = \sum_{i \in MG(x,\theta,S^{ma})} (1 - T_i)$ and $\tau(x) = g^{(1)}(x) - g^{(0)}(x)$ is the CATE estimated of interest. For a weighted Hamming distance with positive weighted vector w of length p, and $0 < \|w\|_2 < \infty$, and define $M = \max_{x, x' \in \{0,1\}^p, t \in \{0,1\}} \frac{|g^t(x') - g^{(t)}(x)|}{w^T 1_{x' \not x}}$, we can have for any θ, we have

$$\left| \frac{1}{n_t(x, \theta, S^{ma})} \sum_{i \in MG(x,\theta,S^{ma})} Y_i T_i + \frac{1}{n_t(x, \theta, S^{ma})} \sum_{i \in MG(x,\theta,S^{ma})} Y_i (1 - T_i) - \tau(x) \right| \boxed{\leq 2Mw^T(1 - \theta)}$$

Duke