September 15, 2020
Online Causal Inference Seminar

Localized Debiased Machine Learning:

Efficient Inference on Quantile Treatment Effects and Beyond

Nathan Kallus† Xiaojie Mao† Masatoshi Uehara
Cornell University
†Presenting today

Paper: `https://arxiv.org/abs/1912.12945`
Code: `https://github.com/CausalML/LocalizedDebiasedMachineLearning`

Quantiles vs Averages 🤔

- Hypothetical new program: teach *everyone nationwide* basic computer + MS Office skills in high school so they can also gain access to pool of in-need low-level office jobs 👩‍💻
- Central question in program evaluation: *What's the **effect**?*

Quantiles vs Averages 🤔

- Hypothetical new program: teach *everyone nationwide* basic computer + MS Office skills in high school so they can also gain access to pool of in-need low-level office jobs 👩‍💻
- Central question in program evaluation: *What's the* ***effect****?*
  - Increase in average income: 0.1% 😕

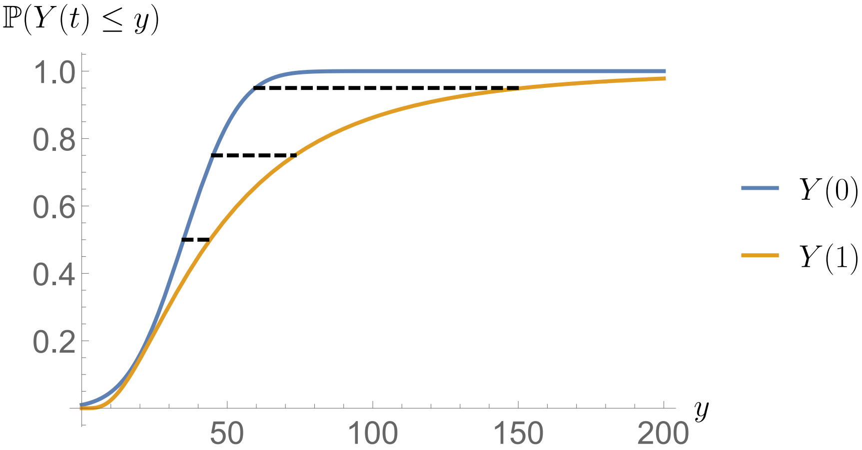
Quantiles vs Averages 🤔

- Hypothetical new program: teach *everyone nationwide* basic computer + MS Office skills in high school so they can also gain access to pool of in-need low-level office jobs 👩‍💻
- Central question in program evaluation: *What's the **effect**?*
    - Increase in average income: 0.1% 😕
        - Aggregate income mostly from unaffected top earners 🛥️🤑

Quantiles vs Averages 🤔

- Hypothetical new program: teach *everyone nationwide* basic computer + MS Office skills in high school so they can also gain access to pool of in-need low-level office jobs 👩‍💻
- Central question in program evaluation: *What's the* ***effect****?*
  - Increase in average income: 0.1% 😕
    - Aggregate income mostly from unaffected top earners 🛥️🤑
  - Increase in 1st quartile of income: 10% 💪

Quantiles vs Averages 🤔



⏩ This Talk in One Sentence ⏩

- *Efficient* estimation of *(L)QTEs* in presence of hi-dim controls/nuisances using just *black-box ML methods for supervised regression** and very lax assumptions

- * means fitting $\mathbb{E}[L \mid X]$ from $X_1, L_1, \ldots, X_N, L_N$
  L can be binary or continuous

How Do We Even Identify QTEs?

- Want to estimate θ^* s.t. $\mathbb{P}(Y(1) \leq \theta^*) = \gamma$ (assume unique)
  - $Y(0)$ quantile and QTE are analogous

How Do We Even Identify QTEs?

- Want to estimate θ^* s.t. $\mathbb{P}(Y(1) \leq \theta^*) = \gamma$ (assume unique)
  - $Y(0)$ quantile and QTE are analogous
- Central problem of causal inference:

  Can't observe counterfactual outcomes

How Do We Even Identify QTEs?

- Want to estimate θ^* s.t. $\mathbb{P}(Y(1) \leq \theta^*) = \gamma$ (assume unique)
  - $Y(0)$ quantile and QTE are analogous
- Central problem of causal inference:

  Can't observe counterfactual outcomes

- Suppose we see $Z = (X, T, Y)$ and we have ignorability $(Y(t) \perp\!\!\!\perp T \mid X)$ and overlap $(0 < \mathbb{P}(T = 1 \mid X) < 1)$ (alternatively: identify *L*QTEs using an IV)
  - Then can use IPW estimating equation to identify θ^*

$$\mathbb{E}[\psi^{\mathsf{IPW}}(Z; \theta^*, \pi^*(X))] = 0$$

$$\text{where} \quad \psi^{\mathsf{IPW}}(Z; \theta, \pi(X)) = \frac{\mathbb{I}[T = 1]}{\pi(X)} \mathbb{I}[Y \leq \theta] - \gamma$$

$$\pi^*(X) = \mathbb{P}(T = 1 \mid X)$$

  - To estimate: replace $\mathbb{E}, \pi$ with $\mathbb{E}_N, \hat{\pi}$ and solve to get $\hat{\theta}^{\mathsf{IPW}}$

The Problem with IPW 😞

- Estimating $\mathbb{P}(T = 1 \mid X)$ is a standard binary regression task
  - 🧰 Lots of flexible ML methods for this task: RF, LASSO, neural nets, CART, BART, xgboost, ... 🛠️ 🤖 🛠️
- Problem: $\hat{\theta}^{\mathsf{IPW}}$ depends *heavily* on how this estimation is done
  - 🦄 In super special cases with extreme smoothness and sieve estimators, $\hat{\theta}^{\mathsf{IPW}}$ can be efficient (Firpo 2007)
  - 🐌 Usually: slowed down by ML estimate's bias (regularization, overparametrization) and sub-$\sqrt{n}$ convergence
- Want to be *insensitive* to how we estimate nuisances

Orthogonality Saves the Day … Or Does It?

- Alternative identification using the *efficient* estimation equation (Robins and Rotnitzky 1994, Tsiatis 2007):

$$\mathbb{E}[\psi(Z;\theta^*,\mu^*(X;\theta^*),\pi^*(X))]=0$$

$$\begin{aligned}\psi(Z;\theta,\mu(X;\theta),\pi(X)) &= \mu(X;\theta)-\gamma \\ &\quad + \frac{\mathbb{I}\left[T=1\right]}{\pi(X)}\left(\mathbb{I}\left[Y\leq\theta\right]-\mu(Z;\theta)\right)\end{aligned}$$

$$\mu^*(X;\theta)=\mathbb{P}\left(Y\leq\theta\mid X,T=1\right)$$

  - **Neyman orthogonality**: $\mathbb{E}[\psi(Z;\theta,\mu(X;\theta),\pi(X))]$ has zero derivative wrt nuisances at θ^*,μ^*,π^*
    - 🥳 *I.e.*: insensitive to errors in nuisances!
  - DML (Chernozhukov et al. 2018): if we split the data into two folds and fit $\hat{\mu},\hat{\pi}$ on the opposite fold of the data Z_i we evaluate at, we get asymptotic behavior akin to using μ^*,π^*

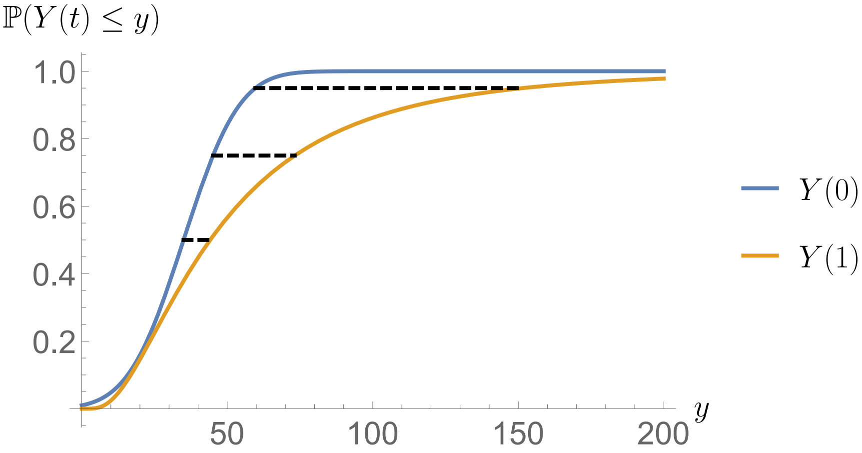
Orthogonality Saves the Day ... Or Does It?

- ... there's just one catch
- Fitting $\mu^*(Z;\theta) = \mathbb{P}\left(Y \leq \theta \mid X, T = 1\right)$ means estimating a *whole* conditional CDF with hi-dim covariates, *i.e.*, a *continuum* of regression functions 😱

$$H = \{\mathbb{P}\left(Y \leq \theta \mid X, T = 1\right) : \theta \in \Theta\}$$

  - Does not exactly fit into standard ML supervised regression
    - Limited options: kernel weights, kNN, mixture of 10 Gaussians param'ed by neural nets
  - Belloni et al. (2017): hypothetical *continuum* of LASSOs 🤯
    And in practice must discretize the range of θ 😞

Orthogonality Saves the Day ... Or Does It?



Orthogonality Saves the Day … Or Does It?

- 🤔 This issue does not appear in ATE estimation using DML
  - 😅 For ATE the efficient estimation equation ψ is *linear* in θ
    Nuisances are π^*, $\mathbb{E}[Y \mid X, T]$ – just regress and plug in
- DML with non-linear equations is *hard*
  - Our work can be understood as a new way to deal with this

Localized DML: the Basic Idea

- We had to estimate a continuum of nuisances because we did not know which θ to use ... what if we *did* know?
  - But θ is what we want to begin with – isn't this a Catch 22?
  - Turns out no: a rough initial guess is enough! 😸
  - *E.g.*, for QTEs, can start with $\hat{\theta}^{\mathsf{IPW}}$ and then refine it using an orthogonal estimating equation where we only estimate $\mu^*(Z; \hat{\theta}^{\mathsf{IPW}})$ (now just a single binary regression task!)

This talk

1. Introduction
2. Method
3. Asymptotic Guarantees
4. Empirical Results
5. Conclusions

The Problem

- Data: $Z_1, \ldots, Z_N$ iid from $\mathbb{P}$
- Target: $\theta^* \in \mathbb{R}^d$ defined by the following d-dimensional moment condition

$$\mathbb{E}\left[\psi(Z; \theta^*, \eta_1^*(Z, \theta^*), \eta_2^*(Z))\right] = 0$$

  - $\eta_1^*(Z, \theta)$ and $\eta_2^*(Z)$ are two unknown but estimable nuisance functions

- Many examples in the paper: QTE under ignorability, LQTE using IV, CVaR, expectiles, equations with incomplete data ...
  - In all cases: efficient estimation equations have *estimand-dependent* nuisances 😬

Oracle Estimating Equation

- *If* knew η_1^*, η_2^*, then could solve the *oracle* empirical equation

$$\tilde{\theta} \text{ solves } \mathbb{E}_N[\psi(Z;\theta,\eta_1^*(Z,\theta),\eta_2^*(Z))] = 0$$

Oracle Estimating Equation

- *If* knew η_1^*, η_2^*, then could solve the *oracle* empirical equation

$$\tilde{\theta} \ \text{ solves } \ \mathbb{E}_N[\psi(Z;\theta,\eta_1^*(Z,\theta),\eta_2^*(Z))] = 0$$

$$\sqrt{N}(\tilde{\theta} - \theta^*) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} J^{*-1} \psi(Z_i;\theta^*,\eta_1^*(Z_i,\theta^*),\eta_2^*(Z_i)) + o_{\mathbb{P}}(1)$$

$$J^* = \partial_{\theta^\top} \mathbb{E}\left[\psi(Z;\theta,\eta_1^*(Z,\theta),\eta_2^*(Z))\right]|_{\theta=\theta^*}$$

Oracle Estimating Equation

- *If* knew η_1^*, η_2^*, then could solve the *oracle* empirical equation

$$\tilde{\theta} \;\; \text{solves} \;\; \mathbb{E}_N[\psi(Z;\theta,\eta_1^*(Z,\theta),\eta_2^*(Z))] = 0$$

$$\sqrt{N}(\tilde{\theta}-\theta^*) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} J^{*-1}\psi(Z_i;\theta^*,\eta_1^*(Z_i,\theta^*),\eta_2^*(Z_i)) + o_{\mathbb{P}}(1)$$

$$J^* = \partial_{\theta^\top}\mathbb{E}\left[\psi(Z;\theta,\eta_1^*(Z,\theta),\eta_2^*(Z))\right]|_{\theta=\theta^*}$$

😕 But actually η_1^*, η_2^* are unknown

Oracle Estimating Equation

- *If* knew η_1^*, η_2^*, then could solve the *oracle* empirical equation

$$\tilde{\theta} \quad \text{solves} \quad \mathbb{E}_N[\psi(Z;\theta,\eta_1^*(Z,\theta),\eta_2^*(Z))] = 0$$

$$\sqrt{N}(\tilde{\theta} - \theta^*) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} J^{*-1} \psi(Z_i;\theta^*,\eta_1^*(Z_i,\theta^*),\eta_2^*(Z_i)) + o_{\mathbb{P}}(1)$$

$$J^* = \partial_{\theta^\top} \mathbb{E}\left[\psi(Z;\theta,\eta_1^*(Z,\theta),\eta_2^*(Z))\right]|_{\theta=\theta^*}$$

😕 But actually η_1^*, η_2^* are unknown

- 🤓 DML = cross-fit η_1^*, η_2^* and plug in estimates

Oracle Estimating Equation

- *If* knew η_1^*, η_2^*, then could solve the *oracle* empirical equation

$$\tilde{\theta} \;\; \text{solves} \;\; \mathbb{E}_N[\psi(Z;\theta,\eta_1^*(Z,\theta),\eta_2^*(Z))] = 0$$

$$\sqrt{N}(\tilde{\theta} - \theta^*) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} J^{*-1} \psi(Z_i;\theta^*,\eta_1^*(Z_i,\theta^*),\eta_2^*(Z_i)) + o_{\mathbb{P}}(1)$$

$$J^* = \partial_{\theta^\top} \mathbb{E}\left[\psi(Z;\theta,\eta_1^*(Z,\theta),\eta_2^*(Z))\right]|_{\theta=\theta^*}$$

😕 But actually η_1^*, η_2^* are unknown

- 🤓 DML = cross-fit η_1^*, η_2^* and plug in estimates
- 😱 But this means estimating a continuum of nuisances!

Invariant Jacobian Assumption

💡 What if we change the oracle equation a little...

$$\tilde{\theta} \quad \text{solves} \quad \mathbb{E}_N[\psi(Z;\theta,\eta_1^*(Z,\theta),\eta_2^*(Z))] = 0$$

Invariant Jacobian Assumption

- What if we change the oracle equation a little...

$$\tilde{\theta} \quad \text{solves} \quad \mathbb{E}_N[\psi(Z;\theta,\eta_1^*(Z,\boxed{\theta^*}),\eta_2^*(Z))] = 0$$

Invariant Jacobian Assumption

- What if we change the oracle equation a little...

$$\tilde{\theta} \quad \text{solves} \quad \mathbb{E}_N[\psi(Z;\theta,\eta_1^*(Z,\boxed{\theta^*}),\eta_2^*(Z))] = 0$$

$$\sqrt{N}(\tilde{\theta}-\theta^*) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} J^{\diamond -1}\psi(Z_i;\theta^*,\eta_1^*(Z_i,\theta^*),\eta_2^*(Z_i)) + o_{\mathbb{P}}(1)$$

$$J^{\diamond} = \partial_{\theta^\top}\mathbb{E}\left[\psi(Z;\theta,\eta_1^*(Z,\boxed{\theta^*}),\eta_2^*(Z))\right]|_{\theta=\theta^*}$$

Invariant Jacobian Assumption

💡 What if we change the oracle equation a little...

$$\tilde{\theta} \quad \text{solves} \quad \mathbb{E}_N[\psi(Z;\theta,\eta_1^*(Z,\theta^*),\eta_2^*(Z))] = 0$$

$$\sqrt{N}(\tilde{\theta}-\theta^*) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} J^{\diamond -1}\psi(Z_i;\theta^*,\eta_1^*(Z_i,\theta^*),\eta_2^*(Z_i)) + o_{\mathbb{P}}(1)$$

$$J^{\diamond} = \partial_{\theta^\top}\mathbb{E}\left[\psi(Z;\theta,\eta_1^*(Z,\theta^*),\eta_2^*(Z))\right]|_{\theta=\theta^*}$$

Assumption (Invariant Jacobian)

$$J^{\diamond} = J^*$$

Fréchet Orthogonality is Sufficient

Proposition (Fréchet Orthogonality)

Assume the map $(\theta, \eta_1(\cdot, \theta')) \mapsto \mathbb{E}\left[\psi(Z; \theta, \eta_1(Z, \theta'), \eta_2^*(Z))\right]$ *is Fréchet differentiable at* $(\theta^*, \eta_1^*(\cdot, \theta^*))$ *and exists* $C > 0$ *such that in a small neighborhood of* $(\theta^*, \eta_1^*(\cdot, \theta^*))$*:*

$$\mathcal{D}_{\eta_1}\mathbb{E}\left[\psi(Z; \theta, \eta_1^*(Z, \theta^*), \eta_2^*(Z))\right]\left[\eta_1'(\cdot, \theta') - \eta_1^*(\cdot, \theta^*)\right] = 0,$$

$$\mathbb{E}[\|\eta_1^*(Z, \theta') - \eta_1^*(Z, \theta^*)\|^2]^{1/2} \leq C\|\theta' - \theta^*\|.$$

Then ***Invariant Jacobian Assumption*** *is satisfied.*

- Essentially: Fréchet version of the η_1 part of the Neyman orthogonality condition (which uses Gâteaux derivative)
  - Holds for *all* of our examples: thanks to double robustness $\mathbb{E}\left[\psi(Z; \theta, \eta_1(Z, \theta'), \eta_2^*(Z))\right]$ does not even depend on η_1!

LDML

- Let $\mathcal{D}_1, \ldots, \mathcal{D}_K$ be a random even K-fold split of the data
- For $k = 1, \ldots, K$:

LDML

- Let $\mathcal{D}_1, \ldots, \mathcal{D}_K$ be a random even K-fold split of the data
- For $k = 1, \ldots, K$:
  - Split $\mathcal{D}_k^C$ into two halves $\mathcal{D}_k^{C,1}, \mathcal{D}_k^{C,2}$

LDML

- Let $\mathcal{D}_1, \ldots, \mathcal{D}_K$ be a random even K-fold split of the data
- For $k = 1, \ldots, K$:
  - Split $\mathcal{D}_k^C$ into two halves $\mathcal{D}_k^{C,1}, \mathcal{D}_k^{C,2}$
  - Use $\mathcal{D}_k^{C,1}$ to construct an initial estimator $\hat{\theta}_{\text{init}}^{(k)}$ of θ (*E.g.*, in certain problems like QTE can use IPW)

LDML

- Let $\mathcal{D}_1, \ldots, \mathcal{D}_K$ be a random even K-fold split of the data
- For $k = 1, \ldots, K$:
  - Split $\mathcal{D}_k^C$ into two halves $\mathcal{D}_k^{C,1}, \mathcal{D}_k^{C,2}$
  - Use $\mathcal{D}_k^{C,1}$ to construct an initial estimator $\hat{\theta}_{\text{init}}^{(k)}$ of θ (*E.g.*, in certain problems like QTE can use IPW)
  - Use $\mathcal{D}_k^{C,2}$ to construct estimator $\hat{\eta}_1^{(k)}(\cdot; \hat{\theta}_{\text{init}}^{(k)})$ of $\eta_1(\cdot; \hat{\theta}_{\text{init}}^{(k)})$ ($\hat{\theta}_{\text{init}}^{(k)}$ is fixed wrt $\mathcal{D}_k^{C,2}$ – estimating only a single nuisance)

LDML

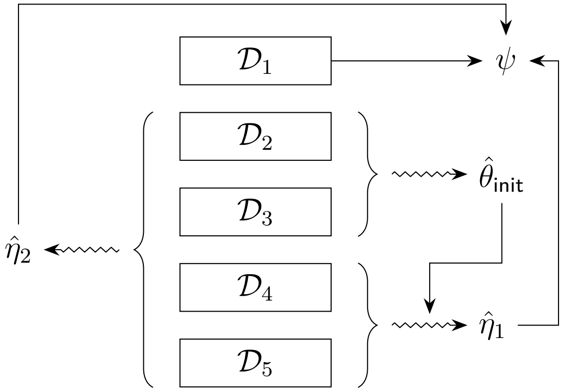
- Let $\mathcal{D}_1, \ldots, \mathcal{D}_K$ be a random even K-fold split of the data
- For $k = 1, \ldots, K$:
  - Split $\mathcal{D}_k^C$ into two halves $\mathcal{D}_k^{C,1}, \mathcal{D}_k^{C,2}$
  - Use $\mathcal{D}_k^{C,1}$ to construct an initial estimator $\hat{\theta}_{\text{init}}^{(k)}$ of θ (*E.g.*, in certain problems like QTE can use IPW)
  - Use $\mathcal{D}_k^{C,2}$ to construct estimator $\hat{\eta}_1^{(k)}(\cdot; \hat{\theta}_{\text{init}}^{(k)})$ of $\eta_1(\cdot; \hat{\theta}_{\text{init}}^{(k)})$ ($\hat{\theta}_{\text{init}}^{(k)}$ is fixed wrt $\mathcal{D}_k^{C,2}$ – estimating only a single nuisance)
  - Use $\mathcal{D}_k^C$ to construct estimator $\hat{\eta}_2^{(k)}$ of η_2

LDML

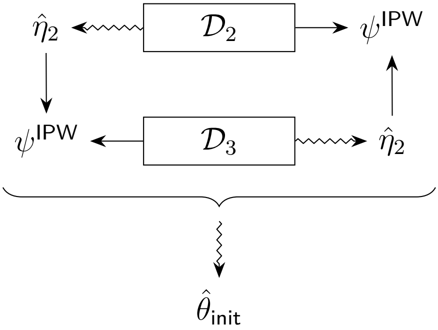
- Let $\mathcal{D}_1, \ldots, \mathcal{D}_K$ be a random even K-fold split of the data
- For $k = 1, \ldots, K$:
  - Split $\mathcal{D}_k^C$ into two halves $\mathcal{D}_k^{C,1}, \mathcal{D}_k^{C,2}$
  - Use $\mathcal{D}_k^{C,1}$ to construct an initial estimator $\hat{\theta}_{\text{init}}^{(k)}$ of θ (*E.g.*, in certain problems like QTE can use IPW)
  - Use $\mathcal{D}_k^{C,2}$ to construct estimator $\hat{\eta}_1^{(k)}(\cdot; \hat{\theta}_{\text{init}}^{(k)})$ of $\eta_1(\cdot; \hat{\theta}_{\text{init}}^{(k)})$ ($\hat{\theta}_{\text{init}}^{(k)}$ is fixed wrt $\mathcal{D}_k^{C,2}$ – estimating only a single nuisance)
  - Use $\mathcal{D}_k^C$ to construct estimator $\hat{\eta}_2^{(k)}$ of η_2
- Let $\hat{\theta}$ solve (within $o(N^{-1/2})$ error)

$$\min_{\theta \in \Theta} \left\| \frac{1}{N} \sum_{k=1}^K \sum_{i \in \mathcal{D}_k} \psi(Z_i; \theta, \hat{\eta}_1^{(k)}(Z_i, \hat{\theta}_{\text{init}}^{(k)}), \hat{\eta}_2^{(k)}(Z_i)) \right\|^2$$

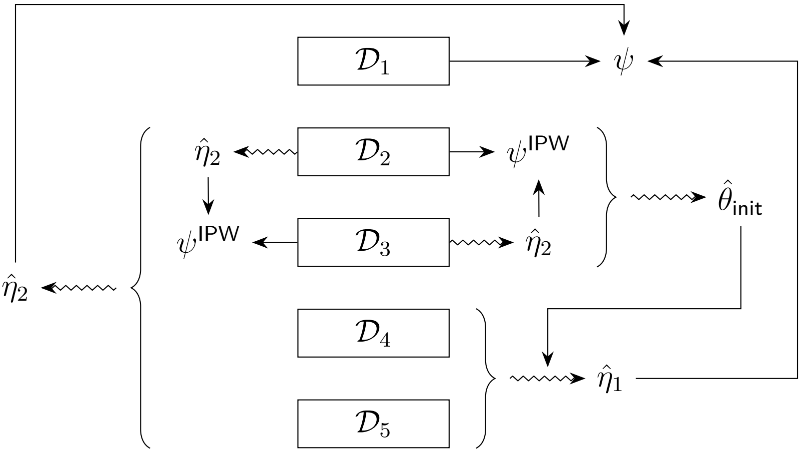
LDML Schematic Overview



LDML Schematic Overview



LDML Schematic Overview



LDML Variance Estimator and Inference

- Given an estimator $\hat{J}$ of J^*, set

$$\widehat{\Sigma} = \frac{1}{N} \sum_{k=1}^{K} \sum_{i \in \mathcal{D}_k} \hat{J}^{-1} \psi\psi^\top (Z_i; \hat{\theta}, \hat{\eta}_1^{(k)}(Z_i, \hat{\theta}_{\text{init}}^{(k)}), \hat{\eta}_2^{(k)}(Z_i)) \hat{J}^{-\top}$$

LDML Variance Estimator and Inference

- Given an estimator $\hat{J}$ of J^*, set

$$\widehat{\Sigma} = \frac{1}{N} \sum_{k=1}^{K} \sum_{i \in \mathcal{D}_k} \hat{J}^{-1} \psi \psi^\top (Z_i; \hat{\theta}, \hat{\eta}_1^{(k)}(Z_i, \hat{\theta}_{\text{init}}^{(k)}), \hat{\eta}_2^{(k)}(Z_i)) \hat{J}^{-\top}$$

- Given contrasts $\zeta \in \mathbb{R}^d$ construct $(1 - \alpha)$ confidence interval

$$\mathrm{CI}_\alpha = \left[\zeta^\top \hat{\theta} \pm \Phi^{-1}(1 - \alpha/2) \sqrt{\zeta^\top \hat{\Sigma} \zeta / N} \right]$$

Average Over the Splitting

- One run of LDML with one random split is enough to ensure our desirable asymptotics
  - Nonetheless the random splitting is just unnecessary noise
  - In theory, would prefer to just average over *all* splits
- Practically, to protect against outliers:
  - Run many iterations of LDML
  - Take a median / winsorized mean of $\hat{\theta}$ and $\widehat{\Sigma}$
  - Can also add the standard error of averaging $\hat{\theta}$ over iterations (but this is $o(N^{-1/2})$)

This talk

1 Introduction

2 Method

3 Asymptotic Guarantees

4 Empirical Results

5 Conclusions

Recall: LDML Estimator for QTE

- Efficient estimation equation $\psi(Z;\theta,\mu(X;\theta),\pi(X))$:

$$\frac{\mathbb{I}\left[T=1\right]}{\pi(X)}\left(\mathbb{I}\left[Y\le\theta\right]-\mu(X;\theta)\right)+\mu(X;\theta)-\gamma.$$

- γ-quantile of $Y(1)$ solves

$$\begin{aligned}&\mathbb{E}\left[\psi(Z;\theta,\mu^*(X;\theta^*),\pi^*(X))\right]=\mathbb{P}\left(Y(1)\le\theta^*\right)-\gamma=0,\\ &\text{where}\quad \pi^*(X)=\mathbb{P}(T=1\mid X)\ge\varepsilon>0,\\ &\qquad\qquad \mu^*(X;\theta^*)=\mathbb{P}\left(Y\le\theta^*\mid T=1,X\right).\end{aligned}$$

Recall: LDML Estimator for QTE

- Efficient estimation equation $\psi(Z;\theta,\mu(X;\theta),\pi(X))$:

$$\frac{\mathbb{I}\left[T=1\right]}{\pi(X)}\left(\mathbb{I}\left[Y\leq\theta\right]-\mu(X;\theta)\right)+\mu(X;\theta)-\gamma.$$

- γ-quantile of $Y(1)$ solves

$$\mathbb{E}\left[\psi(Z;\theta,\mu^*(X;\theta^*),\pi^*(X))\right]=\mathbb{P}\left(Y(1)\leq\theta^*\right)-\gamma=0,$$

where $\pi^*(X)=\mathbb{P}(T=1\mid X)\geq\varepsilon>0,$

$$\mu^*(X;\theta^*)=\mathbb{P}\left(Y\leq\theta^*\mid T=1,X\right).$$

- LDML estimator $\hat{\theta}$ solves

$$\frac{1}{N}\sum_{k=1}^{K}\sum_{i\in\mathcal{D}_k}\psi(Z;\theta,\hat{\mu}^{(k)}(X;\hat{\theta}_{\text{init}}),\hat{\pi}^{(k)}(X))=0$$

Assumptions: Nuisance Estimation

Assumption (Nuisance Estimation)

With probability $1 - o(1)$, for $k = 1, \ldots, K$

$$\begin{aligned}
\left\| \hat{\mu}^{(k)}(1, X; \hat{\theta}^{(k)}_{\text{init}}) - \mu^*(1, X; \hat{\theta}^{(k)}_{\text{init}}) \right\|_2 &\leq \rho_{\mu,N}, \\
\|\hat{\pi}^{(k)}(X) - \pi^*(X)\|_2 &\leq \rho_{\pi,N}, \\
|\hat{\theta}^{(k)}_{\text{init}} - \theta^*| &\leq \rho_{\theta,N}, \\
\|1/\hat{\pi}^{(k)}(X)\|_\infty &\leq 1/\varepsilon.
\end{aligned}$$

Assumptions: Distribution Regularity

Assumption (Distribution Regularity)

$\theta^* \in \text{int}(\Theta)$ for a compact Θ. For any $\theta \in \Theta$:

1. $F_1(\theta) = \mathbb{P}(Y(1) \leq \theta)$ is twice differentiable with
$$0 < c \leq F_1'(\theta) \leq C, \quad |F_1''(\theta)| \leq C.$$
2. $F_1(\theta \mid X) = \mathbb{P}(Y(1) \leq \theta \mid X)$ is almost surely twice differentiable with
$$F_1'(\theta \mid X) \leq C, \quad |F_1''(\theta \mid X)| \leq C \text{ almost surely.}$$

Asymptotic Normality

Theorem (Asymptotic Normality of LDML Quantile Estimator)

Under previous assumptions, if further

$$\rho_{\pi,N} = o(1), \ \ \rho_{\mu,N} = o(1), \ \ \rho_{\theta,N} = o(1)$$
$$\rho_{\pi,N}(\rho_{\mu,N} + \rho_{\theta,N}) = o(N^{-1/2}),$$

then for $J^* = F_1'(\theta^*)$,

$$\sqrt{N}(\hat{\theta} - \theta^*) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{1}{J^*} \psi(Z_i; \theta^*, \mu^*(X_i; \theta^*), \pi^*(X_i)) + o_P(1)$$
$$\rightsquigarrow \mathcal{N}\left(0, \frac{1}{J^{*2}} \mathbb{E}\left[\psi^2(Z_i; \theta^*, \mu^*(X_i; \theta^*), \pi^*(X_i))\right]\right)$$

Remarks

- Stronger uniform convergence results in our paper.

Remarks

- Stronger uniform convergence results in our paper.
- New analysis to enable $\rho_{\pi,N}(\rho_{\mu,N} + \rho_{\theta,N}) = o(N^{-1/2})$.

Remarks

- Stronger uniform convergence results in our paper.
- New analysis to enable $\rho_{\pi,N}(\rho_{\mu,N} + \rho_{\theta,N}) = o(N^{-1/2})$.
- Corollary: LDML QTE estimator is also asymptotically linear, asymptotically normal and semiparametrically efficient.

Remarks

- Stronger uniform convergence results in our paper.
- New analysis to enable $\rho_{\pi,N}(\rho_{\mu,N} + \rho_{\theta,N}) = o(N^{-1/2})$.
- Corollary: LDML QTE estimator is also asymptotically linear, asymptotically normal and semiparametrically efficient.
- IPW Initial estimator $\hat{\theta}_{\text{init}}$ has $\rho_{\theta,N} = O(\rho_{\pi,N})$:

$$\rho_{\pi,N}\rho_{\theta,N} = o(N^{-1/2}) \implies \rho_{\pi,N} = o(N^{-1/4}).$$

Remarks

- Stronger uniform convergence results in our paper.
- New analysis to enable $\rho_{\pi,N}(\rho_{\mu,N} + \rho_{\theta,N}) = o(N^{-1/2})$.
- Corollary: LDML QTE estimator is also asymptotically linear, asymptotically normal and semiparametrically efficient.
- IPW Initial estimator $\hat{\theta}_{\text{init}}$ has $\rho_{\theta,N} = O(\rho_{\pi,N})$:

$$\rho_{\pi,N}\rho_{\theta,N} = o(N^{-1/2}) \implies \rho_{\pi,N} = o(N^{-1/4}).$$

- IPW kernel estimator for $J^* = F_1'(\theta^*)$:

$$\hat{J} = \frac{1}{Nh}\sum_{k=1}^{K}\sum_{i\in\mathcal{D}_k}\frac{\mathbb{I}\left[T_i = 1\right]}{\hat{\pi}^{(k)}(X_i)}\kappa((Y_i - \hat{\theta})/h).$$

For stability, divide $\hat{J}$ by $\frac{1}{N}\sum_{k=1}^{K}\sum_{i\in\mathcal{D}_k}\frac{\mathbb{I}[T_i=1]}{\hat{\pi}^{(k)}(1|X_i)}$.

Inference

- Plug-in variance estimator:

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{k=1}^{K} \sum_{i \in \mathcal{D}_k} \frac{1}{\hat{J}^2} \psi^2(Z_i; \hat{\theta}, \hat{\mu}^{(k)}(X_i; \hat{\theta}^{(k)}_{\textsf{init}}), \hat{\pi}^{(k)}(X_i)).$$

Inference

- Plug-in variance estimator:

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{k=1}^{K} \sum_{i \in \mathcal{D}_k} \frac{1}{\hat{J}^2} \psi^2(Z_i; \hat{\theta}, \hat{\mu}^{(k)}(X_i; \hat{\theta}_{\text{init}}^{(k)}), \hat{\pi}^{(k)}(X_i)).$$

- $(1-\alpha) \times 100\%$ Confidence interval:

$$\mathrm{CI} := [\hat{\theta} \pm \Phi^{-1}(1-\alpha/2)\hat{\sigma}/\sqrt{N}]$$

Inference

- Plug-in variance estimator:

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{k=1}^{K}\sum_{i\in\mathcal{D}_k}\frac{1}{\hat{J}^2}\psi^2(Z_i;\hat{\theta},\hat{\mu}^{(k)}(X_i;\hat{\theta}^{(k)}_{\text{init}}),\hat{\pi}^{(k)}(X_i)).$$

- $(1-\alpha)\times 100\%$ Confidence interval:

$$\mathrm{CI} := [\hat{\theta} \pm \Phi^{-1}(1-\alpha/2)\hat{\sigma}/\sqrt{N}]$$

Theorem (Confidence Interval)

Under all previous assumptions, if further $\hat{J}$ is consistent for J^*, then

$$\mathbb{P}(\theta^* \in \mathrm{CI}) \to (1-\alpha),\ \textit{as } n \to \infty.$$

IV and Local Quantiles

- What if the unconfoundedness no longer holds?

$$Y(t) \not\perp T \mid X.$$

One common solution: instrumental variable (IV).

IV and Local Quantiles

- What if the unconfoundedness no longer holds?

$$Y(t) \not\perp T \mid X.$$

  One common solution: instrumental variable (IV).
- $Z = (X, W, T, Y)$ with an IV $W \in \{0, 1\}$ (Angrist, Imbens, Rubin 1996).
  - Exclusion restriction: $Y(t) = Y(t, w) = Y(t, 1 - w)$.
  - Exogeneity $(Y(t), T(w)) \perp W \mid X$.
  - Monotonicity: $T(1) \geq T(0)$;
  - Relevance: $\nu^* = \mathbb{P}(T(1) = 1) - \mathbb{P}(T(0) = 1) > 0$.
  - Overlap: $0 < \tilde{\pi}^*(X) = \mathbb{P}(W = 1 \mid X) < 1$.

IV and Local Quantiles

- What if the unconfoundedness no longer holds?

$$Y(t) \not\perp T \mid X.$$

  One common solution: instrumental variable (IV).
- $Z = (X, W, T, Y)$ with an IV $W \in \{0, 1\}$ (Angrist, Imbens, Rubin 1996).
  - Exclusion restriction: $Y(t) = Y(t, w) = Y(t, 1 - w)$.
  - Exogeneity $(Y(t), T(w)) \perp W \mid X$.
  - Monotonicity: $T(1) \geq T(0)$;
  - Relevance: $\nu^* = \mathbb{P}(T(1) = 1) - \mathbb{P}(T(0) = 1) > 0$.
  - Overlap: $0 < \tilde{\pi}^*(X) = \mathbb{P}(W = 1 \mid X) < 1$.
- Goal: local $\gamma-$ quantile θ^* that solves

$$\mathbb{P}(Y(1) \leq \theta \mid T(1) > T(0)) - \gamma = 0.$$

Efficient Estimation Equation for Local Quantiles

- Efficient estimation equation (Belloni et al. 2017)

$$
\begin{aligned}
&\psi(Z;\theta,\tilde{\mu}(1,X;\theta),\tilde{\mu}(0,X;\theta),\tilde{\pi}(X),\nu) \\
=&\frac{1}{\nu}\Bigg(\tilde{\mu}(1,X;\theta)+\frac{W}{\tilde{\pi}(X)}\left(\mathbb{I}\left[T=1,Y\leq\theta\right]-\tilde{\mu}(1,X;\theta)\right) \\
-&\ \tilde{\mu}(0,X;\theta)-\frac{1-W}{1-\tilde{\pi}(X)}\left(\mathbb{I}\left[T=1,Y\leq\theta\right]-\tilde{\mu}(0,X;\theta)\right)\Bigg)-\gamma.
\end{aligned}
$$

- Local $\gamma-$quantile θ^* solves

$$
\begin{aligned}
&\mathbb{E}\left[\psi(Z;\theta,\tilde{\mu}^*(1,X;\theta^*),\tilde{\mu}^*(0,X;\theta^*),\tilde{\pi}^*(X),\nu^*)\right] \\
&\qquad\qquad=\mathbb{P}\left(Y(1)\leq\theta \mid T(1)>T(0)\right)-\gamma=0,
\end{aligned}
$$

where $\tilde{\mu}^*(w,X;\theta^*)=\mathbb{P}\left(T=1,Y\leq\theta^* \mid W=w,X\right)$,

$$
\begin{aligned}
\tilde{\pi}^*(X)&=\mathbb{P}\left(W=1 \mid X\right), \\
\nu^*&=\mathbb{E}\left[\mathbb{P}\left(T=1 \mid X,W=1\right)-\mathbb{P}\left(T=1 \mid X,W=0\right)\right].
\end{aligned}
$$

LDML Estimator for Local Quantiles

LDML estimator $\hat{\theta}$ solves

$$\frac{1}{N}\sum_{k=1}^{K}\sum_{i\in\mathcal{D}_k}\psi(Z;\theta,\hat{\tilde{\mu}}^{(k)}(1,X;\hat{\theta}_{\text{init}}),\hat{\tilde{\mu}}^{(k)}(0,X;\hat{\theta}_{\text{init}}),\hat{\tilde{\pi}}^{(k)}(X),\hat{\nu})=0.$$

- $\hat{\nu}$ only needs to be consistent for theoretical guarantees.
- Asymptotic normality and inferential results analogously hold.

General Theory

General theory for LDML estimation and inference with

$$\mathbb{E}\left[\psi(Z;\theta,\eta_1^*(Z,\theta_1),\eta_2^*(Z))\right]=0$$

under Neyman-orthogonality condition and generic rate conditions for nuisance estimation.

General Theory

General theory for LDML estimation and inference with

$$\mathbb{E}\left[\psi(Z;\theta,\eta_1^*(Z,\theta_1),\eta_2^*(Z))\right]=0$$

under Neyman-orthogonality condition and generic rate conditions for nuisance estimation.

- Unconfoundedness setting:

$$\mathbb{E}\left[U(Y(t);\theta_1)+V(\theta_2)\right]=0, t=0,1.$$

General Theory

General theory for LDML estimation and inference with

$$\mathbb{E}\left[\psi(Z;\theta,\eta_1^*(Z,\theta_1),\eta_2^*(Z))\right]=0$$

under Neyman-orthogonality condition and generic rate conditions for nuisance estimation.

- Unconfoundedness setting:

$$\mathbb{E}\left[U(Y(t);\theta_1)+V(\theta_2)\right]=0, t=0,1.$$

- IV setting:

$$\mathbb{E}\left[U(Y(t);\theta_1)+V(\theta_2)\mid T(1)>T(0)\right]=0, t=0,1.$$

This talk

1 Introduction

2 Method

3 Asymptotic Guarantees

4 Empirical Results

5 Conclusions

The Effect of 401(k) Participation On Net Assets

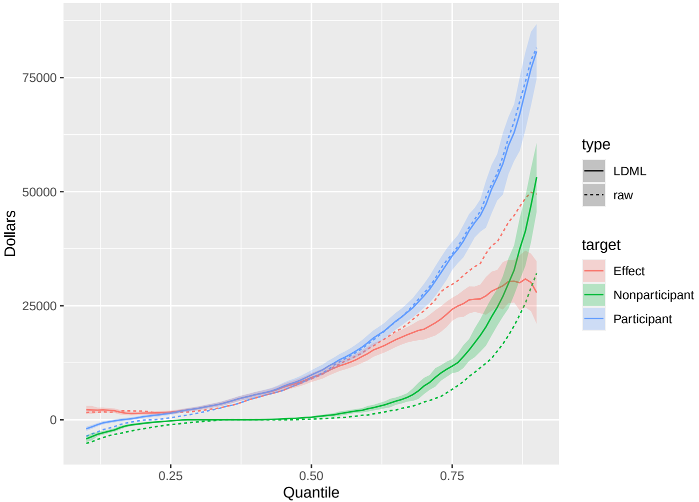
- Wealth is generally a very skewed distribution
  - Quantiles potentially more informative than averages
- 401(k) participation confounded with wealth
  - Might be instrumented using 401(k) eligibility
  - Eligibility also non-random but may be ignorable given age, income, education, family size, marital status, ...
- Chernozhukov and Hansen (2004) consider lo-dim linear spec
  - Belloni et al. (2017) include high-order terms + interactions but use many LASSOs in a discretized grid of θ's; asymptotic results may not apply to generic black-box methods
  - Chernozhukov et al. (2018) use generic black-box methods but only tackle *averages* (linear estimating equation)
- In contrast: we will use LDML to conduct inference on (L)QTEs using a variety of black-box regression methods

LQTE: 401(k) Participation

γ	K	LASSO	Neural Net	Boosting	Raw
25%	5	1.74 (0.23)	1.77 (0.26)	1.53 (0.25)	4.18 (0.37)
	15	1.70 (0.22)	1.81 (0.27)	1.46 (0.24)	
	25	1.68 (0.22)	1.94 (0.27)	1.41 (0.24)	
50%	5	8.93 (0.60)	8.93 (0.66)	7.59 (0.59)	15.05 (0.67)
	15	9.27 (0.60)	8.51 (0.67)	7.65 (0.57)	
	25	9.42 (0.61)	8.65 (0.67)	7.63 (0.56)	
75%	5	23.11 (1.71)	26.74 (1.93)	20.71 (2.04)	38.59 (1.71)
	15	24.20 (1.53)	24.48 (2.04)	20.91 (1.98)	
	25	24.73 (1.48)	25.77 (2.05)	20.99 (1.96)	

In \$1k

LQTE: 401(k) Participation

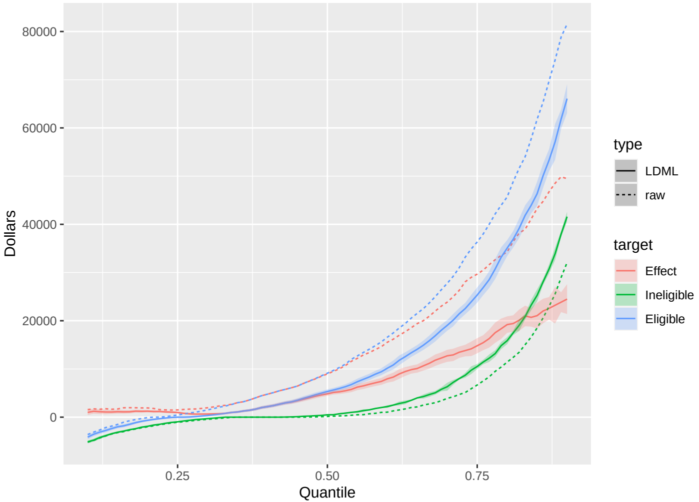


QTE: 401(k) Eligibility

γ	K	LASSO	Neural Net	Boosting	Raw
	5	0.95 (0.14)	1.01 (0.12)	1.00 (0.11)	
25%	15	0.96 (0.12)	1.02 (0.13)	0.99 (0.11)	1.50 (0.25)
	25	0.97 (0.12)	1.04 (0.13)	0.99 (0.10)	
	5	4.79 (0.26)	4.94 (0.30)	4.45 (0.28)	
50%	15	4.83 (0.26)	5.01 (0.32)	4.42 (0.28)	9.98 (4.15)
	25	4.87 (0.26)	5.14 (0.32)	4.42 (0.28)	
	5	14.49 (0.95)	15.40 (1.03)	13.39 (0.95)	
75%	15	14.81 (0.96)	15.23 (1.06)	13.39 (0.96)	29.67 (1.35)
	25	14.88 (0.96)	15.19 (1.06)	13.44 (0.95)	

In $1k

QTE: 401(k) Eligibility



This talk

1 Introduction

2 Method

3 Asymptotic Guarantees

4 Empirical Results

5 Conclusions

Conclusions

- (L)QTEs are important in empirical studies with skewed distributions and/or where important to understand risk
  - But difficult to assess in high-dimensional/complex settings
  - SotA DML requires we estimate a *continuum* of nuisances
- Instead we proposed *Localized* DML
  - Localized the nuisance estimation to a single point using a rough initial guess
  - Asymptotically behaves like oracle estimation equation under lax conditions that allow using black-box regression methods
- More generally relevant with any nonlinear orthogonal estimating equation with estimand-dependent nuisances
  - Just need a slightly strong Fréchet-derivative orthogonality

Thank you!