Reading Group: Causal Isotonic Regression

Ted Westling, Peter Gilbert, Marco Carone (JRSSB, 2020)

Agenda

Terminologies

There are some common terminologies to be introduced.

Terminoloies

1. **Observational Studies:** Conduct inference based on the data set we observed, i.e. we did not make any control on the experiment.
2. **Experimental Design:** Randomly assigning the values of the exposure and observe the corresponding response.
3. **Association:** X is positively associated with Y iff X increases together with Y.
4. **Causation:** Increase in X results in increase in Y.

In general, we cannot draw causation conclusion under framework of observation studies due to the existence of confounders.
AIM 1: Accessing causality effect by using observational data set.

Literature Review

1. Non-parametric methods with **binary/categorical exposure**
   1. Inverse-probability-weighted estimators
      (Horvitz and Thompson, 1952)
   2. Augmented inverse-probability-weighted (AIPW) estimators
      (Scharfstein et al., 1999; Bang and Robins, 2005)
   3. Targeted minimum-loss-based estimators
      (van der Laan and Rose, 2011)
2. **Continuous exposure**
   Common approach: Discretize the continuous interval into two or more region and apply above methods under categorial case.

Remark: Undesirable method as we are commonly interested in the causal dose response curver, which describes the causal relationship between the exposure and outcome across a **continuum** of the exposure

Literature Review

We further consider existing method dealing with continuous exposures.

1. Applying parametric models *(Robins, 2000 and Zhang et al., 2016)*
2. Inference on parameters obtained by projecting a causal dose-response curve onto a parametric working model. *(Neugebauer and van der Laan, 2007)*
3. nonparametric estimation using flexible data-adaptive algorithms *(Rubin and van der Laan, 2006 and Diaz and van der Laan, 2011)*
4. estimator based on local linear smoothing *(Kennedy et al, 2017)*
5. general framework for inference on parameters that fail to be sufficiently smooth as a function of the data-generating distribution and for which regular root-n-estimation theory is therefore not available. *(van der Laan et al., 2018)*

Comment on existing methodologies

For the literature on non-parametric estimation of causal dose-response curves, the large sample inference is not valid and is sensitive to selection of certain tuning parameters.
Smoothing-based methods are sensitive to choice of kernel function and bandwidth. The estimators commonly have non-negligible bias.

Motivation for isotonic regression

In many setting, the causal dose-response curve is a monotone function For example, exposures such as daily exercise performed, cigarettes smoked per week and air pollutant levels are all known to have monotone relationships with various health outcomes. Recall in the setting of linear regression, we are interested in estimating the function g s.t.

$$\mathbb{E}(Y|X_1, \cdots, X_p) = g(X_1, \cdots, X_p).$$

For linear regression, we assumed linearity on the model, i.e. restrict the class of estimating cuver for g to class of the linear function $\mathbb{E}(Y|X_1, \cdots, X_p) = g(X_1, \cdots, X_p) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$ for some $\beta_0, \beta_1, \cdots, \beta_p$. However, linearity is usually too strong as the model assumption. In contrast, we add only trivial restriction on the class of function under non-parametric function for estimation g.

Motivation for isotonic regression

However, as mentioned, the estimator in non-parametric methods might be easily affected by choice of kernel and bandwidth. And as an intermediate level of assumption, we may assume g to be a monotone function. Notice that linearity implies monotonicity but the converse does not holds. The regression with monotonicity assumption is known as **Isotonic Regression**.
The following are some advantages of isotonic regression

1. Does not require linearity assumption.
2. Does not require selection of kernel and bandwidth.
3. Invariance to strictly increasing transformation of exposure and on centring and scaling by factor of $n^{-1/3}$.

If there is not confounding variables, we can directly estimate the causal isotonic curve. However as mentioned, the authur attempted access causality under confounded setting in this paper.

Notation

1. Y: Response.
2. A: Continuous exposure.
3. W: Vector of covariates.
4. $O \triangleq (Y, A, W)$: Data unit.
5. P_0: The true data-generating distribution
6. $\mathcal{O} = \mathcal{Y} \times \mathcal{A} \times \mathcal{W}$: Support of P_0, where $\mathcal{Y}, \mathcal{A} \subseteq \mathbb{R}$ are intervals and $\mathcal{W} \subseteq \mathbb{R}^p$.

Parameter of interest

Denote E_0 as the expectation under the law P_0. Our parameter of interest, known as **G-computed regression function** from $\mathcal{A}$ to $\mathbb{E}$ is defined as

$$a \mapsto \theta_0(a) \triangleq \mathbb{E}_0\{\mathbb{E}_0(Y|A=a,W)\}$$

, which has causation interpretation in some scientific contexts. It kind of measure the **effect of** A **on** Y **after accounted for** W.

Causal parameter and unadjusted regression function

Adopting the idea of randomnized experiment, for each fixed $a \in \mathcal{A}$, we denote by $Y(a)$ the potential outcome (i.e. a R.V.) under exposure level $A = a$. The causal parameter is thus defined as $m_0 : \mathcal{A} \to \mathbb{R}$ through

$$m_0(a) \triangleq \mathbb{E}_0\{Y(a)\}$$

annd the resulting curve is known as ***causal dose-response curve***.
We can also denote the **unadjusted** regression function as $r_0(a) \triangleq \mathbb{E}_0(Y|A = a)$, which is the standard quantity of interest in regression analysis.
Under different set of conditions, we can claim that $m_0(a) = r_0(a)$ and $m_0(a) = \theta_0(a)$, meaning that those causal conditions allows us to draw causation conclusion from observational data set.

Comparison among quantities

1. Causal dose-response curve: $m_0(a) \triangleq \mathbb{E}_0\{Y(a)\}$ measures the intrinsic casual effect of A on Y.
2. G computed regression function: $\theta_0(a) \triangleq \mathbb{E}_0\{\mathbb{E}_0(Y|A=a,W)\}$ measures the effect of A on Y after accounted for covariates W
3. Unadjusted regression function: $r_0(a) \triangleq \mathbb{E}_0(Y|A=a)$ measures the association between A on Y

Therefore, the strength of causation statement can be made through these functions are different. Even though we may not be able to contain all covariates in W, it gives more informative conclusion than only relying on $r_0(a)$.

Causal condition: $m_0(a) = r_0(a)$

Causal condition 1

Suppose that for some $a \in \mathcal{A}$,

1. Each unit's potential outcomes are independent of all other units' exposures.
2. $Y = Y(A)$, where Y is the response and $Y(A)$ is defined as in setting of randomnized experiment.
3. A and $Y(a)$ are independent.
4. The marginal density of A is positive at a.

then for such $a \in \mathcal{A}$, we have $m_0(a) = r_0(a)$, i.e. the causal effect for exposure level $A = a$.

Remark: Condition (3) is commonly only satisfied in randomnized trials as usually ther exist some confounders affecting both A and $Y(a)$, which induce dependency.

Causal condition: $m_0(a) = \theta_0(a)$

Condition (3) and (4) are being too strong to be practical, it is natural to find sufficient condition s.t. inference of $\theta_0(a)$ give meaningful result.

Causal condition 2

Suppose that for some $a \in \mathcal{A}$,

1. Each unit's potential outcomes are independent of all other units' exposures.
2. $Y = Y(A)$, where Y is the response and $Y(A)$ is defined as in setting of randomnized experiment.
3. A and $Y(a)$ are **conditionally** independent given W.
4. The marginal density of A **given** $\boldsymbol{W}$ is a.s. positive at a

then for such $a \in \mathcal{A}$, we have $m_0(a) = \theta_0(a)$, i.e. the causal effect for exposure level $A = a$.

Remark: Whenever $m_0(a) = \theta_0(a)$, the conclusion drawn from inference can be interpreted as causal statement.

Notations

1. $F_P : \mathcal{A} \to \mathbb{R}$ the distribution function of A under P.
2. $\mathscr{F}_\theta$: class of non-decreasing real-valued functions on $\mathcal{A}$.
3. $\mathscr{F}_T$: class of strictly increasing and continuous distribution functions supported on $\mathcal{A}$.

The statistical model to work on is given by

$$\mathcal{M} \triangleq \{P : \theta_P \in \mathscr{F}_\theta, F_P \in \mathscr{F}_T\}$$

Recall our target: Making inference about

$$\theta_0(a) = \mathbb{E}_0\{E_0(Y|A = a, W)\}$$

for cts exposure A and monotone θ_0 by using **independent** (NOT randomnized) observations $O_1, \cdots, O_n$ drawn from $P_) \in \mathcal{M}$, which is an extension of classical isotonic regression in sense of allowing the existence of confounders.

Contribution of paper

Contributions

1. generalizes the unadjusted isotonic regression estimator to the more realistic scenario in which there is confounding by recorded covariates.
2. investigate finite sample and asymptotic properties of the estimator proposed, including invariance to strictly increasing transformations of the exposure, doubly robust consistency and doubly robust convergence in distribution to a non-degenerate limit.
3. derive practical methods for constructing pointwise confidence intervals, including intervals that have valid doubly robust calibration.
4. illustrate numerically the practical performance of the estimator.

Agenda

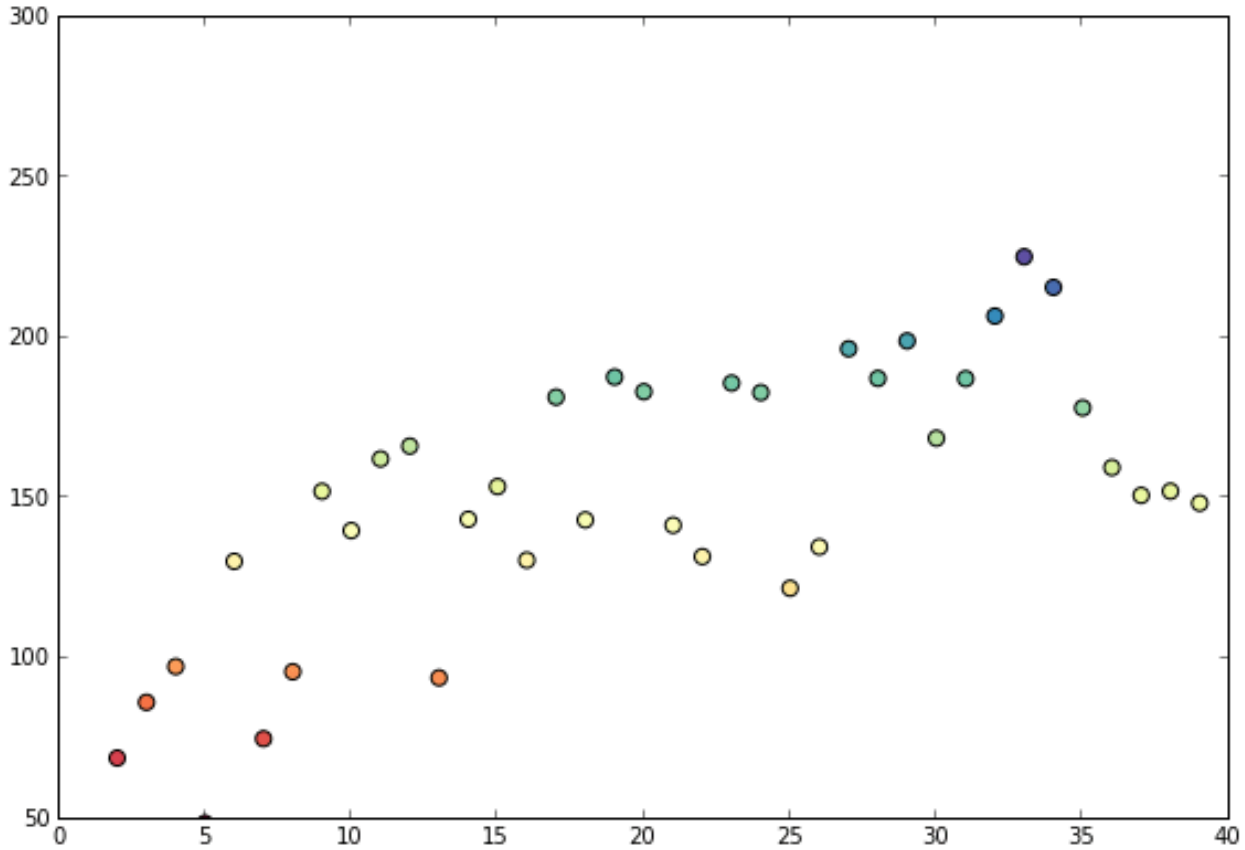
Classical least square isotonic regression

Linear regression: find $\beta = \hat{\beta}$ s.t. $\sum_{i=1}^{n} \left(Y_i - \beta A_i\right)^2$ is minimized

Isotonic regression: find $r = r_n$ s.t. $\sum_{i=1}^{n} \left[Y_i - r(A_i)\right]^2$ is minimized

- $Y_{1:n}$: responses
- $A_{1:n}$: continuous exposures
- r: any monotone non-decreasing function
- r_n can be obtained via pool adjacent violators algorithm (PAVA)
  - Not true without assuming piecewise linearity of r?
  - PAVA can be used to find best monotone fit $\hat{Y}_i$ only
- r_n can also be represented by greatest convex minorants (GCMs)
  - Probably because isotonic regression can be formulated as a convex programming problem
  - See section 2.3 of the R package *isotone*'s vignette

Pool adjacent violators algorithm



Source: Pedregosa, Fabian (2013)

Pool adjacent violators algorithm

Target: find best monotone fit $\hat{Y}_i$ of response Y_i

- Exposure A_i are ordered in i first, i.e. $A_1 \leq A_2 \leq \cdots \leq A_n$
- Response Y_i may not be monotone as the sorting is done on A_i
- Fit $\hat{Y}_i = r_n(A_i)$ must be monotone in i
  - That's why $r = r_n$ should be a monotone non-decreasing function
  - Identifiable without assuming piecewise linearity of r?

Algorithm (sketch):

1. Initialize $l := 0$, $B^{(0)} := n$, $\hat{Y}_r^{(0)} := Y_r$ for $r = 1, \ldots, n$
2. Merge $\hat{Y}^{(l)}$-values into blocks if $\hat{Y}_{r+1}^{(l)} < \hat{Y}_r^{(l)}$ for $r = 1, \ldots, B^{(l)}$
3. Minimize the loss function for each block r, which gives $\hat{Y}_r^{(l+1)}$
4. If $\hat{Y}_{r+1}^{(l)} < \hat{Y}_r^{(l)}$ for some r, set $l = l + 1$ and go back to step 2
5. Expand the block values w.r.t. to $i = 1, \ldots, n$

Greatest convex minorant

GCM of a function f bounded on $[a, b]$: supremum over all convex functions g such that $g \leq f$

Let F_n be the empirical distribution function of $A_{1:n}$. It can be shown that the isotonic regression estimator $r_n(a)$ is

- the left derivative, evaluated at $F_n(a)$,
- of the GCM over the interval $[0, 1]$ of the linear interpolation,
- of the cumulative sum diagram $\left\{ \frac{1}{n} \left[i, \sum_{j=0}^{i} Y^*_{(i)} \right] : i = 0, 1, \ldots, n \right\}$,
- where $Y^*_{(0)} := 0$ and $Y^*_{(i)}$ is the response Y sorted by value of exposure A

Attractive properties of isotonic regression estimator

No need to choose kernel, bandwidth or any other tuning parameter

- The monotone fit restriction is kind of like a kernel already
- but it is true that no choice of tuning parameter is needed

Invariant to strictly increasing transformations of A

Uniform consistency on any strict subinterval of A

Limit distribution available

- $n^{\frac{1}{3}}\left[r_n(a) - r_0(a)\right] \xrightarrow{d} \left[4r_0'(a)\sigma_0^2(a)/f_0(a)\right]^{\frac{1}{3}}\mathbb{W}$ for any interior point $a \in \mathcal{A}$ at which $r_0'(a)$, $f_0(a) := F_0'(a)$ and $\sigma_0^2(a) := E_0\left[\{Y - r_0(a)\}^2 | A = a\right]$ exist and are positive and continuous in a neighbourhood of a
- $\mathbb{W}$ follows Chernoff's distribution, which often appears in the limit distribution of monotonicity-constrained estimators

Definition of proposed estimator

Definition: pointwise outcome

$$\mu_P(a, w) := E_P(Y|A = a, W = w)$$

for any given $P \in \mathcal{M}$

Definition: normalized exposure density

$$g_P(a, w) := \pi_P(a|w)/f_P(a)$$

where $\pi_P(a|w)$ is the conditional density evaluated at a given $W = w$, f_P is the marginal density of A under P

Definition: pseudo-outcome

$$\xi_{\mu,g,Q}(y, a, w) := \frac{y - \mu(a, w)}{g(a, w)} + \int \mu(a, z)Q(dz)$$

Monotonicity of proposed estimator

Kennedy *et al.* (2017) used pseudo-outcome to develop local linear regression for inference of $\theta_0(a)$. In the setting of this paper, $\theta_0(a)$ is known to be monotone

- I think the monotonicity of $\theta_0(a)$ is an assumption
- Yet it seems reasonable as continuous treatment usually has monotone causal effect (if effective) within certain range
- Example: daily exercise time (0-2 hours) on life expectancy
- Counterexample: daily exercise time (0-12 hours)
- So the reasonability of monotonicity may depend on the range of treatment A (experiemental) or data exploration (observational)

Under monotonicity, it is natural to consider the isotonic regression of the pseudo-outcomes on $A_{1:n}$

Proposed estimation procedure

Estimation of $\theta_n(a)$

1. Construct estimators μ_n, g_n of μ_0, g_0 respectively
2. For each a in the unique values of $A_{1:n}$, compute and set

$$\Gamma_n(a) := \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,a]}(A_i)\frac{Y_i - \mu_n(A_i,W_i)}{g_n(A_i,W_i)} + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} I_{(-\infty,a]}(A_i)\mu_n(A_i,W_j) \quad (1)$$

3. Compute the GCM $\bar{\Psi}_n$ of the set of points $\{(0,0)\} \cup \{(F_n(A_i), \Gamma_n(A_i)) : i = 1, 2, \ldots, n\}$ over $[0,1]$
4. Define $\theta_n(a)$ as the left derivative of $\bar{\Psi}_n$ evaluated at $F_n(a)$

Asymptotic framework for the proposed estimator

As in Kennedy *et al.* (2017), $\theta_n(a)$ deviate from classical results

- Pseudo-outcomes $\xi_{\mu,g,Q}(y,a,w)$ are dependent because they depend on the estimator μ_n, g_n, Q_n estimated with all observations
- Hence classical results from isotonic regression do not apply
- However, θ_n is of generalized Grenander type
- Asymptotic results of Westling and Carone (2020) can be used

We skip the proof of θ_n to be Grenander type here, which is to show θ_n falls in the class of estimator discussed in Westling and Carone (2020)

Some remarks on monotonicity and generality

Monotonicity: if $\theta_0(a)$ were only known to be monotone on a fixed subinterval $\mathcal{A}_0 \subset \mathcal{A}$

- We discuss this assumption on page 24
- The estimation procedure is still valid by first defining $F_p(a) := P(A \leq a | A \in \mathcal{A}_0)$ and F_n as its empirical counterpart
- Then replace $I_{(-\infty,a]}(A_i)$ in equation 1 by $I_{(-\infty,a]\cap\mathcal{A}_0}(A_i)$

Generality: the proposed estimator θ_n generalizes the classical r_n

- Condition 1: $A \perp\!\!\!\perp W \implies g_0(a,w) = 1$, so we may take $g_n = 1$
- Condition 2: $Y|A \perp\!\!\!\perp W|A \implies \mu_n(a,w) = \mu_n(a)$
- Under these conditions, equation 1 becomes

$$\Gamma_n(a) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,a]}(A_i)Y_i - \mu_n(A_i)$$

- As a result, $\theta_n(a) = r_n(a)$ for each a

Agenda

Invariance to strictly increasing transform of exposure

$\theta_n(a)$ is invariant to strictly increasing transform $H(\cdot)$ of exposure A

- Intuition: composition preserve monotonicity
- Desirable property since scale of exposure is often arbitrary
- Example: temperature in degrees Fahrenheit or Celsius or in kelvins
- Change of scale does not affect available information

We skip the proof because the intuition is simple (composition of monotone functions is also monotone)

Condtions for consistency

Notations:

- $\mathcal{F}$: a uniformly bounded class of functions
- Q: a finite discrete probability measure
- $N\{\epsilon, \mathcal{F}, L_2(Q)\}$: the ϵ-covering-number, i.e. the smallest number of $L_2(Q)$ balls of radius less than or equal to ϵ needed to cover $\mathcal{F}$
- $\log\left[\sup_Q N\{\epsilon, \mathcal{F}, L_2(Q)\}\right]$: the uniform ϵ-entropy of $\mathcal{F}$

Condition 1

There exist constants $C, \delta, K_0, K_1, K_2 \in (0, \infty)$ and $V \in [0, 2)$ s.t., almost surely as $n \to \infty$, μ_n and g_n are contained in classes of functions $\mathcal{F}_0$ and $\mathcal{F}_1$ respectively, satisfying

1. $|\mu| \leq K_0, \forall \mu \in \mathcal{F}_0$, and $K_1 \leq g \leq K_2, \forall g \in \mathcal{F}_1$
2. $\log\left[\sup_Q N\{\epsilon, \mathcal{F}_0, L_2(Q)\}\right] \leq C\epsilon^{-V/2}$ and $\log\left[\sup_Q N\{\epsilon, \mathcal{F}_1, L_2(Q)\}\right] \leq C\epsilon^{-V}, \forall \epsilon \leq \delta$

Condtions for consistency

P_0: the true data-generating distribution but not projection

Condition 2

There exists $\mu_\infty \in \mathcal{F}_0$ and $g_\infty \in \mathcal{F}_1$ s.t. $P_0(\mu_n - \mu_\infty)^2 \xrightarrow{p} 0$ and $P_0(g_n - g_\infty)^2 \xrightarrow{p} 0$

Condition 3

There exist subsets $\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3$ of $\mathcal{A} \times \mathcal{W}$ s.t. $P_0(\mathcal{S}_1 \cup \mathcal{S}_2 \cup \mathcal{S}_3) = 1$ and

1. $\mu_\infty(a, w) = \mu_0(a, w), \forall (a, w) \in \mathcal{S}_1$
2. $g_\infty(a, w) = g_0(a, w), \forall (a, w) \in \mathcal{S}_2$
3. $\mu_\infty(a, w) = \mu_0(a, w)$ and $g_\infty(a, w) = g_0(a, w), \forall (a, w) \in \mathcal{S}_3$

These conditions control the uniform entropy of certain classes of functions, which is related to empirical process theory. A thorough treatment is provided in van der Vaart and Wellner (1996)

Consistency

Theorem 1

If Conditions 1-3 hold, then $\theta_n(a) \xrightarrow{p} \theta_0(a)$ for any value $a \in \mathcal{A}$ s.t. $F_0(a) \in (0,1)$, θ_0 is continuous at a and F_0 is strictly increasing in a neighbourhood of a.

If θ_0 is uniformly continuous and F_0 is strictly increasing on $\mathcal{A}$, then $\sup_{a \in \mathcal{A}_0} \left[\theta_n(a) - \theta_0(a)\right] \overset{p}{0}$ for any bounded strict subinterval $\mathcal{A}_0 \subset \mathcal{A}$.

(Well-known) boundary issues with Grenander-type estimators:

- In the pointwise statement, $F_0(a)$ is required to be in $[0,1]$
- Similarly, the uniform statement only covers strict subintervals of $\mathcal{A}$
- Various remedies have been proposed before to mitigate this
- Potential direction for future research

Remark on Condtions for consistency

Remark on Condition 1

Condition 1 requires that μ_n and g_n eventually be contained in uniformly bounded function classes that are sufficiently small for certain empirical process terms to be controlled. This is satisfied by parametric classes and many infinite dimensional function classes. See chapter 2.6 of van der Vaart and Wellner (1996).

There is also an asymmetry between the entropy requirements for $\mathcal{F}_0$ and $\mathcal{F}_1$ in part 2 of Condition 1. This is due to the term $\int \int_{-\infty}^{a} \mu_n(u, w) F_n(du) Q_n(dw)$ appearing in $\Gamma_n(a)$. To control this term, an upper bound of the form $\int_0^1 \log \left[\sup_Q N\{\epsilon, \mathcal{F}_0, L_2(Q)\} \right] d\epsilon$ from the theory of empirical U-process is used (Nolan and Pollard, 1987).

The later part of this paper (section 3.7) considers the use of cross-fitting to avoid these entropy conditions in Condition 1.

Remark on Condtions for consistency

Remark on Condition 2 and 3

Condition 2 requires that μ_n and g_n tend to limit functions μ_∞ and g_∞, and Condition 3 requires that requires that either $\mu_\infty(a, w) = \mu_0(a, w)$ or $g_\infty(a, w) = g_0(a, w)$ for $(F_0 \times Q_0)$ almost every (a, w).

- This is equivalent to saying that μ_n or g_n is consistent?
- Seems to be in line with Kennedy *et al.* (2017)

If either

1. $\mathcal{S}_1$ and $\mathcal{S}_3$ are null sets or
2. $\mathcal{S}_2$ and $\mathcal{S}_3$ are null sets,

then Condition 3 is known simply as double robustness of the estimator θ_n relative to the nuisance functions μ_0 and g_0: θ_n is consistent as long as $\mu_\infty = \mu_0$ or $g_\infty = g_0$. However, Condition 3 is more general than classical double robustness as at least one of μ_n or g_n tends to the truth for **only** almost every point in the domain.

Double robustness

Multiply robustness: preserve consistency even if a subset of the N nuisance models is mispecified in the procedure
Double robustness: $N = 2$, so one model can be mispecified

Example: inverse probability weighted (IPW) estimator

- Y_i: response; $A_i \in \{0, 1\}$: treatment; W: covariates
- Estimator: $\hat{\mu}^{i-IPW} = \frac{1}{n}\sum_{i=1}^{n} \frac{A_i Y_i}{\pi_0(W_i)}$ where $\pi_0(W) = P(A = 1|W)$
  - Often infeasible since functional form $\pi_0(W)$ is unknown
  - (Example) nuisance model 1: $\pi_0(W) = \pi(W; \alpha_0) = \frac{\exp(\alpha_0^T \tilde{W})}{1+\exp(\alpha_0^T \tilde{W})}$
  - $\hat{\mu}^{f-IPW} = \frac{1}{n}\sum_{i=1}^{n} \frac{A_i Y_i}{\pi(W_i;\hat{\alpha})}$
- Augmented IPW (AIPW) estimator:
  - $\hat{\mu}^{f-\phi-IPW} = \frac{1}{n}\sum_{i=1}^{n} \left[\frac{A_i Y_i}{\pi(W_i;\hat{\alpha})} + \left\{1 - \frac{A_i}{\pi(W_i;\hat{\alpha})}\right\}\phi(W_i)\right]$
  - Nuisance model 2: $\phi(W) = E(Y|W, A = 1)$ is the most efficient

See Daniel (2017) for a simple introduction to this topic

Conditions for convergence in distribution

Notations:

- $d(h_1, h_2; a, \epsilon, \mathcal{S})$: pseudodistance; $\sigma_0^2(a, w)$: conditional variance
- $d(h_1, h_2; a, \epsilon, \mathcal{S}) := \sqrt{\sup_{|u-a| \leq \epsilon} E_0\left[I_{\mathcal{S}}(u,W)\{h_1 u,W) - h_2(u,W)\}^2\right]}$
- $\sigma_0^2(a, w) := E_0\left[\{Y - \mu_0(A, W)\}^2 | A = a, W = w\right]$

Condition 4

There exists $\epsilon_0 > 0$ s.t.

1. $\max\left[d(\mu_n, \mu_\infty; a, \epsilon_0, \mathcal{S}_1), d(g_n, g_\infty; a, \epsilon_0, \mathcal{S}_2)\right] = o_p(n^{-1/3})$
2. $\max\left[d(\mu_n, \mu_\infty; a, \epsilon_0, \mathcal{S}_2), d(g_n, g_\infty; a, \epsilon_0, \mathcal{S}_1)\right] = o_p(1)$
3. $d(\mu_n, \mu_\infty; a, \epsilon_0, \mathcal{S}_3) d(g_n, g_\infty; a, \epsilon_0, \mathcal{S}_3) = o_p(n^{-1/3})$

Condition 5

$F_0, \mu_0, \mu_\infty, g_0, g_\infty$ and σ_0^2 are continuously differentiable in a neighbourhood of a uniformly over $w \in \mathcal{W}$

Convergence in distribution

Theorem 2

If Conditions 1-5 hold, then

$$n^{1/3}\{\theta_n(a) - \theta_0(a)\} \xrightarrow{d} \left\{ \frac{4\theta_0'(a)\kappa_0(a)}{f_0(a)} \right\}^{1/3} \mathbb{W}$$

for any $a \in \mathcal{A}$ such that $F_0(a) \in (0,1)$, where $\mathbb{W}$ follows the standard Chernoff distribution and

$$\kappa_0(a) := E_0\left(E_0\left[\left\{ \frac{Y - \mu_\infty(a,W)}{g_\infty(a,W)} + \theta_\infty(a) - \theta_0(a) \right\}^2 \middle| A=a, W \right] g_0(a,W) \right)$$

with $\theta_\infty(a)$ denoting $\int \mu_\infty(a,w) Q_0(dw)$.

We skip the comparison between the limit distributions of θ_n and r_n as it is paritally discussed in p.27. In short, their limit distributions only differ in concentration, which is analogous to findings in linear regression

Remark on Conditions for convergence in distribution

Remark on Condition 4 and 5

The requirements of Condition 4 is equivalent to

1. On $\mathcal{S}_1$ where μ_n is consistent but g_n is not, μ_n converges faster than $n^{-1/3}$ uniformly in a neighbourhood of a,
2. Similarly for g_n on $\mathcal{S}_2$ and
3. On $\mathcal{S}_3$ where both μ_n and g_n are consistent, only the product of their rates of convergence must be faster than $n^{-1/3}$

This suggests the possibility of performing doubly robust inference for $\theta_0(a)$, which is explored in section 4. Note that as discussed in p.34, these conditions are more general than the classical double robustness

We skip the discussion of plug-in estimator $\theta_{\mu_n}(a)$, which can achieve faster rate of convergence than $\theta_n(a)$ but hinges entirely on the consistency of μ_n and may not admit a tractable limit theory

Grenander-type estimation without domain transform

The proposed estimator $\theta_n(a)$ coincides with a generalized Grenander-type estimator for which the marginal exposure empirical distribution function is used as domain transformation

An alternative estimator $\bar{\theta}_n$ could be constructed via Grenander-type estimation **without** the use of any domain transformation. We skip its construction here but there are several points to note:

- $\bar{\theta}_n$ does not generalize the classical isotonic regression
- $\bar{\theta}_n$ is not invariant to strictly increasing transform of A
- Domain of $\mathcal{A}$ needs to be known/chosen in defining $\bar{\theta}_n$
- When $\mu_\infty = \mu_0$, $\theta_n(a)$ and $\bar{\theta}_n$ may have the same limit distribution
- When $\mu_\infty \neq \mu_0$, $\bar{\theta}_n$ is dominated by $\theta_n(a)$ in AMSE sense
  - The transformation improves statistical efficiency in this case
  - Relative gain in efficiency is directly related to the asymptotic bias

Discrete domains

When A is discrete, $\theta_n(a)$ is asymptotically equivalent to the AIPW estimator, which is paritally discussed in p.35

As a result, the large sample properties of $\theta_n(a)$ can be derived from the large sample properties of the AIPW estimator and asymptotically valid inference can be obtained by using standard influence-function-based techniques

We skip the proof here as it is like realizing the isotonic regression of pseudo-outcome under discrete exposure coincides with the AIPW estimator. Instead, we shall have a short discussion on influence function

Influence function

Definition of influence function (Hampel *et al.*, 1986)

Let $T(F)$ be a statistical functional where F is a distribution. The influence function of T at F is given by

$$IF(x; T, F) := \lim_{t \downarrow 0} \frac{T\big[(1-t)F + t\delta_x\big] - T(F)}{t}$$

in those $x \in \mathcal{X}$ where this limit exists.

A complete discussion of this definition usually requires Gâteaux differentiability. We cover some of its usage instead:

- An estimator $\hat{\theta} \approx \theta(P_0) + E_n\big[IF(X)\big]$ can be dominated by a single outlier **unless** IF is bounded
- Asymptotic efficiency bound (Bias, Variance)
- Distributional decomposition, partial identification etc.

See this note for a quick summary

Large sample results for causal effects

The result so far concerns about the causal dose-response $a \mapsto m_0(a)$, which may not hold for the causal effect $(a_1, a_2) \mapsto m_0(a_1) - m_0(a_2)$

If the identification conditions discussed in Section 1.2 applied to each of a_1 and a_2, such causal effects can be identified with the observed data parameter $\theta_0(a_1) - \theta_0(a_2)$

If the condtions of Theorem 1 hold for both a_1 and a_2, we can establish consistency via the use of continuous mapping theorem

However, Theorem 2 only provides marginal distributional results. Joint convergence result is thus required for inference of causal effect

(Joint) convergence for causal effects

Theorem 3

Define $Z_n(a_1, a_2) := \left(n^{1/3}\{\theta_n(a_1) - \theta_0(a_1)\}, n^{1/3}\{\theta_n(a_2) - \theta_0(a_2)\}\right)$. If Conditions 1-5 hold for $a \in \{a_1, a_2\} \subset \mathcal{A}$ and $F_0(a_1), F_0(a_2) \in (0, 1)$, then

$$Z_n(a_1, a_2) \xrightarrow{d} \left(\{4\tau_0(a_1)\}^{1/3}\mathbb{W}_1, \{4\tau_0(a_2)\}^{1/3}\mathbb{W}_2\right)$$

where $\mathbb{W}_1, \mathbb{W}_2$ are independent standard Chernoff distributions and the scale parameter $\tau_0 = \frac{\theta_0'(a)\kappa_0(a)}{f_0(a)}$ is as defined in theorem 2.

Note that Theorem 3 implies

$$n^{1/3}\left[\left\{\theta_n(a_1) - \theta_n(a_2)\right\} - \left\{\theta_0(a_1) - \theta_0(a_2)\right\}\right] \xrightarrow{d} \left\{4\tau_0(a_1)\right\}^{1/3}\mathbb{W}_1 - \left\{4\tau_0(a_2)\right\}^{1/3}\mathbb{W}_2$$

Cross-fitting to avoid empirical process conditions

In observational studies, researchers can rarely specify a *priori* correct parametric models for μ_0 and g_0. This motivates use of data-adaptive estimators to meet Conditions 2 and 3

However, such estimators often leads to violation of Condition 1, or it may be onerous to determine that they do not. See slide p.33

In the context of asymptotically linear estimators, it has been noted that cross-fitting nuisance estimators can resolve this challenge by eliminating empirical process conditions

Therefore, this paper proposes cross-fitting of μ_n and g_n to avoid entropy conditions in Theorem 1 and 2

Estimation with cross-fitting

Estimation procedure with cross-fitting

1. Fix $V \in \{2, 3, \ldots, n/2\}$
2. Randomly partition the indices $\{1, 2, \ldots, n\}$ into V sets $\mathcal{V}_{n,1}, \mathcal{V}_{n,2}, \ldots, \mathcal{V}_{n,V}$
3. Assume $N := n/V \in \mathbb{Z}^+$. For each $v \in \{1, 2, \ldots, V\}$:
   1. Define $\mathcal{T}_{n,v} := \{O_i : i \notin \mathcal{V}_{n,v}\}$ as the *training set* for fold v
   2. Construct $\mu_{n,v}$ and $g_{n,v}$ using only observations from $\mathcal{T}_{n,v}$
4. Define pointwise the cross-fitted estimator Γ_n° of Γ_0 as

$$\Gamma_n^\circ(a) := \frac{1}{V}\sum_{v=1}^{V}\left[\frac{1}{N}\sum_{i\in\mathcal{V}_{n,v}} I_{(-\infty,a]}(A_i)\frac{Y_i-\mu_{n,v}(A_i,W_i)}{g_{n,v}(A_i,W_i)}\right.$$
$$\left. + \frac{1}{N^2}\sum_{i,j\in\mathcal{V}_{n,v}} I_{(-\infty,a]}(A_i)\mu_{n,v}(A_i,W_j)\right]$$

5. Construct the cross-fitted estimator θ_n° as in p.25

Remark: all results hold as long as $\max_v n/|\mathcal{V}_{n,v}| = O_p(1)$

Conditions for convergence under cross-fitting

Condition 6

There exist constants $C', \delta', K_0', K_1', K_2', K_3' \in (0, \infty)$ s.t., almost surely as $n \to \infty$ and for all v, $\mu_{n,v}$ and $g_{n,v}$ are contained in classes of functions $\mathcal{F}_0'$ and $\mathcal{F}_1'$ respectively, satisfying

1. $|\mu| \leq K_0', \forall \mu \in \mathcal{F}_0'$, and $K_1' \leq g \leq K_2', \forall g \in \mathcal{F}_1'$, and
2. $\sigma_0^2(a, w) \leq K_3'$ for almost all a and w

Condition 7

There exist $\mu_\infty \in \mathcal{F}_0'$ and $g_\infty \in \mathcal{F}_1'$ s.t. $\max_v P_0(\mu_{n,v} - \mu_\infty)^2 \xrightarrow{p} 0$ and $\max_v P_0(g_{n,v} - g_\infty)^2 \xrightarrow{p} 0$

Conditions for convergence under cross-fitting

Condition 8

There exists $\epsilon_0 > 0$ s.t.

1. $\max\left[d(\mu_{n,v}, \mu_\infty; a, \epsilon_0, \mathcal{S}_1), d(g_{n,v}, g_\infty; a, \epsilon_0, \mathcal{S}_2)\right] = o_p(n^{-1/3})$
2. $\max\left[d(\mu_{n,v}, \mu_\infty; a, \epsilon_0, \mathcal{S}_2), d(g_{n,v}, g_\infty; a, \epsilon_0, \mathcal{S}_1)\right] = o_p(1)$
3. $d(\mu_{n,v}, \mu_\infty; a, \epsilon_0, \mathcal{S}_3) d(g_{n,v}, g_\infty; a, \epsilon_0, \mathcal{S}_3) = o_p(n^{-1/3})$

Remark: Conditions 6, 7 and 8 are analogue of Conditions 1, 2 and 4 respectively under cross-fitting

Convergence under cross-fitting

Theorem 4

If Conditions 6, 7 and 3 hold, then $\theta_n^\circ(a) \xrightarrow{p} \theta_0(a)$ for any value $a \in \mathcal{A}$ s.t. $F_0(a) \in (0,1)$, θ_0 is continuous at a and F_0 is strictly increasing in a neighbourhood of a.

If θ_0 is uniformly continuous and F_0 is strictly increasing on $\mathcal{A}$, then $\sup_{a \in \mathcal{A}_0} \left[\theta_n^\circ(a) - \theta_0(a)\right] \overset{p}{0}$ for any bounded strict subinterval $\mathcal{A}_0 \subset \mathcal{A}$.

Theorem 5

If Conditions 6, 7, 3, 8, 5 hold, then

$$n^{1/3}\{\theta_n^\circ(a) - \theta_0(a)\} \xrightarrow{d} \{4\tau_0(a)\}^{1/3}\,\mathbb{W}$$

for any $a \in \mathcal{A}$ such that $F_0(a) \in (0,1)$, where $\mathbb{W}$ follows the standard Chernoff distribution.

Agenda

Wald-type Confidence Interval

Wald-type CI

Since the limit distribution of $n^{1/3}\Big[\theta_n(a) - \theta_0(a)\Big]$ is symmetric around zero by the result of Theorem 2, writing $\tau_0(a) := \theta_0'(a)\kappa_0(a)/f_0(a)$ and denoting by $\tau_n(a)$ any consistent estimator of $\tau_0(a)$, then a Wald-type $1-\alpha$ level asymptotic confidence interval for $\theta_0(a)$ is given by

$$\Big[\theta_n(a) - \Big[\frac{4\tau_n(a)}{n}\Big]^{1/3} q_{1-\alpha/2},\ \theta_n(a) + \Big[\frac{4\tau_n(a)}{n}\Big]^{1/3} q_{1-\alpha/2}\Big],$$

where q_p denotes the p^{th} quantile of $\mathbb{W}$.

Wald-type Confidence Interval

Estimation of $\tau_0(a)$

Since estimation of $\tau_0(a)$ involves, it can be estimated either direcly or indirectly, estimation of $\theta_0^{'}(a)/f_0(a)$ and $\kappa_0(a)$. For direct estimation, we note that $\theta_0^{'}(a)/f_0(a) = \psi_0^{'}(F_0(a)$ with $\psi_0 := \theta_0 \circ F_0^{-1}$.It suggests that

1. Estimate $\theta_0^{'}$ and f_0 separately and consider the ratio of these of estimators; or
2. Estimate $\psi_0^{'}$ directly and compose it with the estimator of F_0.

For indirect estimation, it involves the invariance property of the scale estimator to strictly monotone transformations of the exposure. To estimate $\psi_0^{'}$, we recall that the estimator ψ_n from Section 2 is a step function and is not differentiable. A Natural Solution consists of computing the derivative of a smoothed version of ϕ_n. We have found local quadratic kernel smoothing of points $\{(u_j, \psi_n(u_j)) : j = 1, 2, \cdots, \}$, for u_j the midpoints of the jump points of ψ_n, to work well in practice.

Wald-type Confidence Interval

Wald-type CIs for causal effects

Theorem 3 is used to construct the Wald-type CIs for causal effects in the form $\theta_0(a_1) - \theta_0(a_2)$, then a Wald-type $1-\alpha$ level asymptotic confidence interval for $\theta_0(a_1) - \theta_0(a_2)$ is given by

$$\left[\theta_n(a_1) - \theta_n(a_2) - \bar{q}_{n,1-\alpha/2} n^{-1/3},\ \theta_n(a_1) - \theta_n(a_2) + \bar{q}_{n,1-\alpha/2} n^{-1/3}\right]$$

where $\bar{q}_{n,1-\alpha/2}$ denotes the $(1-\alpha/2)$ quantile of $\left[4\tau_n(a_1)\right]^{1/3}\mathbb{W}_1 - \left[4\tau_n(a_2)\right]^{1/3}\mathbb{W}_2$, $\mathbb{W}_1$ and $\mathbb{W}_2$ are independent Chernoff distributions, using Monte Carlo simulations.

Scale estimation relying on consistent nuisance estimation

Plug-in estimator of $\kappa_0(a)$

Consider settings in which both μ_n and g_n are consistent estimators. In such cases, we have $\kappa_0(a) = E_0\left[\sigma_0^2(a, W)/g_0(a, W)\right]$ with $\sigma_0^2(a, W)$ denoting the conditional variance $E_0\{\left[Y - \mu_0(a, W)\right]^2 | A = a, W = w\}$. Then, a plug-in estimator of $\kappa_0(a)$ is given by

$$\kappa_n(a) := \frac{1}{n} \sum_{i=1}^{n} \frac{\sigma_n^2(a, W_i)}{g_n(a, W_i)}$$

Any regression technique could be used to estimate the conditional expectation of $Z_n := \left[Y - \mu_(a, W)\right]^2$ given A and W. If μ_n, g_n and σ_n^2 are consistent estimator, then $\kappa_n(a)$ is a consistent estimator.

Doubly-robust scale estimation

Aim: To construct an estimator of $\kappa_0(a)$ consistent even of either $\mu_\infty \neq \mu_0$ or $g_\infty \neq g_0$

Doubly-robust estimator of $\kappa_0(a)$

Note that $\kappa_0(a) = lim_{h \downarrow 0} E_0\big[K_h(F_0(A) - F_0(a))\eta_\infty(Y, A, W)\big]$, where $K_h : u \mapsto h^{-1}K(uh^{-1})$ for some bounded density function K with bounded support, and defined

$$\eta_\infty : (y, a, w) \mapsto \left[\frac{y - \mu_\infty(a,w)}{g_\infty(a,w)} + \theta_\infty(a) - \theta_0(a)\right]^2$$

Setting $\theta_{\mu_n}(a) := \int \mu_n(a, w)Q_n(dw)$ with Q_n the empirical distribution based on $W_1, W_2, \cdots, W_n$, and define

$$\kappa^*_{n,h}(a) = \frac{1}{n}\sum_{i=1}^n K_h(F_n(A_i) - F_n(a))\eta_n(Y_i, A_i, W_i)$$

by substituting $\mu_\infty, g_\infty, \theta_\infty, \theta_0$ by $\mu_n, g_n, \theta_{\mu_n}, \theta_n$.

Doubly-robust scale estimation

Doubly-robust estimator of $\kappa_0(a)$(Con'd)

Under conditions (A1)-(A5), it can be shown that $\kappa^*_{n,h_n}(a) \xrightarrow{P} \kappa_0(a)$ by standard kernel smoothing arguments for any sequence $h_n \to 0$. In particular, $\kappa^*_{n,h_n}(a)$ is consistent under the general form of doubly-robustness specified by condition(A3).

Appropriate value of bandwidth h

We propose the following empirical criterion:

1. Define the integrated scale $\gamma_0 := \int \kappa_0(a) F_0(da)$ and construct the estimator $\gamma_n(h) := \int \kappa_{n,h}(a) F_n(da)$ for any candidate $h > 0$.
2. Observe that $\gamma_0 = E_0[\eta_\infty(Y, A, W)]$, which suggest $\bar{\eta}_n := \frac{1}{n}\sum_{i=1}^n \eta_n(Y_i, A_i, W_i)$.

It motivates us to define $h^*_n := argmin_h[\gamma_n(h) - \bar{\eta_n}]^2$.

The proposed doubly-robust estimator of $\kappa_0(a)$ is $\kappa_{n,DR}(a) := \kappa_{n,h^*_n}(a)$.

Doubly-robust scale estimation

Remarks

1. $\kappa_{n,DR}(a)$ only depends on A and a through the ranks $F_n(A)$ and $F_n(a)$. Hence, the estimator is invariant to strictly monotone transformations of the exposure.
2. If $\mu_n(a,w) = \mu_n(a)$ does not depend on w and $g_n = 1$, $\kappa_{n,DR}(a)$ tends to the conditional variance $Var_0(Y|A=a)$, which is precisely the scale parameter appearing in standard isotonic regression.

Confidence intervals via sample splitting

Sample splitting method recently proposed by Banerjee er al.(2019) could also be used to perform inference. To implement the approach in our context: (1)we randomly split the sample into m subsets of roughly equal size;(2) perform the estimation procedure on each subset to form subset-specific estimates $\theta_{n,1}, \theta_{n,2}, \cdots, \theta_{n,m}$; and (3) define $\bar{\theta}_{n,m}(a) := \frac{1}{m}\sum_{j=1}^{m}\theta_{n,j}(a)$. If $m > 1$ is fixed, then under mild conditions $\bar{\theta}_{n,m}(a)$ has strictly better asymptotic MSE than $\theta_n(a)$.

CI via sample splitting

For moderate m, the asymptotic $1-\alpha$ level confidence interval for $\theta_0(a)$ is given by

$$\left[\bar{\theta}_{n,m}(a) - \frac{\sigma_{n,m}(a)}{\sqrt{m}n^{1/3}}t_{1-\alpha/2,m-1},\ \bar{\theta}_{n,m}(a) + \frac{\sigma_{n,m}(a)}{\sqrt{m}n^{1/3}}t_{1-\alpha/2,m-1}\right]$$

where $\sigma^2_{n,m}(a) := \frac{1}{m-1}\sum_{j=1}^{m}[\theta_{n,j}(a) - \bar{\theta}_{n,m}(a)]^2$ and $t_{1-\alpha/2,m-1}$ is the $(1-\alpha/2)$ quantile of the t-distribution with $m-1$ degrees of freedom.

Agenda

Numerical Studies

Goals: Perform numerical experiments to assess the performance of the proposed estimators of $\theta_0(a)$ and of the three approaches for constructing confidence intervals, which we also compare to that of the local linear estimator and associated confidence intervals proposed in Kennedy et al.(2017).

Idea of Data Generating Process

1. Generate $W \in \mathbb{R}^4$ as a vector of four independent standard normal variate.
2. Generate U given W, and then transform U to obtain A. ($\because$ the estimation procedures requires estimating the conditional density of $U := F_0(A)$ given W)

Data Generating Process

1. $W \sim N(0, \sum)$
   where $\sum$ is the 4×4 diagonal matrix with value 1.
2. $U \sim \bar{g}_0(u|w) = I_{[0,1]}(u)\{\lambda(w) + 2u[1 - \lambda(w)]\}$
   where $\lambda(w) := 0.1 + 1.8\ expit(\beta^\top w)$.
3. $A \sim 0.5N(-2, 1) + 0.5N(2, 1)$
4. $[Y|A = a, W = w] \sim Bern(\mu_0(a, w))$
   where $\mu_0(a, w) := expit(\gamma_1^\top \underline{w} + \gamma_2^\top \underline{w}a + \gamma_3 a^2)$
5. $\beta = (-1, -1, 1, 1)^\top, \gamma_1 = (-1, -1, -1, 1, 1)^\top, \gamma_1 = (3, -1, -1, 1, 1)^\top$
   and $\gamma_3 = 3$

Estimation and Conditions

Estimator used

1. Causal isotonic regression estimator θ_n.
2. Local linear estimator of Kennedy et al.(2017) with data-driven bandwidth selection procedure proposed in Section 3.5.
3. Sample-splitting version of θ_n with $m = 5$ splits.

Settings

1. Both μ_n and g_n are consistent.
2. Only μ_n is consistent.
3. Only g_n is consistent.

Estimation and Conditions

Estimator under settings

1. If μ_n and g_n are consistent, then logistic regression model and maximum likelihood estimator based on the correctly specified parametric model are used.
2. If μ_n is inconsistent, then logistic regression model is used but omit covariates W_3, W_4 and all interactions.
3. If g_n is inconsistent, then posit the parametric model as before but omit W_3 and W_4.

Confidence intervals for causal effect

Construction of CIs

Constructing pointwise confidence intervals for θ_0 in each setting from section 4 using Plug-in and doubly-robust estimators of $\kappa_0(a)$. We expect that

1. asymptotically correct coverage rates for each of the three settings when doubly-robust estimator of $\kappa_0(a)$ is used; while
2. asymptotically correct coverage rates for the first setting when Plug-in estimator of $\kappa_0(a)$ is used.

As before, pointwise confidence intervals for the local linear estimator and the sample splitting procedure will be demonstrated also. And, we consider the performance of these inferential procedures for values of a between -3 and 3.

Simulation Results

Adjusted and unadjusted regression

For $n = 5000$, note that $\theta_0(a) \neq \tau_0(a)$ for $a \neq 0$. Since the relationship between Y and A is confounding by W, the unadjusted regression curve dose not have a causal interpretation. Therefore, the marginal isotonic regression estimator will not be consistent for the true causal parameter.

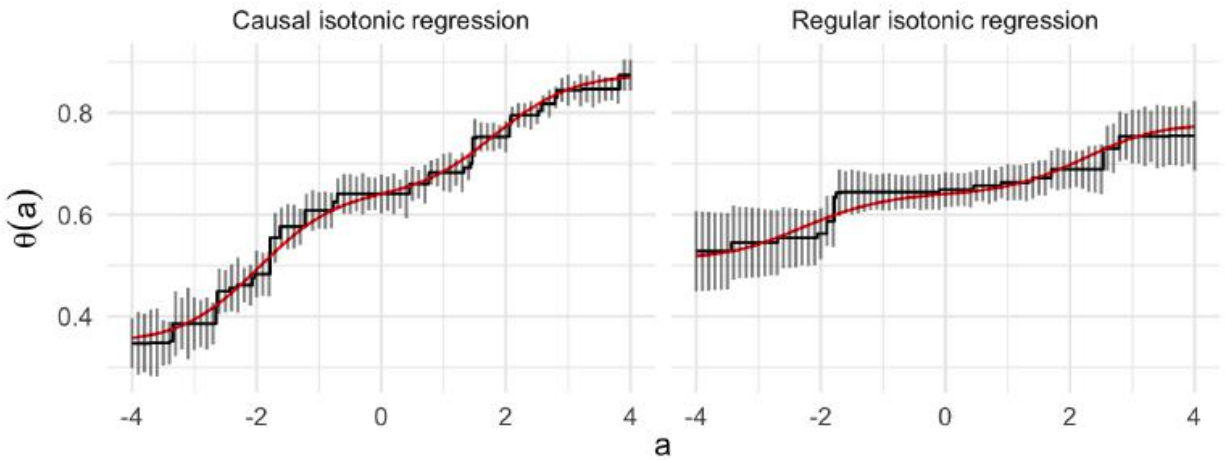


Figure 1: Causal isotonic regression estimate using consistent nuisance estimators μ_n and g_n (left), and regular isotonic regression estimate (right). Pointwise 95% confidence intervals constructed using the doubly-robust estimator are shown as vertical bars. The true functions are shown in red.

Simulation Results

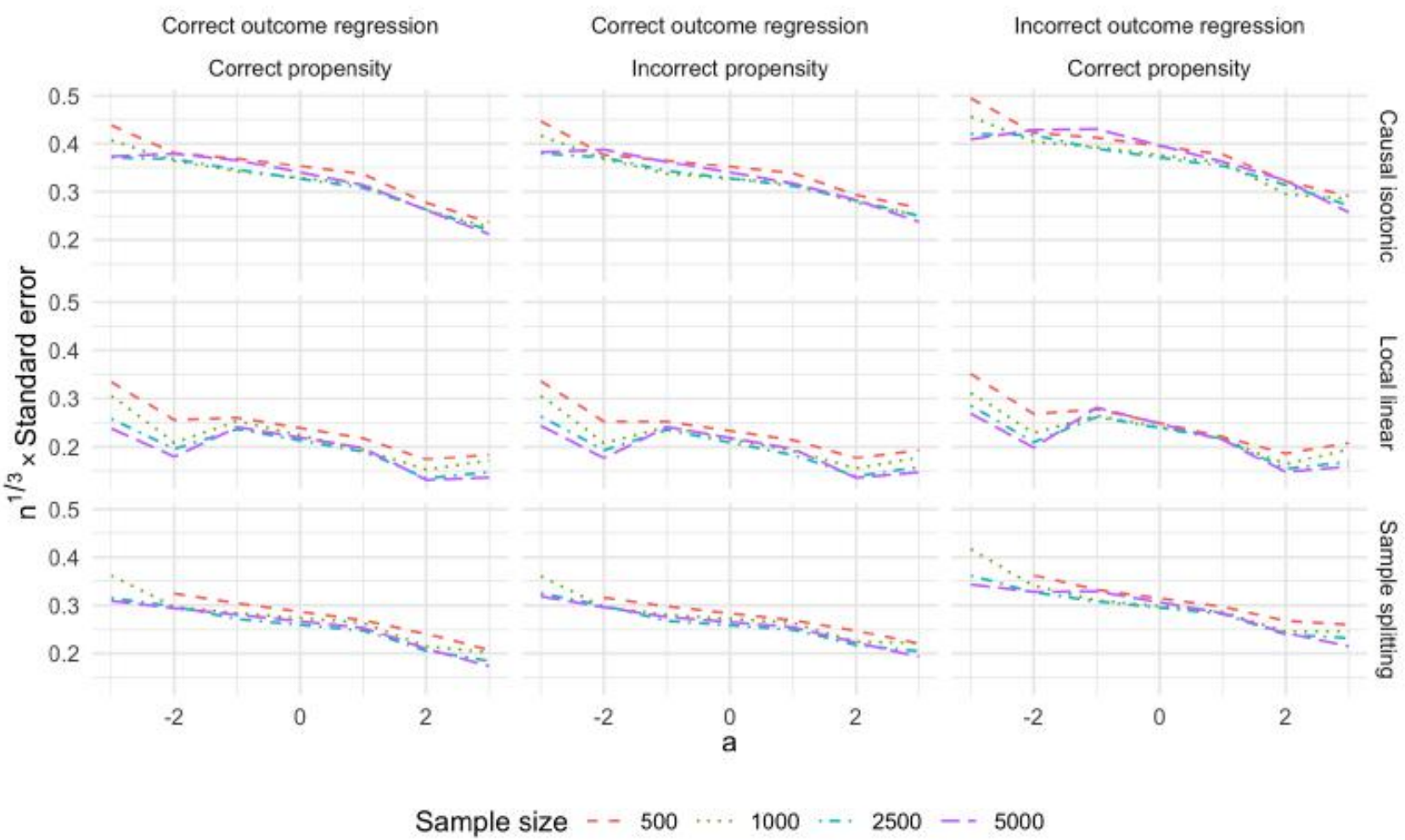


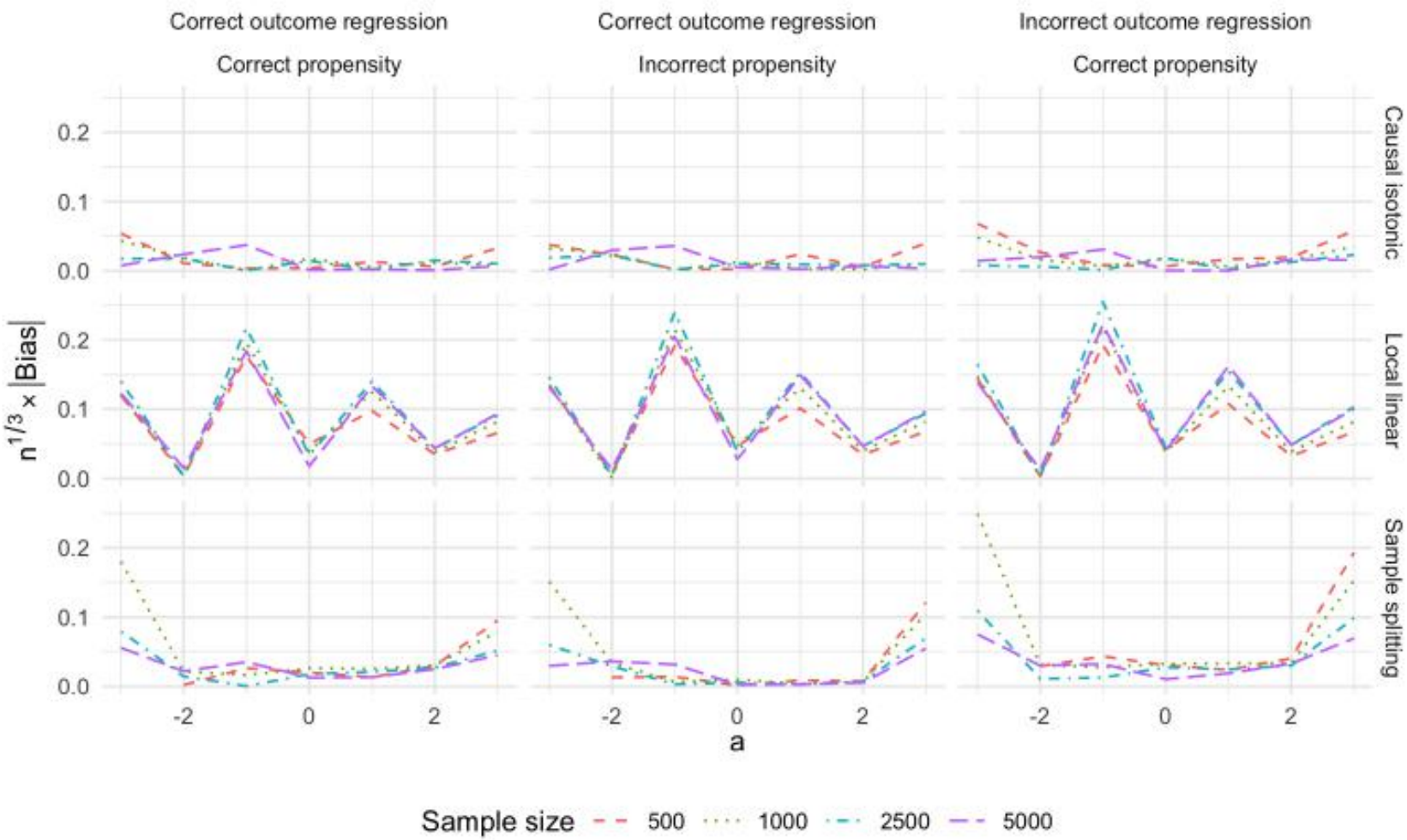
Figure 2: Standard error of the three estimators scaled by $n^{1/3}$ as a function of n for different values of a and in contexts in which μ_n and g_n are either consistent or inconsistent, computed empirically over 1000 simulated datasets of different sizes.

Simulation Results

Standard error

1. The standard error of the local linear estimator is smaller than that of θ_n, is expected due to the fast rate of convergence.
2. The standard deviation of the local linear estimator appears to decrease faster than $n^{-1/3}$.
3. Inconsistent estimation of the propensity has little impact on the standard errors of any of the estimators but inconsistent estimation of the outcome regression not.

Simulation Results

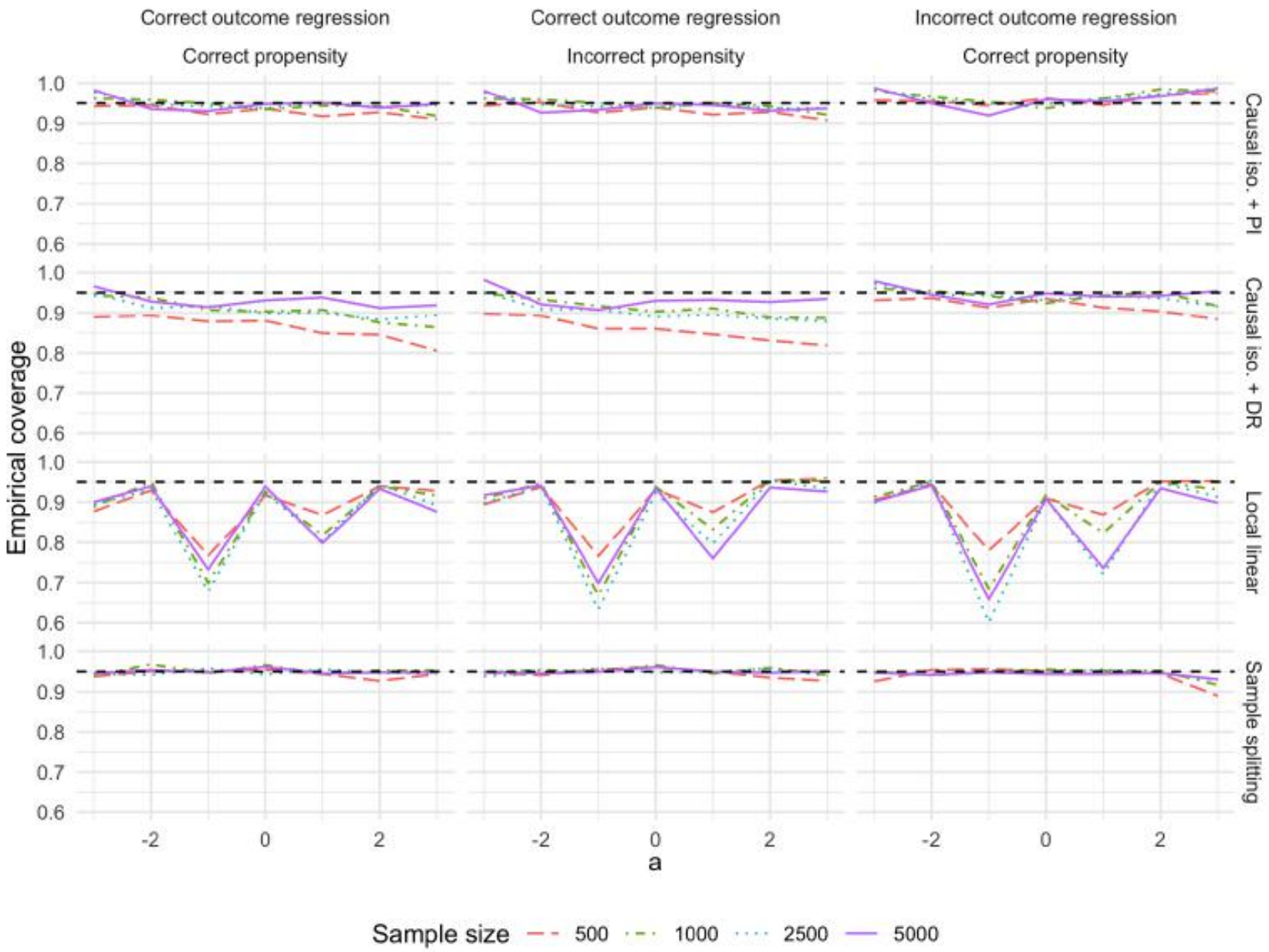


Simulation Results

Absolute bias

1. The estimator of θ_n has smaller absolute bias than the local linear estimator, and its absolute bias decreases faster than $n^{-1/3}$.
2. The absolute bias of the local linear estimator depends strongly on a, and in particular is largest where the second derivative of θ_0 is larger in absolute value.
3. The sample splitting estimator has larger absolute bias than θ_n since it inherits the bias of $\theta_{n/m}$. The bias is large for values of a in the tails of the marginal distribution of A.

Simulation Results



Simulation Results

Coverage of pointwise 95% confidence intervals

1. Both plug-in and doubly-robust estimator intervals centered around θ_n, the coverage improves as n grows.
2. Under correct specification of outcome and propensity regression models, the plug-in method attains close to nominal coverage. When the propensity estimator is inconsistent, the plug-in method still performs well in this case. When μ_n is consistent, the plug-in method is very conservative for positive values of a.
3. The doubly-robust method attains close to nominal coverage for large samples as one of g_n or μ_n is consistent.
4. The local linear estimator has poor coverage for values of a where the bias of the estimator is large.
5. Sample-splitting method performs excellent except perhaps the value of a in the tails when n is small or moderate.

Simulation Results

A simulation study is conducted to illustrate the performance of the proposed procedures when machine learning techniques are used to construct μ_n and g_n.

Estimator under settings

Consider the estimator θ_n° obtained via cross-fitting (mentioned in section 3.7).

1. If μ_n is consistent, then use Super Learner (van der Laan et al. 2017)
2. If g_n is consistent, then used the method proposed by Diaz and van der Laan(2011) with covariate vector (W_1, W_2, W_3, W_4).
3. If μ_n or g_n is inconsistent, then omit covariates W_1 and W_2.

Simulation Results

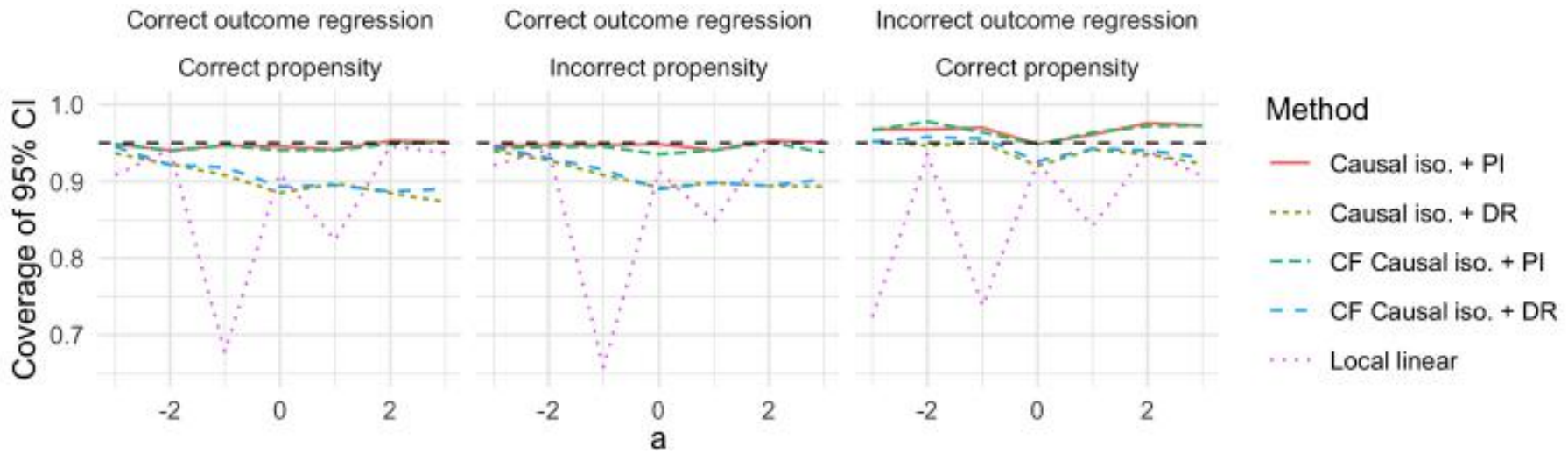


Figure 5: Observed coverage of pointwise 95% doubly-robust and plug-in confidence intervals using machine learning estimators based on simulated data including $n = 1000$ observations. Columns indicate whether μ_n and g_n are consistent or not. Black dashed lines indicate the nominal coverage rate. CF stands for cross-fitted; PI for plug-in; DR for doubly-robust.

Simulation Results

Coverage of 95% of confident intervals

1. The plug-in method performs well which achieve very close to nominal coverage under both consistent settings, even propensity is inconsistently estimated.
2. The doubly-robust method is anti-conservative under both inconsistent settings and also when the propensity is inconsistently estimated. Good coverage rates are also achieved when the outcome regression is inconsistently estimated.
3. Cross fitting has little impact on coverage.

Simulation Results

Conclusion of results

1. For plug-in method, under the inconsistent estimation of any nuisance function, the scale parameter is biased and its variance decreases relatively quickly with sample size by the simple empirical average of estimated functions.
2. For doubly-robust method, the scale parameter is asymptotically unbiased but its variance decreases much slower with sample size.

Agenda

Literature Review on BMI and cells response

Some previous scientific literature indicates that

1. BMI is inversely associated with immune responses to vaccination
2. higher BMI might lead to impaired immune responses
3. obesity reduced hepatitis B immune responses through leptin-induced systemic and B cell intrinsic inflammation, impaired T cell responses

AIM: assess the covariate-adjusted relationship between BMI and CD4+ T-cell responses using data from a collection of clinical trials of candidate HIV vaccines

Comment on previous literature

1. Jin et al. (2015): low BMI participants had a statistically significantly greater response rate than high BMI participants by using Fisher?s exact test.
   Comment: Marginal comparison can be misleading due to existence of confounders such as age and sex.
2. Jin et al. (2015): logistic regression of the binary CD4+ responses against sex, age, BMI (not discretized), vaccination dose and number of vaccinations.
   Comment: adjusted odds ratio has a formal causal interpretation only under strong parametric assumptions.

Comparetively, the method proposed in this paper can identify the covariate-adjusted dose-response function θ_0 with the causal dose-response curve without making parametric assumptions.

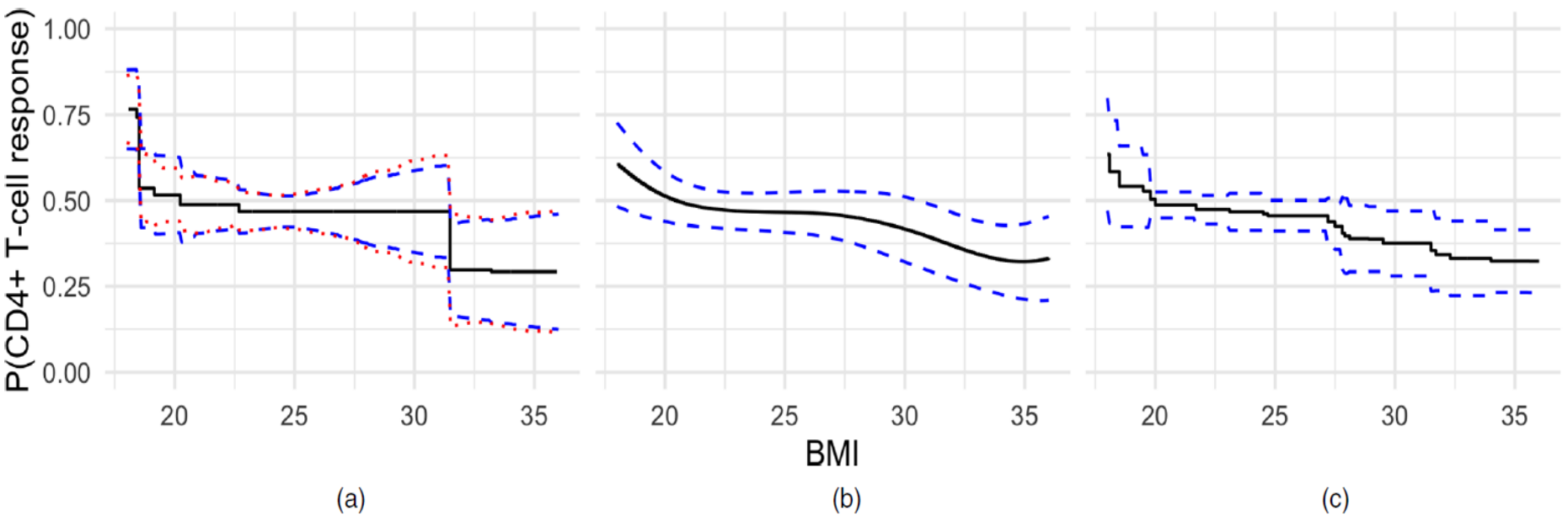
Argument in Causal inference

Some researchers suggest that causal model should always be tied to hypothetical randomized experiments. However, randominzed experiments are commonly not practical. For example,

1. Not ethical to force someone to smoke or not to smoke
2. **Impossible** to assign the BMI to the participants randomly.

one may think it is not appropriate to interpret the G-computed regression function $\theta_0(a)$ in causal manner.
However, it provides a meaningful summary of the relationship between BMI and immune response accounting for measured potential confounders.

Result



The graph refer to the estimated probabilities of CD4+ T-cell response and 95% pointwise confidence intervals as a function of BMI, adjusted for sex, age, number of vaccinations received, vaccine dose and study with

1. Estimator proposed here
2. Local linear estimator of Kennedy et al. (2017)
3. Sample splitting version of our estimator with $m = 5$ splits

Agenda

Furture direction

1. Inference on a monotone causal dose-response curve when outcome data are only observed subject to potential coarsening, such as censoring, truncation or missingness
2. Develop tests of the monotonicity assumption.
3. Develop methods for uniform inference.
4. Inferential methods that do not require estimation of additional nuisance parameters or sample splitting

Comparison between methods

There is some tradeoff between local linear smoothing and monotonicity-based methods.

1. Convergence rate of regression estimator:
   faster for local linear regression estimator; $n^{-2/5}$ VS $n^{-1/3}$ for monotonicity based method.
2. Local linear smoothing: Limit distribution involves an asymptotic bias term depending on the second derivative of the true function, so confidence intervals based on optimally chosen tuning parameters provide asymptotically correct coverage only for a smoothed parameter rather than the true parameter of interest.
3. Monotonicity based: Do not require choosing a tuning parameter, are invariant to strictly increasing transformations of the exposure and their limit theory does not include any asymptotic bias.