Causal Clustering: A new approach to analysis of treatment effect heterogeneity

Kwangho Kim

{Statistics, Machine Learning} Department
Carnegie Mellon University

Joint work with
Edward Kennedy, Jisu Kim, Larry Wasserman

Contents

1. Motivation
2. Causal Clustering
3. Adaptation to three widely-used clustering algorithms
    - k-means
    - hierarchical
    - density (level-set)
4. Efficient k-means causal clustering
5. Application

Motivation

Average Treatment Effect

We begin with considering data structure

$$Z = (X, A, Y) \sim \mathbb{P}$$

where we have covariates $X \in \mathbb{R}^d$, treatment $A \in \{0, 1\}$, and outcome $Y \in \mathbb{R}$.

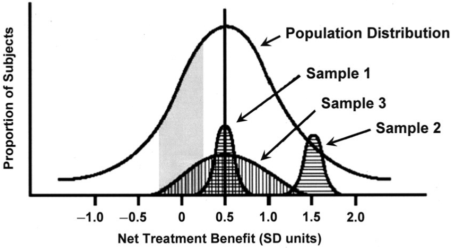
Y^a: potential outcome under treatment a.

The population-level *average treatment effect* (ATE) is defined by

$$\mathbb{E}_{\mathbb{P}}(Y^1 - Y^0). \tag{1}$$

Heterogeneity in treatment effects

In many cases, we have a non-random variability in direction/magnitude of treatment effects



In this case, the standard ATE does not help to find an optimal policy.

Figure: Kravitz et al. 2004

Heterogeneity in treatment effects

Identifying treatment effect heterogeneity and corresponding subgroups are of great importance

- cancer treatment [Zhang et al. 2017]
- efficacy of social programs [Imai and Ratkovic 2013]

Previous approaches

Conditional average treatment effects (CATE):

$$\tau(X) = \mathbb{E}_{\mathbb{P}}[Y^1 - Y^0 \mid X] \tag{2}$$

Goal: find subgroups whose units have similar CATE

Previous attempts:

- simple parametric regression [e.g. Imai and Ratkovic 2013, Robins 1991]
- recursive partitioning via tree-based methods [e.g. Athey and Imbens 2015, Doove 2014]
- other supervised-learning [e.g. Kunzel 2017, van der Laan and Luedtke 2014]

Limitations

- parametric restrictions
- not directly expandable to outcome-wide study [e.g. VanderWeele et al 2017, 2016, Li et al 2016] or multiple treatments [e.g. Lopez et al 2017]
- some drawbacks of the widely-used recursive partitioning methods
  - inefficient when lots of leafs have same effects
  - perform not very well for continuous variables [e.g. Lee et al 2017]
  - trade-off between reducing noise and decreasing bias

Causal Clustering

Setup & Assumptions

Consider i.i.d samples from data structure $Z = (X, A, Y) \sim \mathbb{P}$, where

$$\mathcal{X} \in \mathbb{R}^d, \quad \mathcal{A} = \{0, 1, ..., p-1\}, \quad \mathcal{Y} \in \mathbb{R}.$$

Causal & Boundedness assumptions: for $\forall a \in \mathcal{A}$

- (A1) (consistency) $Y = \sum_a \mathbb{1}\{A = a\} Y^a$
- (A2) (no unmeasured confounding) $A \perp\!\!\!\perp Y^a \mid X$
- (A3) (positivity) $\mathbb{P}(A = a \mid X)$ is bounded away from 0 a.s.
- (A4) $\mathbb{E}[Y^a|X]$ is globally bounded $\forall a$.

All the pairwise CATE's are identified under (A1)-(A3).

Representation map

Definition (Representation map)

We define a map $\Phi : \mathcal{X} \to \mathbb{R}^p$ by

$$\Phi(X) = \left(\mathbb{E}[Y^0 \mid X], \dots, \mathbb{E}[Y^{p-1} \mid X]\right). \tag{3}$$

Let $\mu_a \equiv \mathbb{E}[Y \mid X, A = a]$. Under (A1)-(A3), $\Phi(X)$ can be constructed by estimating μ_a for $a = 0, \dots, p-1$.

Representation map: implication

On the image of Φ,

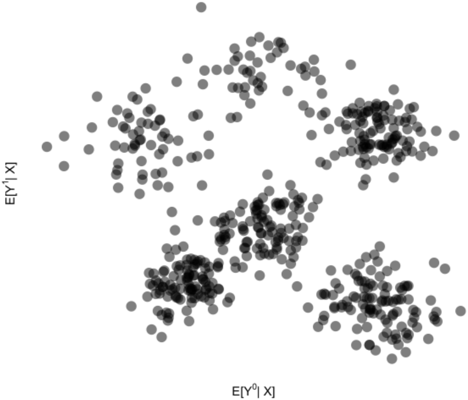
- a point whose coordinates are mostly the same
  $\Rightarrow$ no treatments bring any visible effect
- for two unites i, j,

$$\Phi(X_i) \approxeq \Phi(X_j) \Rightarrow \tau_{a,0}(X_i) \approxeq \tau_{a,0}(X_j) \text{ for } \forall a \in \mathcal{A}$$

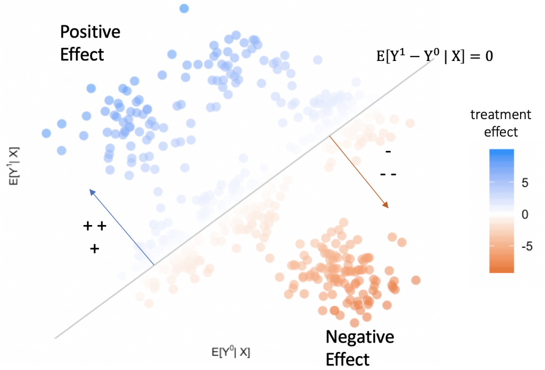
where $\tau_{a,0}(X) = \mathbb{E}[Y^a - Y^0 \mid X]$: i.e., the effect of receiving treatment a over placebo (a=0).

Illustrating example

Consider samples projected through the representation map, where $\mathcal{A} = \{0, 1\}$ and $\mathbb{E}[Y^1 - Y^0] = 0$.

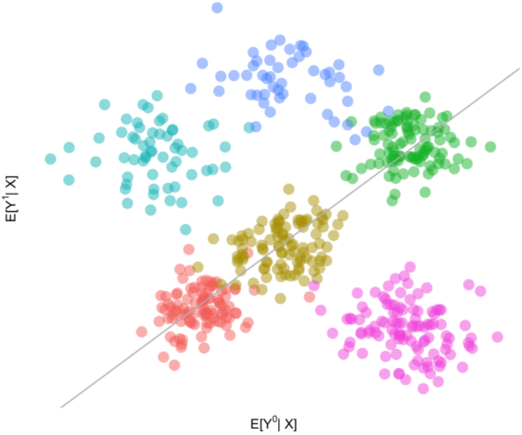


Illustrating example



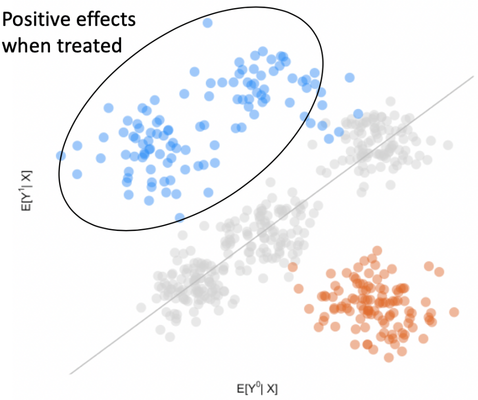
Illustrating example

It would be worth analyzing each *cluster* separately (e.g. k-means),



Illustrating example

or based on the distance from $\mathbb{E}[Y^1 - Y^0 \mid X] = 0$ line.



Causal Clustering: the idea

Analysis of treatment effect heterogeneity:

- need to ascertain a subgroup that shows similar responses towards given treatments (in terms of CATE)

$\Rightarrow$ Perform **cluster analysis** on the **image of Φ**.

Adaptation to three widely-used clustering algorithms

Main result I

Challenges

- every coordinate $\mu_a = \mathbb{E}[Y^a|X]$ in Φ is a *random function* that needs to be estimated

Our result

- We show that for three widely-used clustering algorithms (k-means, hierarchical, density),
  the additional cost comes out to be the cost of estimating μ_a's (as a linearly additive error).

k-means causal clustering

$\widehat{C}$: sample splitting → plug-in → empirical risk minimizer

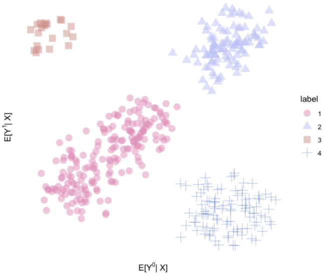
Theorem (Error bound for k-means causal clustering)

Under the same conditions of Linder et al (1994), there exists an N such that for every $n > N$

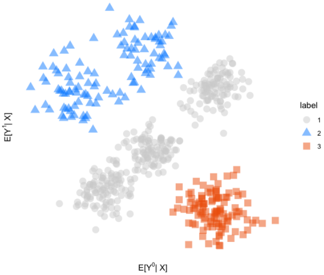
$$\mathbb{E}\left|R(\widehat{C}) - R(C^*)\right| \leq \underbrace{64\, B^2 \sqrt{\frac{k(d+1)\log n}{n}}}_{\textit{Linder et al (1994)}} + \underbrace{4\sqrt{2}B \sum_{a \in \mathcal{A}} \|\widehat{\mu_a} - \mu_a\|}_{\textit{additional cost}}.$$

Hierarchical & (level-set) Density clustering

We also verify *Hierarchical* and *Density* causal clustering can be done at the additional error/risk of $O\left(\sum_a \|\widehat{\mu_a} - \mu_a\|\right)$



Density-based clustering



Hierarchical clustering

Efficient k-means causal clustering

Nonparametric condition on nuisance parameters

- Cost of $\sum_a \|\widehat{\mu_a} - \mu_a\|$ seems expensive; to attain $n^{-1/2}$ rates overall, we need to estimate each μ_a at $\boldsymbol{n^{-1/2}}$ rate which is infeasible in nonparametric modeling
- We may want to utilize information about treatment process (i.e., propensity score)

Semiparametric approach

$$R(C) \underbrace{\xrightarrow{\qquad}}_{\substack{\text{kernel}\\ \text{smoothing}}} R_h(C) \underbrace{\xrightarrow{\qquad}}_{\substack{\text{efficient}\\ \text{influence}\\ \text{function}}} \Psi(C;\mu,\pi) \underbrace{\xrightarrow{\qquad}}_{\substack{\text{sample}\\ \text{splitting}}} \widehat{\Psi}(C;\hat{\mu},\hat{\pi})$$

$$\widehat{C} = \arg\min_{C} \widehat{\Psi}(C;\hat{\mu},\hat{\pi})$$

We will focus on *k-means causal clustering*

Main result II: Efficient k-means causal clustering

Theorem (Error bound)

Under the margin condition (Levrard 2015, 2018) and other weak conditions, if

- $\sum_{a,a' \in \mathcal{A}} \|\pi_a - \widehat{\pi_a}\| \|\mu_{a'} - \widehat{\mu_{a'}}\| = o_{\mathbb{P}}(n^{-1/2})$
- $\sum_{a,a' \in \mathcal{A}} \|\mu_a - \widehat{\mu_a}\| \|\mu_{a'} - \widehat{\mu_{a'}}\| = o_{\mathbb{P}}(n^{-1/2})$

then

$$R(\widehat{C}) - R(C^*) = O_{\mathbb{P}}\left(\frac{1}{\sqrt{n}}\right).$$

Sufficient condition: now μ, π can be estimated at $\boldsymbol{n^{-1/4}}$ rates.

Efficient k-means causal clustering

Theorem (Asymptotic normality)

Under the stronger version of the margin condition along with the other proper assumptions, we have

$$\sqrt{n}(\widehat{C} - C^*) \rightsquigarrow N\left(0, \Sigma'_{C^*,\eta}\right)$$

where $\eta = (\pi, \mu)$ *and* $\Sigma'_{C^*,\eta}$ *is* $kp \times kp$ *covariance matrix.*

- Our estimate of $\widehat{C}$ satisfies $\sqrt{n}$-consistent, asymptotic normality property, under weak NP conditions.

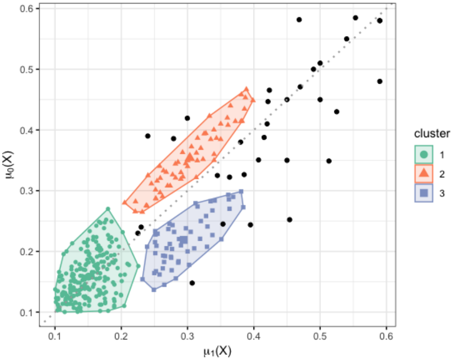
Application

Application: the EAGeR aspirin data[1]

Goal: study the effect of aspirin on pregnancy loss

$A \in \{0,1\}$: low-dose aspirin, $Y \in \mathbb{R}$: indicator of pregnancy loss,

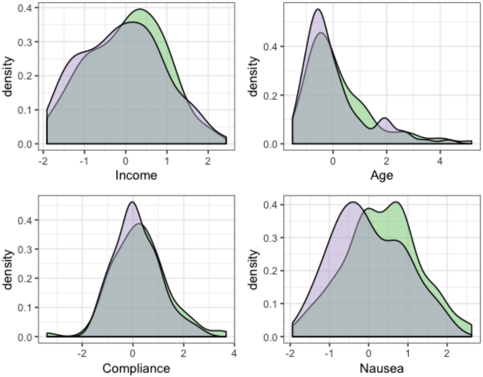
$X \in \mathbb{R}^d$: pretreatment covariates $\Rightarrow \widehat{\mathbb{E}}[Y^1 - Y^0] \approx 0$



[1] https://www.nichd.nih.gov/about/org/diphr/officebranch/eb/effects-aspirin

Application: aspirin data

- seems 'Nausea' drives the difference



Conclusion

- Causal Clustering: a new framework for the analysis of treatment effect heterogeneity by leveraging tools in clustering analysis
    - pursue an intuitive way of ascertaining subgroups with similar treatment effects based on *unsupervised* method
- show that three widely-used clustering methods can be successfully adopted into our framework
- develop efficient k-means causal clustering algorithm that attains fast convergence rates/asymptotic normality even when incorporating flexible machine learning methods

End of Talk

Thank you

Appendix

Appendix

Margin condition (Levrard 2015, 2018)

Definition (Margin condition)

Let us define $p(t) := \sup_{C \in \mathcal{M}^*} \mathbb{P}(W \in N_C(t))$. We assume that there exists a fixed $\kappa > 0$ such that for all $0 \leq t \leq \kappa$

$$p(t) \lesssim t^{\alpha}$$

for some $\alpha > 0$.

hierarchical clustering

Theorem (Balcan et al (2014))

Suppose each $\widehat{\mu}_a$ is estimated in the separate sample set D^n *and let similarity function K (induced from Euclidean distance d) satisfy the (α, ν)-good neighborhood property for the clustering problem (S, l). Then under the additional set of assumptions (A1)-(A4), we have robust hierarchical clustering (Balcan et al, 2014) on $(\hat{S}, l)$ with a pruning that have error at most $\nu + \xi + \delta$ with respect to the true target clustering on (S, l) with probability at least $1 - \delta$, where $\xi = O(\sum_{a \in \mathcal{A}} \|\widehat{\mu_a} - \mu_a\|_\infty)$.*

.

(level-set) density clustering

Theorem (Rinaldo et al (2010), Kim et al (2018))

Suppose that $L_{h,t}$ is stable and let $H(\cdot,\cdot)$ be the Hausdorff distance between two sets. Suppose each $\widehat{\mu}_a$ is estimated in the separate sample set D^n*, and suppose Assumptions (A1)-(A6). Let $\{h_n\}_{n\in\mathbb{N}} \subset (0, h_0)$ be satisfying*

$$\limsup_n \frac{(\log(1/h_n))_+}{nh_n^2} < \infty.$$

Then,

$$H(\widehat{L}_t, L_{h,t}) = O_P\left(\sqrt{\frac{(\log(1/h_n))_+}{nh_n^2}} + \frac{1}{h_n^3}\min\left\{\sum_a \|\widehat{\mu_a} - \mu_a\|_1\,,\, h_n\right\}\right)$$