NONPARAMETRIC INFERENCE ON DOSE-RESPONSE CURVES WITHOUT THE POSITIVITY CONDITION

Yikun Zhang†, Yen-Chi Chen, and Alexander Giessing

Department of Statistics, University of Washington

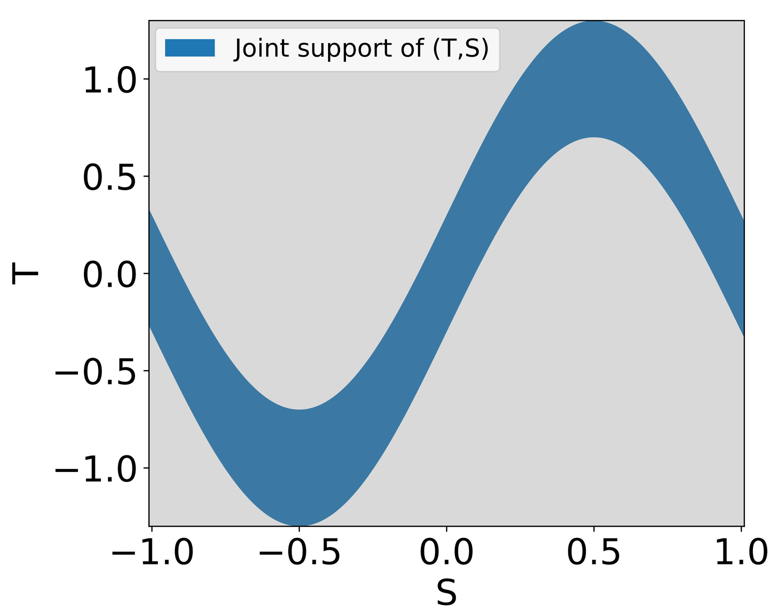
† yikun@uw.edu

INTRODUCTION

Estimating the causal effects for continuous treatments (*i.e.*, the dose-response curves) often relies on the **positivity condition**:

Every subject has some chance of receiving any treatment level $T = t$ regardless of its covariates $\boldsymbol{S} = \boldsymbol{s} \in \mathbb{R}^d$.

- This condition **could fail** in observational studies with continuous treatments.



- We propose a novel integral estimator of the dose-response curve without assuming the positivity condition.
  1. It is based on a localized derivative estimator and the fundamental theorem of calculus.
  2. It can be efficiently computed in practice via Riemann sum approximations.
  3. It can be combined with bootstrap methods for valid inference on the dose-response curve and its derivative.

IDENTIFICATION CONDITIONS

Assume that $\{(Y_i, T_i, \boldsymbol{S}_i)\}_{i=1}^n$ are IID from the model:

$$Y = \mu(T, \boldsymbol{S}) + \epsilon \quad \text{and} \quad T = f(\boldsymbol{S}) + E,$$

where $E \perp\!\!\!\perp \boldsymbol{S}, \epsilon$, $\epsilon \perp\!\!\!\perp \boldsymbol{S}$, $\mathbb{E}(E) = \mathbb{E}(\epsilon) = 0$, $\mathbb{E}(E^2) > 0$, and $\mathbb{E}(\epsilon^4) < \infty$.

Dose-response curve and its **derivative function** can be identified with observed data as:

$$m(t) = \mathbb{E}\left[\mu(t, \boldsymbol{S})\right] \quad \text{and} \quad \theta(t) = m'(t) = \frac{d}{dt}\mathbb{E}\left[\mu(t, \boldsymbol{S})\right]$$

under *consistency* and *ignorability* assumptions.

Interchangability Assumption: The function $\mu(t, \boldsymbol{s})$ is continuously differentiable with respect to t and

$$\mathbb{E}\left[\mu(T, \boldsymbol{S})\right] = \mathbb{E}\left[m(T)\right],$$

$$\theta(t) = \mathbb{E}\left[\frac{\partial}{\partial t}\mu(t, \boldsymbol{S})\right] = \mathbb{E}\left[\frac{\partial}{\partial t}\mu(t, \boldsymbol{S}) \Big| T = t\right].$$

MOTIVATING EXAMPLE

Consider the following additive confounding model:

$$Y = m(T) + \eta(\boldsymbol{S}) + \epsilon \quad \text{and} \quad T = f(\boldsymbol{S}) + E$$

with $\mathbb{E}\left[\eta(\boldsymbol{S})\right] = 0$. This model satisfies our interchangability assumption and is known as the geoadditive structural equation in spatial statistics.

THREE KEY INSIGHTS

1. $\mu(t, \boldsymbol{s})$ and $\frac{\partial}{\partial t}\mu(t, \boldsymbol{s})$ can be consistently estimated at each observation $(T_i, \boldsymbol{S}_i)$.
2. $\theta(t)$ can be consistently estimated by the localized form $\theta_C(t) = \mathbb{E}\left[\frac{\partial}{\partial t}\mu(t, \boldsymbol{S}) \big| T = t\right]$.
3. By the fundamental theorem of calculus,

$$m(t) = m(T) + \int_{\tilde{t}=T}^{\tilde{t}=t} m'(\tilde{t})\, d\tilde{t} = m(T) + \int_{\tilde{t}=T}^{\tilde{t}=t} \theta(\tilde{t})\, d\tilde{t}.$$

$\Rightarrow$ Taking the expectation on both sides yield that

$$m(t) = \mathbb{E}\left[\mu(T, \boldsymbol{S})\right] + \mathbb{E}\left[\int_{\tilde{t}=T}^{\tilde{t}=t} \theta_C(\tilde{t})\, d\tilde{t}\right] = \mathbb{E}(Y) + \mathbb{E}\left[\int_{\tilde{t}=T}^{\tilde{t}=t} \theta_C(\tilde{t})\, d\tilde{t}\right].$$

PROPOSED ESTIMATORS

Proposed Integral Estimator of $m(t)$:

$$\widehat{m}_\theta(t) = \frac{1}{n}\sum_{i=1}^n \left[Y_i + \int_{\tilde{t}=T_i}^{\tilde{t}=t} \widehat{\theta}_C(\tilde{t})\, d\tilde{t}\right],$$

where $\widehat{\theta}_C(t)$ is a consistent estimator of $\theta_C(t) = \int \beta_2(t, \boldsymbol{s})\, d\mathrm{P}(\boldsymbol{s}|t)$ with $\beta_2(t, \boldsymbol{s}) \equiv \frac{\partial}{\partial t}\mu(t, \boldsymbol{s})$.

- Fit $\beta_2(t, \boldsymbol{s})$ by local polynomial regression;
- Estimate $\mathrm{P}(\boldsymbol{s}|t)$ by Nadaraya-Watson conditional CDF estimator.

Proposed Localized Estimator of $\theta(t)$:

$$\widehat{\theta}_C(t) = \frac{\sum_{i=1}^n \widehat{\beta}_2(t, \boldsymbol{S}_i) \cdot \bar{K}_T\left(\frac{T_i - t}{\hbar}\right)}{\sum_{j=1}^n \bar{K}_T\left(\frac{T_j - t}{\hbar}\right)}.$$

FAST COMPUTING ALGORITHM

Let $T_{(1)} \leq \cdots \leq T_{(n)}$ be the order statistics of $T_1, ..., T_n$ and $\Delta_j = T_{(j+1)} - T_{(j)}$ for $j = 1, ..., n-1$.

- Approximate $\widehat{m}_\theta(T_{(j)})$ for $j = 1, ..., n$ as:

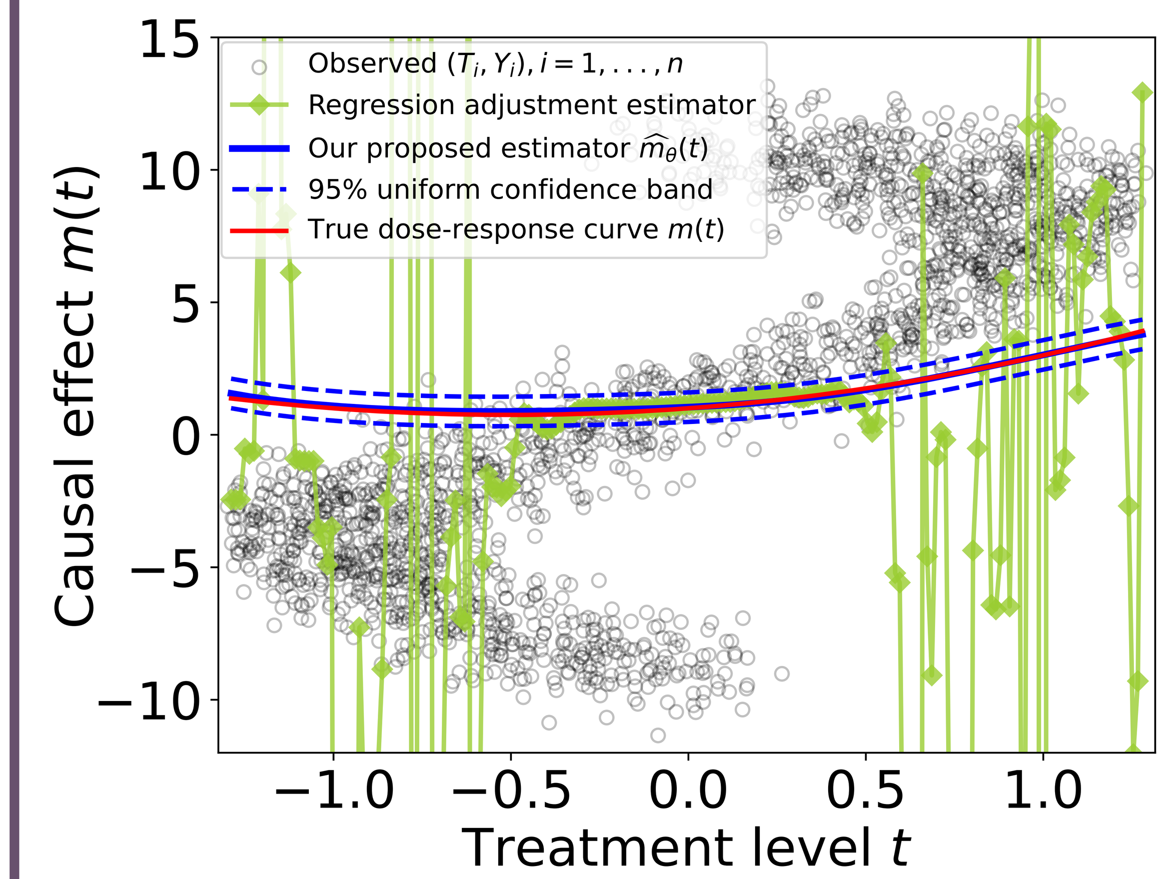
$$\widehat{m}_\theta(T_{(j)}) \approx \frac{1}{n}\sum_{i=1}^n Y_i + \frac{1}{n}\sum_{i=1}^{n-1} \Delta_i \Big[i \cdot \widehat{\theta}_C(T_{(i)}) \mathbb{1}_{\{i<j\}} - (n-i) \cdot \widehat{\theta}_C(T_{(i+1)}) \mathbb{1}_{\{i \geq j\}}\Big].$$

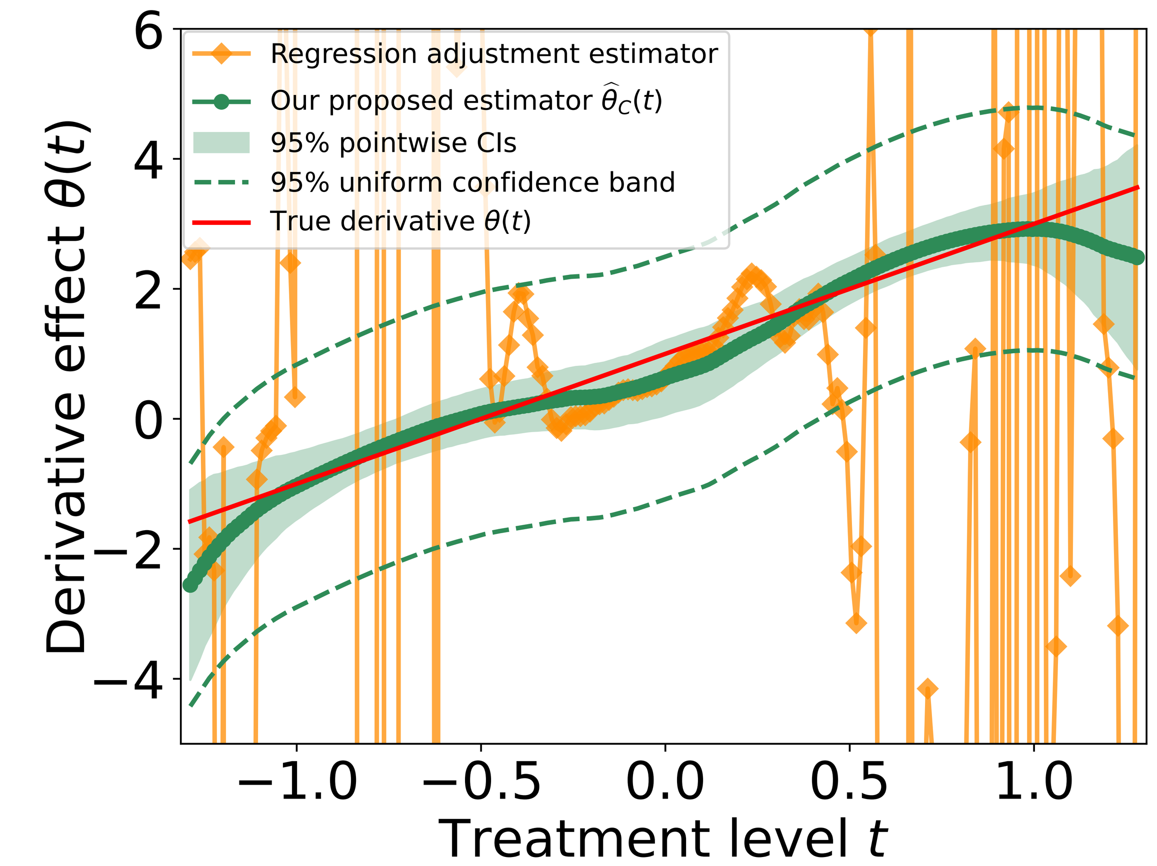
- Evaluate $\widehat{m}_\theta(t)$ at any $t \in \left[T_{(j)}, T_{(j+1)}\right]$ by a linear interpolation between $\widehat{m}_\theta(T_{(j)})$ and $\widehat{m}_\theta(T_{(j+1)})$.

SIMULATION STUDIES

- **Single Confounder Model:**

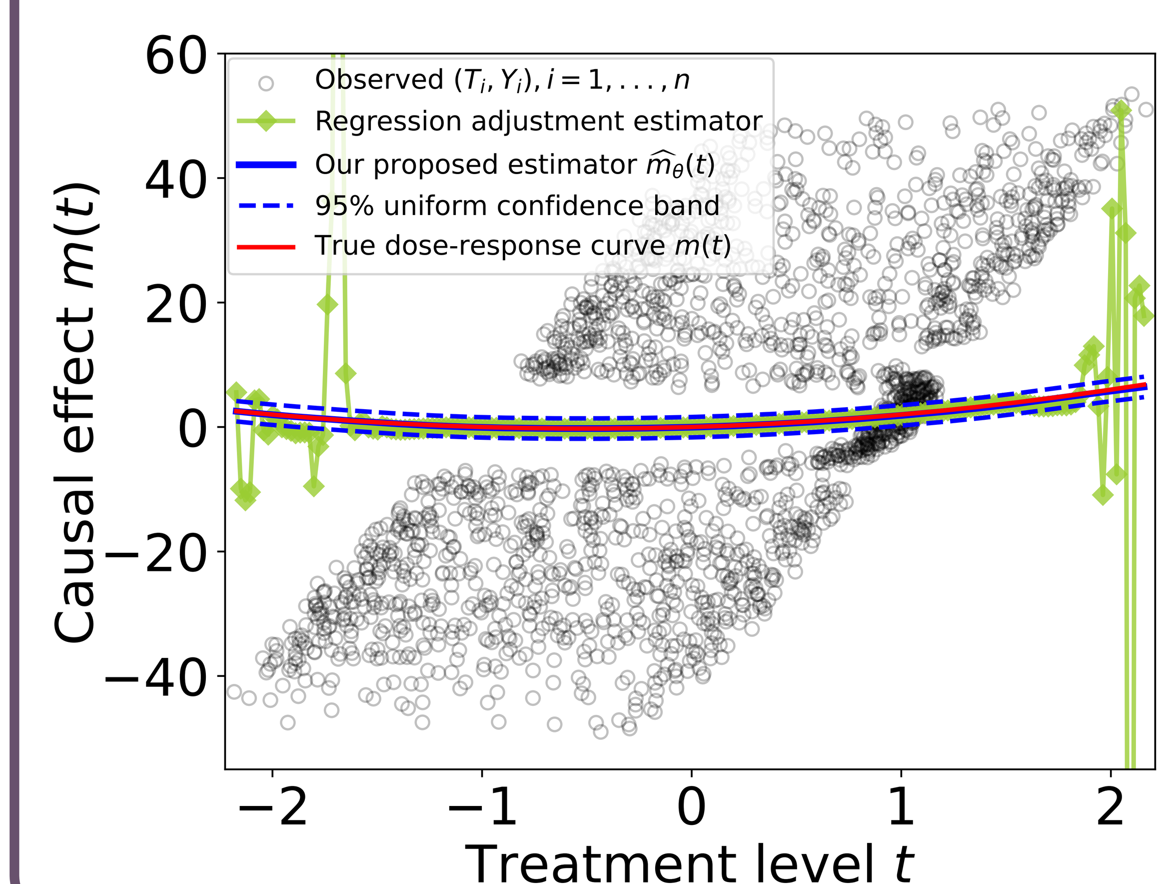
$$Y = T^2 + T + 1 + 10S + \epsilon,\ T = \sin(\pi S) + E,\ S \sim \mathrm{Unif}[-1,1] \subset \mathbb{R},\ E \sim \mathrm{Unif}[-0.3, 0.3],\ \text{and}\ \epsilon \sim \mathcal{N}(0,1).$$

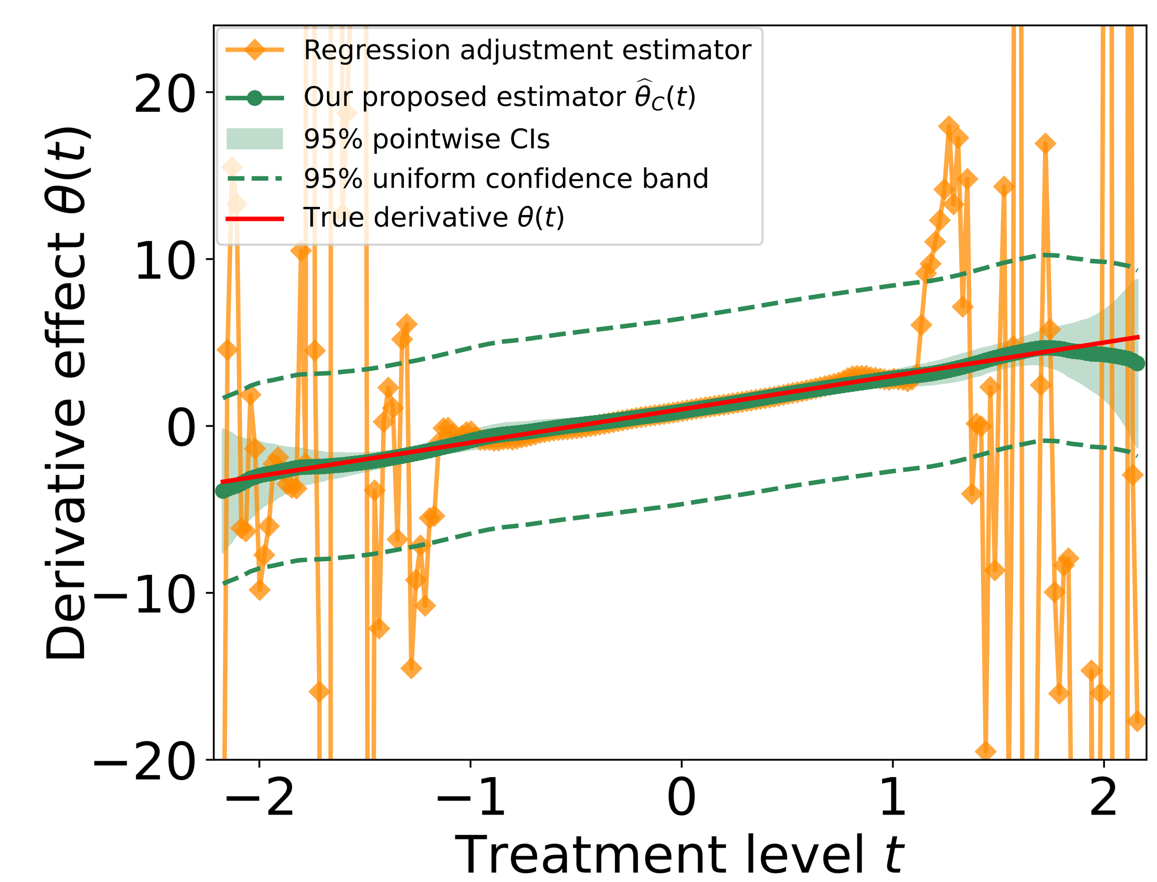




- **Nonlinear Confounding Model:**

$$Y = T^2 + T + 10Z + \epsilon, \quad T = \cos\left(\pi Z^3\right) + Z/4 + E, \quad Z = 4S_1 + S_2,$$
$$\boldsymbol{S} = (S_1, S_2) \sim \mathrm{Unif}[-1,1]^2 \subset \mathbb{R}^2, \quad E \sim \mathrm{Unif}[-0.1, 0.1], \quad \text{and} \quad \epsilon \sim \mathcal{N}(0,1).$$





EFFECT OF $PM_{2.5}$ ON CARDIOVASCULAR MORTALITY RATE (CMR)

The covariate vector $\boldsymbol{S} \in \mathbb{R}^{10}$ includes spatical locations (longitude, latitude) and eight socioeconomic factors.