# Causal Learning in Social Sciences

## Theory and Applications

Jingzou Ron Huang[1]
Jiuyao Joe Lu[2]
Milan Mossé[3]
Alexander Williams Tolbert[4]

[1]Senior Undergraduate, Quantitative Sciences and Economics, Emory University
[2]PhD Student, Statistics, Wharton School, University of Pennsylvania
[3]PhD Student, Philosophy, University of California at Berkeley
[4]Assistant Professor, Quantitative Theory and Methods, Emory University

April 16, 2025

# Background and Motivations

- The problem of variable choice (Woodward, 2016):
  - given a pre-selected stock of candidate variables, which should be incorporated into a causal model for some system?
  - construct or define new previously unconsidered variables either de novo or by transforming or combining or aggregating old variables?
- Causal Feature Learning (CFL) is an algorithm designed to construct macrovariables that preserve the causal relationships between variables (Chalupka, 2016).
- CFL has been used with neural data as cause and behavioral data as effect or with climate data on both the cause and the effect side.
- We would like to apply it to social science data.
- What are some good properties of the CFL algorithm in the causal inference of social sciences?

# Preliminaries

Prediction Versus Causation

- In prediction we are interested in

$$P(Y \in A \mid X = x)$$

  which means: the probability that $Y \in A$ given that we observe that $X$ is equal to $x$.

- For causation we are interested in

$$P(Y \in A \mid \text{ do } X = x)$$

  which means: the probability that $Y \in A$ given that we set $X$ equal to $x$.

- Prediction is about passive observation. Causation is about active intervention.

# Theory of CFL

## Preliminaries and assumptions

- We can learn higher-level observational/causal features by partitioning the covariate space based on those conditional probabilities.

**Definition (Observational Partition)**

The observational partition $\Pi_o(\mathcal{X})$ is induced by the equivalence relation

$$X_1 \sim X_2 \iff \forall Y \in \mathcal{Y},\ P(Y \mid X = X_1) = P(Y \mid X = X_2)\,.$$

**Definition (Causal Partition)**

The causal partition $\Pi_c(\mathcal{X})$ is induced by the equivalence relation

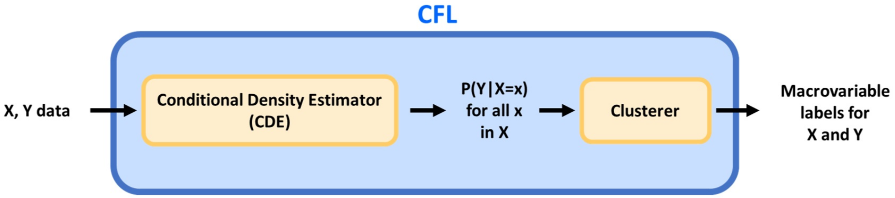
$$X_1 \sim X_2 \iff \forall Y \in \mathcal{Y},\ P(Y \mid \mathrm{do}(X_1)) = P(Y \mid \mathrm{do}(X_2))\,.$$

# The CFL Algorithm

## Preliminaries and assumptions

### Assumption (Discrete Macrovariables)

All macrovariables are discretized into a finite number of states. This assumption ensures that the state space is finite, enabling the application of clustering methods for partitioning the data space.




An overview of the CFL pipeline

- Mixture density network to estimate conditional probabilities
- Cluster observations by KMeans

# The CFL Algorithm

## Extension to Social Sciences

- In social science research, we can have a few treatments, but it is hard to intervene on all covariates of interest.
- Causal partition is almost impossible in social science data
- But we can still make use of observation partition, which has some good properties.
- Two major application of observation partition we explore in this project
  - Heterogeneity detection
  - Dimensionality Reduction

# Heterogeneity on Average

Heterogeneity arises when

$$\mathbb{E}[Y \mid D=1, X=x_j] - \mathbb{E}[Y \mid D=0, X=x_j]$$
$$\neq \mathbb{E}[Y \mid D=1, X=x_i] - \mathbb{E}[Y \mid D=0, X=x_i]$$

for some $i, j$. If

$$P(Y \mid D=1, X=x_j) - P(Y \mid D=0, X=x_j)$$
$$\neq P(Y \mid D=1, X=x_i) - P(Y \mid D=0, X=x_i),$$

If we assume expectation is different as long as the distribution is different, the second inequality can imply the previous inequality. Thus, if CFL can detect some $i, j$ such that the second inequality holds, CFL will manifest heterogeneity.

# Heterogeneity on Average

1. For some $j$, $(D = 1, X = x_j) \sim (D = 0, x = x_j)$, which means they are clustered into one macrostate, so that

$$P(Y \mid D = 1, X = x_j) = P(Y \mid D = 0, X = x_j),$$

but for some $i$, $(D = 1, X = x_i)$ is not in the same equivalence class as $(D = 0, X = x_i)$, so

$$P(Y \mid D = 1, X = x_i) \neq P(Y \mid D = 0, X = x_i),$$

meaning that treatment has no effect on some subpopulations but has an effect on others.

# Heterogeneity on Average

2. For some $i, j$, $(D = 1, X = x_i) \sim (D = 1, X = x_j)$, so

$$P(Y \mid D = 1, X = x_j) = P(Y \mid D = 1, X = x_i),$$

but $(D = 0, X = x_i)$ is not in the same equivalence class as $(D = 0, X = x_j)$, so

$$P(Y \mid D = 0, X = x_i) \neq P(Y \mid D = 0, X = x_j),$$

meaning that $P(Y \mid D = 1, X) - P(Y \mid D = 0, X)$ is not constant across all values of $X$.

# The National Supported Work (NSW) Dataset

- Originally used to study the impact of labor training program on earning
- Treatment dummy + 6 demographic/socioeconomic variables (age, education, race) + 2 income variables (pre-intervention income in 1975 and post-intervention income in 1978, all in 1982 dollar)
- Randomized treatment

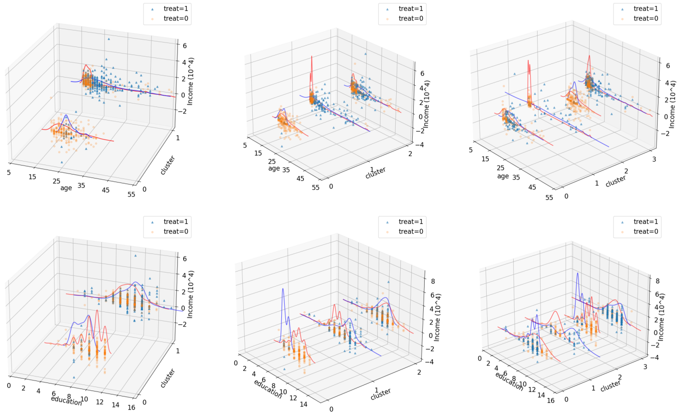
$$Y_i = \beta_0 + \beta_1 \text{Treat}_i + \beta_2 \mathbb{1}[age_i > \overline{age}] + \beta_3 \text{Treat}_i \mathbb{1}[age_i > \overline{age}] + \epsilon_i$$

Table 1: NSW Regression Results: Heterogeneity Identification

| **Variable** | **Coef.** | **Std. Err.** | **t** | **P > \|t\|** |
|---|---|---|---|---|
| Intercept | 2790.36 | 476.31 | 5.85 | 0.00 |
| treat | -409.66 | 742.27 | -0.55 | 0.58 |
| $\text{age}_{\text{dummy}}$ | -1670.13 | 721.93 | -2.31 | 0.02 |
| $\text{age}_{\text{dummy}} \times$ treat | 2889.34 | 1125.78 | 2.56 | 0.01 |

# Heterogeneity Detection by CFL in Practice

Figure 1: Distribution of Treated and Untreated Units Across Clusters with Kernel Density Estimates



# Heterogeneity Detection by CFL in Practice

- Randomized treatment is necessary for CFL to correctly identify heterogeneity

$$
\begin{aligned}
&\mathbb{E}[Y(1) \mid D = 1, X = x_j] - \mathbb{E}[Y(0) \mid D = 0, X = x_j] \\
&\neq \mathbb{E}[Y(1) \mid D = 1, X = x_i] - \mathbb{E}[Y(0) \mid D = 0, X = x_i]
\end{aligned}
$$

$$
\begin{aligned}
\Rightarrow &\mathbb{E}[Y(1) - Y(0) \mid D = 1, X = x_j] \\
&\underbrace{+\mathbb{E}[Y(0) \mid D = 1, X = x_j] - \mathbb{E}[Y(0) \mid D = 0, X = x_j]}_{\text{Selection Bias}} \\
\neq\ &\mathbb{E}[Y(1) - Y(0) \mid D = 1, X = x_i] \\
&+ \underbrace{\mathbb{E}[Y(0) \mid D = 1, X = x_i] - \mathbb{E}[Y(0) \mid D = 0, X = x_i]}_{\text{Selection Bias}}
\end{aligned}
$$

- Therefore, there could be no heterogeneity e.g. if $P[Y(1) - Y(0) \mid D = 1, x = x_j] = P[Y(1) - Y(0) \mid D = 1, x = x_i]$, but different degrees of selection biases across values of covariate will be confounding.

# CFL as Dimension Reduction Techinique

**Lemma**

*Let $X$ be the random vector for all relevant covariates such that the unconfoundedness assumption $D \perp\!\!\!\perp (Y(1), Y(0)) \mid X$ holds, then the observational coarsening, $M$, of the covariate space of $X$ by CFL also satisfy $D \perp\!\!\!\perp Y(\cdot) \mid M$*

- Suppose $\{x_i\}_{i=1}^{\infty}$ is the sequence of all possible values in $\mathcal{X}$, and $\{x_{ik}\}_{k=1}^{n}$ is a subsequence that includes all the representatives of $n$ equivalent classes. Define a new random vector $M$ such that

$$M = \begin{bmatrix} 1[X \in [x_{i1}]] \\ 1[X \in [x_{i2}]] \\ \vdots \\ 1[X \in [x_{in-1}]] \end{bmatrix}.$$

# CFL as Dimension Reduction Technique

- $D \perp\!\!\!\perp (Y(1), Y(0)) \mid X \Rightarrow P(Y(\cdot) \mid D, X) = P(Y(\cdot) \mid X)$
- Within each equivalence class $[x_{ik}]$, the conditional distribution of $Y(\cdot)$ given $X$ is constant by definition of observational partition $\Rightarrow P(Y(\cdot) \mid X, M) = P(Y(\cdot) \mid M) \to D \perp\!\!\!\perp Y(\cdot) \mid M$.
- Together with the law of total probability:

$$
\begin{aligned}
P(Y(\cdot) \mid D, M) &= \int P(Y(\cdot) \mid D, X, M)\, P(X \mid D, M)\, dX \\
P(Y(\cdot) \mid D, M) &= \int P(Y(\cdot) \mid M)\, P(X \mid D, M)\, dX \\
P(Y(\cdot) \mid D, M) &= P(Y(\cdot) \mid M) \int P(X \mid D, M)\, dX. \\
P(Y(\cdot) \mid D, M) &= P(Y(\cdot) \mid M).
\end{aligned}
$$

- $P(Y(\cdot) \mid D, M) = P(Y(\cdot) \to D \perp\!\!\!\perp Y(\cdot) \mid M$

# CFL as Dimension Reduction Technique

- Therefore, the coarsening of the covariate space does not affect the conditional independence between the outcome and treatment
- The treatment is still as if randomized after controlling for all the macrovariables created by the CFL algorithm

$$
\begin{aligned}
&\mathbb{E}[\mathbb{E}[Y \mid D=1, M] - \mathbb{E}[Y \mid D=0, M]] \\
=&\mathbb{E}[\underbrace{\mathbb{E}[Y(1) - Y(0) \mid D=1, M]}_{\text{ATT}}] \\
&+ \mathbb{E}[\underbrace{\mathbb{E}[Y(0) \mid D=1, M] - \mathbb{E}[Y(0) \mid D=0, M]}_{\mathbb{E}[Y(0)|M]-\mathbb{E}[Y(0)|M]=0}] \\
=&\underbrace{\mathbb{E}[Y(1) - Y(0) \mid M]}_{\text{ATE}}
\end{aligned}
$$

# Future Work

- How well does the clustering technique approximate the partition induced by the equivalence relation?
  - Currently we only assume that those clustering techniques are good approximations of partition induced by the defined equivalence relation.
  - How to ensure a good approximation or to lower bound the approximation?
- Continous macrovariable exists, such as temperature. How to extend the CFL framework to continuous cases and apply it in social science research?
  - What about the hybrid case?