Nonparametric Inference on Dose-Response Curves Without the Positivity Condition

Yikun Zhang

Joint work with *Professor Yen-Chi Chen*

Department of Statistics,
University of Washington

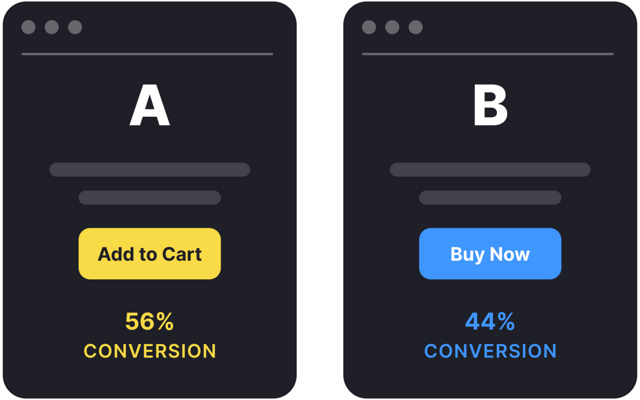
CIMA Lab, North Carolina State University
March 17, 2025

Department of STATISTICS

UNIVERSITY of WASHINGTON

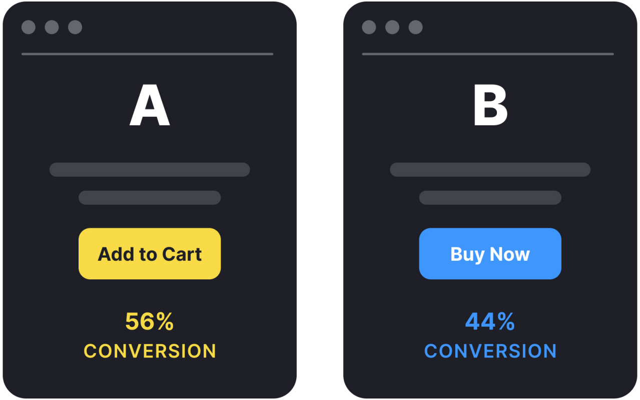
Introduction

- Study the causal effect of a treatment $T \in \mathcal{T}$ on the outcome of interest $Y \in \mathcal{Y}$.



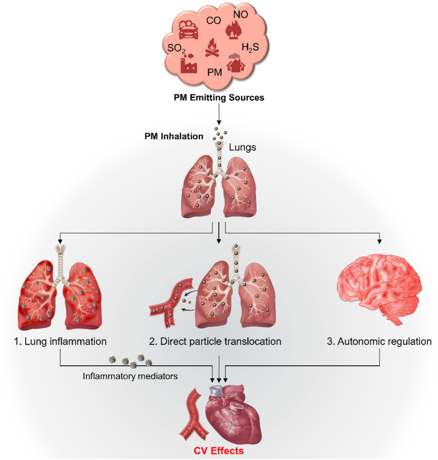
Fundamental Problem of Causal Inference

- Study the causal effect of a treatment $T \in \mathcal{T}$ on the outcome of interest $Y \in \mathcal{Y}$.



- The treatment variable T is *binary*, *i.e.*, $\mathcal{T} = \{0, 1\}$.
- Only one potential outcome, $Y(1)$ or $Y(0)$, can be observed for each individual.
- The common causal estimand is the average treatment effect $\mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$.

- We want to study the causal effects of $PM_{2.5}$ levels on Cardiovascular Mortality Rates (CMRs).



Biological pathways associated with particulate matter (PM) and cardiovascular disease (Miller and Newby, 2020; Basith et al., 2022).

FIPS	County name	Longitude	Latitude	PM2.5	CMR
1025	Clarke	-87.830772	31.676955	6.766443	379.421713
1061	Geneva	-85.839330	31.094869	8.254272	378.524698
1073	Jefferson	-86.896571	33.554343	10.825441	352.790427
1077	Lauderdale	-87.654117	34.901500	9.208783	332.594557
5085	Lonoke	-91.887917	34.754412	8.213144	365.061085
8045	Garfield	-107.903621	39.599420	2.601772	250.781477

The dataset contains the average annual cardiovascular mortality rates (CMRs) and $PM_{2.5}$ levels across $n = 2132$ U.S. counties from 1990 to 2010 (Wyatt et al., 2020a,b).

FIPS	County name	Longitude	Latitude	PM2.5	CMR
1025	Clarke	-87.830772	31.676955	6.766443	379.421713
1061	Geneva	-85.839330	31.094869	8.254272	378.524698
1073	Jefferson	-86.896571	33.554343	10.825441	352.790427
1077	Lauderdale	-87.654117	34.901500	9.208783	332.594557
5085	Lonoke	-91.887917	34.754412	8.213144	365.061085
8045	Garfield	-107.903621	39.599420	2.601772	250.781477

The dataset contains the average annual cardiovascular mortality rates (CMRs) and $PM_{2.5}$ levels across $n = 2132$ U.S. counties from 1990 to 2010 (Wyatt et al., 2020a,b).

- The treatment variable T, *i.e.*, the $PM_{2.5}$ level at each county, is a quantitative measure. In other words, it is *not a binary but continuous variable*!

For *binary* treatment (*i.e.*, $\mathcal{T} = \{0, 1\}$), common causal estimands are

- $\mathbb{E}[Y(t)]$ = mean counterfactual outcome when we set $T = t$.
- $\mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$ = average treatment effect.

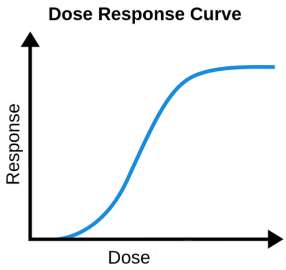
▶ **Question:** What are the counterparts of the above estimands under *continuous* treatment (*i.e.*, $\mathcal{T} \subset \mathbb{R}$)?

For *binary* treatment (*i.e.*, $\mathcal{T} = \{0, 1\}$), common causal estimands are

- $\mathbb{E}\left[Y(t)\right]$ = mean counterfactual outcome when we set $T = t$.
- $\mathbb{E}\left[Y(1)\right] - \mathbb{E}\left[Y(0)\right]$ = average treatment effect.

▶ **Question:** What are the counterparts of the above estimands under *continuous* treatment (*i.e.*, $\mathcal{T} \subset \mathbb{R}$)?

- $t \mapsto m(t) := \mathbb{E}\left[Y(t)\right]$ = (causal) dose-response curve.
- $t \mapsto \theta(t) := m'(t) = \frac{d}{dt}\mathbb{E}\left[Y(t)\right]$ = (causal) derivative effect curve.



Identification of Dose-Response Curves in RCTs

In randomized controlled trials (RCTs),

$$m(t) = \mathbb{E}\left[Y(t)\right] = \mathbb{E}\left(Y|T=t\right) \quad \text{and} \quad \theta(t) = \frac{d}{dt}\mathbb{E}\left(Y|T=t\right).$$

Identification of Dose-Response Curves in RCTs

In randomized controlled trials (RCTs),

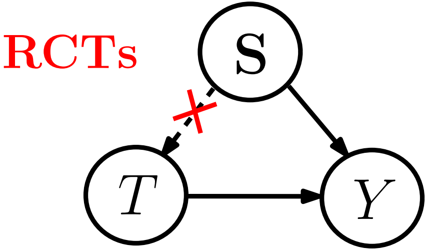
$$m(t) = \mathbb{E}\left[Y(t)\right] = \mathbb{E}\left(Y|T=t\right) \quad \text{and} \quad \theta(t) = \frac{d}{dt}\mathbb{E}\left(Y|T=t\right).$$

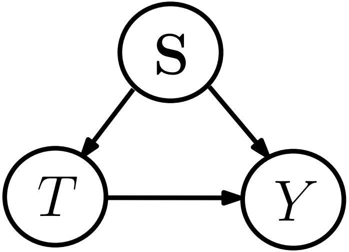
- We can estimate $m(t)$ by fitting a regression on $\{(Y_i, T_i)\}_{i=1}^n$.
- Recovering $\theta(t)$ is a derivative estimation problem (Gasser and Müller, 1984).

In randomized controlled trials (RCTs),

$$m(t) = \mathbb{E}\left[Y(t)\right] = \mathbb{E}\left(Y|T=t\right) \quad \text{and} \quad \theta(t) = \frac{d}{dt}\mathbb{E}\left(Y|T=t\right).$$

- We can estimate $m(t)$ by fitting a regression on $\{(Y_i, T_i)\}_{i=1}^n$.
- Recovering $\theta(t)$ is a derivative estimation problem (Gasser and Müller, 1984).





- Some identification assumptions are required to estimate $m(t)$ and $\theta(t)$ from the observable data $\{(Y_i, T_i, S_i)\}_{i=1}^n$.

Assumption (Identification Conditions)

1. *(Consistency)* $Y = Y(t)$ *whenever* $T = t \in \mathcal{T}$.
2. *(Ignorability)* $Y(t)$ *is conditionally independent of* T *given* $\boldsymbol{S}$ *for all* $t \in \mathcal{T}$.
3. *(**Positivity**) The conditional density satisfies* $p(t|\boldsymbol{s}) \geq p_{\min} > 0$ *for all* $(t, \boldsymbol{s}) \in \mathcal{T} \times \mathcal{S}$.

Assumption (Identification Conditions)

1. *(Consistency)* $Y = Y(t)$ *whenever* $T = t \in \mathcal{T}$.
2. *(Ignorability)* $Y(t)$ *is conditionally independent of* T *given* $\boldsymbol{S}$ *for all* $t \in \mathcal{T}$.
3. ***(Positivity)*** *The conditional density satisfies* $p(t|\boldsymbol{s}) \geq p_{\min} > 0$ *for all* $(t, \boldsymbol{s}) \in \mathcal{T} \times \mathcal{S}$.

$$m(t) = \mathbb{E}\left[Y(t)\right] \overset{(*)}{=} \mathbb{E}\left\{\mathbb{E}\left[Y(t)|\boldsymbol{S}\right]\right\} \qquad \text{(*) Law of total expectation}$$

$$\overset{(**)}{=} \mathbb{E}\left\{\mathbb{E}\left[Y(t)|T = t, \boldsymbol{S}\right]\right\} \qquad \text{(**) Ignorability}$$

$$\overset{(***)}{=} \mathbb{E}\left[\mathbb{E}\left(Y|T = t, \boldsymbol{S}\right)\right] \qquad \text{(***) Consistency \& Positivity}$$

Assumption (Identification Conditions)

1. *(Consistency)* $Y = Y(t)$ *whenever* $T = t \in \mathcal{T}$.
2. *(Ignorability)* $Y(t)$ *is conditionally independent of* T *given* $\boldsymbol{S}$ *for all* $t \in \mathcal{T}$.
3. ***(Positivity)*** *The conditional density satisfies* $p(t|\boldsymbol{s}) \geq p_{\min} > 0$ *for all* $(t, \boldsymbol{s}) \in \mathcal{T} \times \mathcal{S}$.

$$
\begin{aligned}
m(t) = \mathbb{E}\left[Y(t)\right] &\overset{(*)}{=} \mathbb{E}\left\{\mathbb{E}\left[Y(t)|\boldsymbol{S}\right]\right\} && \text{(*) Law of total expectation} \\
&\overset{(**)}{=} \mathbb{E}\left\{\mathbb{E}\left[Y(t)|T = t, \boldsymbol{S}\right]\right\} && \text{(**) Ignorability} \\
&\overset{(***)}{=} \mathbb{E}\left[\mathbb{E}\left(Y|T = t, \boldsymbol{S}\right)\right] && \text{(***) Consistency \& Positivity}
\end{aligned}
$$

- The positivity condition is required in (***) for

$$\mu(t, \boldsymbol{s}) = \mathbb{E}\left(Y|T = t, \boldsymbol{S} = \boldsymbol{s}\right)$$

to be well-defined on $\mathcal{T} \times \mathcal{S}$.

There are three major strategies for estimating

$$m(t) = \mathbb{E}\left[Y(t)\right] = \mathbb{E}\left[\mu(t, \boldsymbol{S})\right] = \lim_{h \to 0} \mathbb{E}\left[\frac{Y \cdot K\left(\frac{T-t}{h}\right)}{h \cdot p(T|\boldsymbol{S})}\right]$$

from the data $\{(Y_i, T_i, \boldsymbol{S}_i)\}_{i=1}^n$, where $K : \mathbb{R} \to [0, \infty)$ is a kernel function.

There are three major strategies for estimating

$$m(t) = \mathbb{E}\left[Y(t)\right] = \mathbb{E}\left[\mu(t, \boldsymbol{S})\right] = \lim_{h\to 0} \mathbb{E}\left[\frac{Y \cdot K\left(\frac{T-t}{h}\right)}{h \cdot p(T|\boldsymbol{S})}\right]$$

from the data $\{(Y_i, T_i, \boldsymbol{S}_i)\}_{i=1}^n$, where $K : \mathbb{R} \to [0, \infty)$ is a kernel function.

1. **Regression Adjustment** (Robins, 1986; Gill and Robins, 2001):

$$\widehat{m}_{\mathrm{RA}}(t) = \frac{1}{n}\sum_{i=1}^n \widehat{\mu}(t, \boldsymbol{S}_i).$$

2. **Inverse Probability Weighting** (Hirano and Imbens, 2004):

$$\widehat{m}_{\mathrm{IPW}}(t) = \frac{1}{nh}\sum_{i=1}^n \frac{K\left(\frac{T_i-t}{h}\right)}{\widehat{p}(T_i|\boldsymbol{S}_i)} \cdot Y_i.$$

3. **Doubly Robust** (Kennedy et al., 2017; Colangelo and Lee, 2020).

There are three major strategies for estimating

$$m(t) = \mathbb{E}\left[Y(t)\right] = \mathbb{E}\left[\mu(t, \boldsymbol{S})\right] = \lim_{h \to 0} \mathbb{E}\left[\frac{Y \cdot K\left(\frac{T-t}{h}\right)}{h \cdot p(T|\boldsymbol{S})}\right]$$

from the data $\{(Y_i, T_i, \boldsymbol{S}_i)\}_{i=1}^{n}$, where $K : \mathbb{R} \to [0, \infty)$ is a kernel function.

1. **Regression Adjustment** (Robins, 1986; Gill and Robins, 2001):

$$\widehat{m}_{\mathrm{RA}}(t) = \frac{1}{n}\sum_{i=1}^{n} \widehat{\mu}(t, \boldsymbol{S}_i).$$

2. **Inverse Probability Weighting** (Hirano and Imbens, 2004):

$$\widehat{m}_{\mathrm{IPW}}(t) = \frac{1}{nh}\sum_{i=1}^{n} \frac{K\left(\frac{T_i - t}{h}\right)}{\widehat{p}(T_i|\boldsymbol{S}_i)} \cdot Y_i.$$

3. **Doubly Robust** (Kennedy et al., 2017; Colangelo and Lee, 2020).

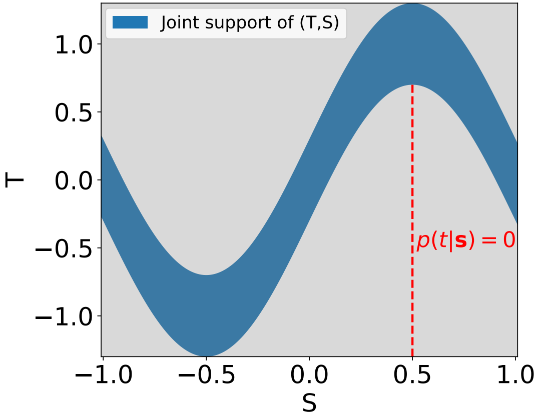
- **Issue:** Positivity is a strong assumption with continuous treatments!

Violation of the Positivity Condition

Assumption (Positivity Condition)

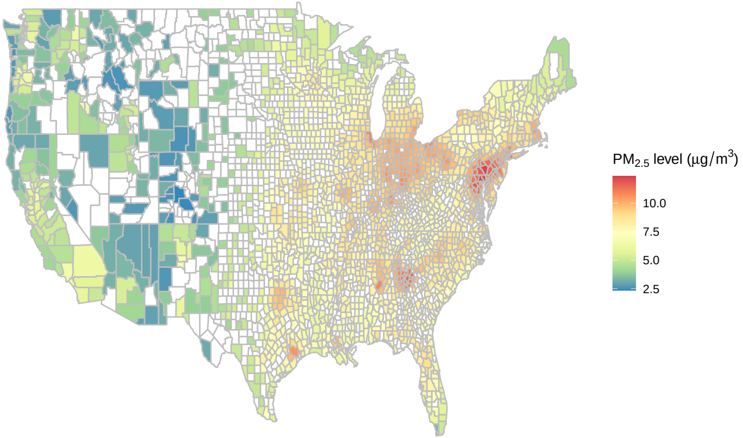
The conditional density $p(t|\mathbf{s})$ is uniformly bounded away from zero for all $(t, \mathbf{s}) \in \mathcal{T} \times \mathcal{S}$.

$$T = \sin(\pi S) + E, \quad E \sim \text{Unif}[-0.3, 0.3], \quad S \sim \text{Unif}[-1, 1], \quad \text{and} \quad E \perp\!\!\!\perp S.$$



- **Note:** $p(t|s) = 0$ in the gray regions, and the positivity condition fails.

$PM_{2.5}$ Distribution at the County Level



Average $PM_{2.5}$ levels from 1990 to 2010 in $n = 2132$ counties.

- T is $PM_{2.5}$ level, and S consists of the county location and socioeconomic factors.
- Only one or several $PM_{2.5}$ levels are available per county in the dataset, and the positivity condition is violated!

Highlight of Today's Talk

$$t \mapsto m(t) = \mathbb{E}\left[Y(t)\right] \quad \text{and} \quad t \mapsto \theta(t) = \frac{d}{dt}\mathbb{E}\left[Y(t)\right] \quad \text{for} \quad t \in \mathcal{T}.$$

1. The positivity condition may fail in some regions of $\mathcal{T} \times \mathcal{S}$.
   - We propose a new identification strategy for $m(t)$ and $\theta(t)$.

Highlight of Today's Talk

$$t \mapsto m(t) = \mathbb{E}\left[Y(t)\right] \quad \text{and} \quad t \mapsto \theta(t) = \frac{d}{dt}\mathbb{E}\left[Y(t)\right] \quad \text{for} \quad t \in \mathcal{T}.$$

1. The positivity condition may fail in some regions of $\mathcal{T} \times \mathcal{S}$.
   - We propose a new identification strategy for $m(t)$ and $\theta(t)$.
2. We derive a novel integral estimator $\widehat{m}_\theta(t)$ of $m(t)$ for all $t \in \mathcal{T}$.

Highlight of Today's Talk

$$t \mapsto m(t) = \mathbb{E}\left[Y(t)\right] \quad \text{and} \quad t \mapsto \theta(t) = \frac{d}{dt}\mathbb{E}\left[Y(t)\right] \quad \text{for} \quad t \in \mathcal{T}.$$

1. The positivity condition may fail in some regions of $\mathcal{T} \times \mathcal{S}$.
   - We propose a new identification strategy for $m(t)$ and $\theta(t)$.
2. We derive a novel integral estimator $\widehat{m}_\theta(t)$ of $m(t)$ for all $t \in \mathcal{T}$.
   - Construct a localized derivative estimator $\widehat{\theta}_C(t)$ of $\theta(t) = m'(t)$ around the observations $T_i, i = 1, ..., n$.
   - Extrapolate $\widehat{\theta}_C(t)$ to any treatment level of interest via the integration.

Highlight of Today's Talk

$$t \mapsto m(t) = \mathbb{E}\left[Y(t)\right] \quad \text{and} \quad t \mapsto \theta(t) = \frac{d}{dt}\mathbb{E}\left[Y(t)\right] \quad \text{for} \quad t \in \mathcal{T}.$$

1. The positivity condition may fail in some regions of $\mathcal{T} \times \mathcal{S}$.
   - We propose a new identification strategy for $m(t)$ and $\theta(t)$.
2. We derive a novel integral estimator $\widehat{m}_\theta(t)$ of $m(t)$ for all $t \in \mathcal{T}$.
   - Construct a localized derivative estimator $\widehat{\theta}_C(t)$ of $\theta(t) = m'(t)$ around the observations $T_i, i = 1, ..., n$.
   - Extrapolate $\widehat{\theta}_C(t)$ to any treatment level of interest via the integration.
   - Both $\widehat{m}_\theta(t)$ and $\widehat{\theta}_C(t)$ are consistent in $\mathcal{T}$ even when the positivity condition fails.

Highlight of Today's Talk

$$t \mapsto m(t) = \mathbb{E}\left[Y(t)\right] \quad \text{and} \quad t \mapsto \theta(t) = \frac{d}{dt}\mathbb{E}\left[Y(t)\right] \quad \text{for} \quad t \in \mathcal{T}.$$

1. The positivity condition may fail in some regions of $\mathcal{T} \times \mathcal{S}$.
   - We propose a new identification strategy for $m(t)$ and $\theta(t)$.
2. We derive a novel integral estimator $\widehat{m}_\theta(t)$ of $m(t)$ for all $t \in \mathcal{T}$.
   - Construct a localized derivative estimator $\widehat{\theta}_C(t)$ of $\theta(t) = m'(t)$ around the observations $T_i, i = 1, ..., n$.
   - Extrapolate $\widehat{\theta}_C(t)$ to any treatment level of interest via the integration.
   - Both $\widehat{m}_\theta(t)$ and $\widehat{\theta}_C(t)$ are consistent in $\mathcal{T}$ even when the positivity condition fails.
3. Nonparametric bootstrap inference with our proposed estimators $\widehat{m}_\theta(t)$ and $\widehat{\theta}_C(t)$ for $m(t)$ and $\theta(t)$ is asymptotically valid.

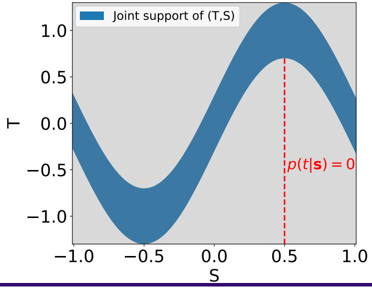
Identification

Assumption (Identification Conditions)

1. *(Consistency)* $Y = Y(t)$ *whenever* $T = t \in \mathcal{T}$.
2. *(Ignorability or Unconfoundedness)* $Y(t) \perp\!\!\!\perp T \,\big|\, \boldsymbol{S}$ *for all* $t \in \mathcal{T}$.
3. ***(Positivity)*** $p_{T|\boldsymbol{S}}(t|\boldsymbol{s}) \geq p_{\min} > 0$ *for all* $(t, \boldsymbol{s}) \in \mathcal{T} \times \mathcal{S}$.

The RA (or G-computation) formulae are given by

$$m(t) = \mathbb{E}\left[Y(t)\right] = \mathbb{E}\left[\mu(t, \boldsymbol{S})\right] \quad \text{and} \quad \theta(t) = \frac{d}{dt}\mathbb{E}\left[Y(t)\right] = \mathbb{E}\left[\frac{\partial}{\partial t}\mu(t, \boldsymbol{S})\right].$$

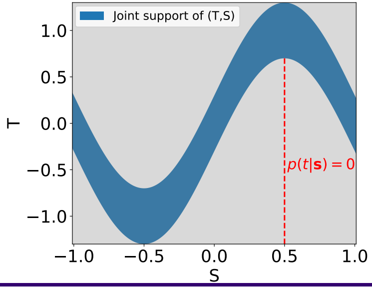


Assumption (Identification Conditions)

1. *(Consistency)* $Y = Y(t)$ *whenever* $T = t \in \mathcal{T}$.
2. *(Ignorability or Unconfoundedness)* $Y(t) \perp\!\!\!\perp T \,\big|\, \mathbf{S}$ *for all* $t \in \mathcal{T}$.
3. ***(Positivity)*** $p_{T|\mathbf{S}}(t|\mathbf{s}) \geq p_{\min} > 0$ *for all* $(t, \mathbf{s}) \in \mathcal{T} \times \mathcal{S}$.

The RA (or G-computation) formulae are given by

$$m(t) = \mathbb{E}\left[Y(t)\right] = \mathbb{E}\left[\mu(t, \mathbf{S})\right] \quad \text{and} \quad \theta(t) = \frac{d}{dt}\mathbb{E}\left[Y(t)\right] = \mathbb{E}\left[\frac{\partial}{\partial t}\mu(t, \mathbf{S})\right].$$



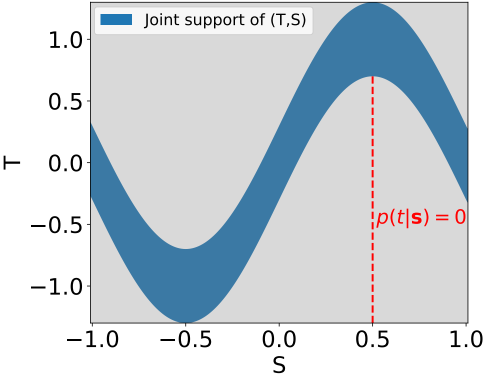
▶ **Identification Issue:** Without positivity,

$$\mu(t, \mathbf{s}) = \mathbb{E}\left(Y | T = t, \mathbf{S} = \mathbf{s}\right)$$

is *not well-defined* outside the support $\mathcal{E} \subset \mathcal{T} \times \mathcal{S}$ of the joint density $p(t, \mathbf{s})$.

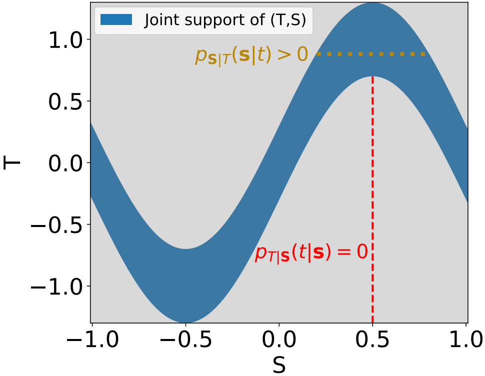
Assumption (Identification Conditions)

1. *(Consistency)* $Y = Y(t)$ *whenever* $T = t \in \mathcal{T}$.
2. *(Ignorability)* $Y(t)$ *is conditionally independent of* T *given* $\boldsymbol{S}$ *for all* $t \in \mathcal{T}$.
3. *(Treatment Variation)* $\text{Var}(T|\boldsymbol{S} = \boldsymbol{s}) > 0$ *for all* $\boldsymbol{s} \in \mathcal{S}$.



Assumption (Identification Conditions)

1. *(Consistency)* $Y = Y(t)$ *whenever* $T = t \in \mathcal{T}$.
2. *(Ignorability)* $Y(t)$ *is conditionally independent of* T *given* $\boldsymbol{S}$ *for all* $t \in \mathcal{T}$.
3. *(Treatment Variation)* $\mathrm{Var}(T|\boldsymbol{S} = \boldsymbol{s}) > 0$ *for all* $\boldsymbol{s} \in \mathcal{S}$.



Assumption (Extrapolation; Zhang et al. 2024)

Assume $(t, \boldsymbol{s}) \mapsto \mathbb{E}[Y(t)|\boldsymbol{S} = \boldsymbol{s}]$ *to be differentiable w.r.to* t *for any* $(t, \boldsymbol{s}) \in \mathcal{T} \times \mathcal{S}$ *with* $p_{\boldsymbol{S}|T}(\boldsymbol{s}|t) > 0$ *and*

$$\theta(t) = \frac{d}{dt}\mathbb{E}[Y(t)] = \mathbb{E}\left[\frac{\partial}{\partial t}\mathbb{E}[Y(t)|\boldsymbol{S}]\right] \stackrel{\star}{=} \mathbb{E}\left[\frac{\partial}{\partial t}\mathbb{E}[Y(t)|\boldsymbol{S}] \,\middle|\, T = t\right].$$

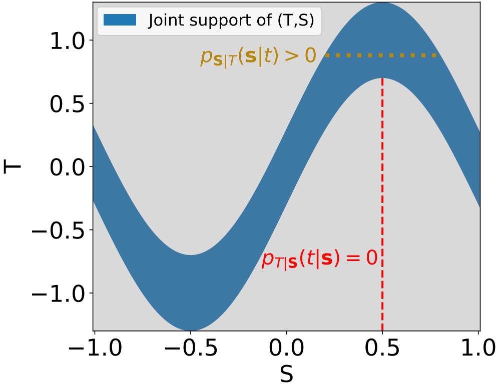
Additionally, it holds true that $\mathbb{E}(Y) = \mathbb{E}[m(T)]$.

Identification Strategy Without Positivity

If $\theta(t) = \frac{d}{dt}\mathbb{E}\left[Y(t)\right] = \mathbb{E}\left[\frac{\partial}{\partial t}\mathbb{E}\left[Y(t)|\boldsymbol{S}\right]\right] = \mathbb{E}\left[\frac{\partial}{\partial t}\mathbb{E}\left[Y(t)|\boldsymbol{S}\right]\Big|T = t\right]$ holds true, then

Identification Strategy Without Positivity

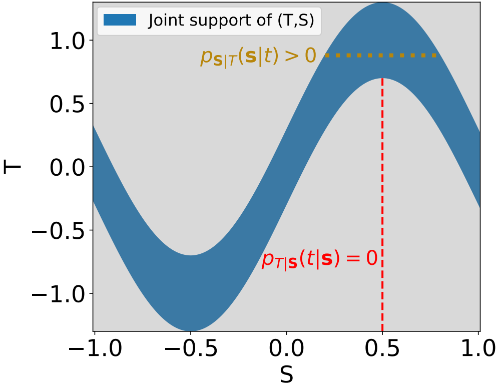
If $\theta(t) = \frac{d}{dt}\mathbb{E}\left[Y(t)\right] = \mathbb{E}\left[\frac{\partial}{\partial t}\mathbb{E}\left[Y(t)|\mathbf{S}\right]\right] = \mathbb{E}\left[\frac{\partial}{\partial t}\mathbb{E}\left[Y(t)|\mathbf{S}\right]\Big|T=t\right]$ holds true, then



$$
\begin{aligned}
\theta(t) &= \mathbb{E}\left[\frac{\partial}{\partial t}\mathbb{E}[Y(t)|\mathbf{S}]\Big|T=t\right] \\
&\overset{(*)}{=} \mathbb{E}\left[\frac{\partial}{\partial t}\mathbb{E}[Y(t)|T=t,\mathbf{S}]\Big|T=t\right] \\
&\overset{(**)}{=} \mathbb{E}\left[\frac{\partial}{\partial t}\mathbb{E}(Y|T=t,\mathbf{S})\Big|T=t\right] \\
&= \mathbb{E}\left[\frac{\partial}{\partial t}\mu(t,\mathbf{S})\Big|T=t\right] := \theta_C(t).
\end{aligned}
$$

(*) Ignorability; (**) Consistency.

If $\theta(t) = \frac{d}{dt}\mathbb{E}\left[Y(t)\right] = \mathbb{E}\left[\frac{\partial}{\partial t}\mathbb{E}\left[Y(t)|\boldsymbol{S}\right]\right] = \mathbb{E}\left[\frac{\partial}{\partial t}\mathbb{E}\left[Y(t)|\boldsymbol{S}\right]\Big|T=t\right]$ holds true, then

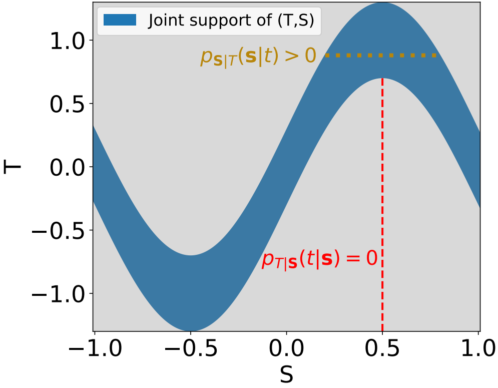


$$\begin{aligned}\theta(t) &= \mathbb{E}\left[\frac{\partial}{\partial t}\mathbb{E}[Y(t)|\boldsymbol{S}]\Big|T=t\right] \\ &\overset{(*)}{=} \mathbb{E}\left[\frac{\partial}{\partial t}\mathbb{E}[Y(t)|T=t,\boldsymbol{S}]\Big|T=t\right] \\ &\overset{(**)}{=} \mathbb{E}\left[\frac{\partial}{\partial t}\mathbb{E}(Y|T=t,\boldsymbol{S})\Big|T=t\right] \\ &= \mathbb{E}\left[\frac{\partial}{\partial t}\mu(t,\boldsymbol{S})\Big|T=t\right] := \theta_C(t).\end{aligned}$$

(*) Ignorability; (**) Consistency.

- By the fundamental theorem of calculus, $m(t) = m(T) + \int_T^t \overbrace{m'(u)}^{=\theta_C(u)}\,du$ so that

If $\theta(t) = \frac{d}{dt}\mathbb{E}\left[Y(t)\right] = \mathbb{E}\left[\frac{\partial}{\partial t}\mathbb{E}\left[Y(t)|\mathbf{S}\right]\right] = \mathbb{E}\left[\frac{\partial}{\partial t}\mathbb{E}\left[Y(t)|\mathbf{S}\right]\Big|T=t\right]$ holds true, then



$$
\begin{aligned}
\theta(t) &= \mathbb{E}\left[\frac{\partial}{\partial t}\mathbb{E}[Y(t)|\mathbf{S}]\Big|T=t\right] \\
&\overset{(*)}{=} \mathbb{E}\left[\frac{\partial}{\partial t}\mathbb{E}[Y(t)|T=t,\mathbf{S}]\Big|T=t\right] \\
&\overset{(**)}{=} \mathbb{E}\left[\frac{\partial}{\partial t}\mathbb{E}(Y|T=t,\mathbf{S})\Big|T=t\right] \\
&= \mathbb{E}\left[\frac{\partial}{\partial t}\mu(t,\mathbf{S})\Big|T=t\right] := \theta_C(t).
\end{aligned}
$$

(*) Ignorability; (**) Consistency.

- By the fundamental theorem of calculus, $m(t) = m(T) + \int_T^t \overbrace{m'(u)}^{=\theta_C(u)}\,du$ so that

$$
m(t) = \mathbb{E}\left[m(t)\right] = \mathbb{E}(Y) + \mathbb{E}\left\{\int_{u=T}^{u=t}\mathbb{E}\left[\frac{\partial}{\partial t}\mu(T,\mathbf{S})\Big|T=u\right]du\right\} \quad \text{for any } t\in\mathcal{T}.
$$

Example: Additive Confounding Model

Consider the additive confounding model, which is commonly assumed in spatial statistics (Paciorek, 2010; Schnell and Papadogeorgou, 2020; Gilbert et al., 2023):

$$Y(t) = \bar{m}(t) + \eta(\boldsymbol{S}) + \epsilon \quad \text{or} \quad Y = \bar{m}(T) + \eta(\boldsymbol{S}) + \epsilon. \tag{1}$$

- $\bar{m} : \mathcal{T} \to \mathbb{R}$ and $\eta : \mathcal{S} \to \mathbb{R}$ are deterministic functions.
- $\epsilon \in \mathbb{R}$ is an independent noise variable with $\mathbb{E}(\epsilon) = 0$ and $\mathrm{Var}(\epsilon) > 0$.
- $m(t) = \mathbb{E}[Y(t)] = \bar{m}(t) + \mathbb{E}[\eta(\boldsymbol{S})]$ and $\theta(t) = m'(t) = \frac{d}{dt}\mathbb{E}[Y(t)] = \bar{m}'(t)$.

Example: Additive Confounding Model

Consider the additive confounding model, which is commonly assumed in spatial statistics (Paciorek, 2010; Schnell and Papadogeorgou, 2020; Gilbert et al., 2023):

$$Y(t) = \bar{m}(t) + \eta(\boldsymbol{S}) + \epsilon \quad \text{or} \quad Y = \bar{m}(T) + \eta(\boldsymbol{S}) + \epsilon. \tag{1}$$

- $\bar{m} : \mathcal{T} \to \mathbb{R}$ and $\eta : \mathcal{S} \to \mathbb{R}$ are deterministic functions.
- $\epsilon \in \mathbb{R}$ is an independent noise variable with $\mathbb{E}(\epsilon) = 0$ and $\text{Var}(\epsilon) > 0$.
- $m(t) = \mathbb{E}[Y(t)] = \bar{m}(t) + \mathbb{E}[\eta(\boldsymbol{S})]$ and $\theta(t) = m'(t) = \frac{d}{dt}\mathbb{E}[Y(t)] = \bar{m}'(t)$.

Proposition (Proposition 2 in Zhang et al. 2024)

Under the additive confounding model (1), the extrapolation condition holds:

$$\theta(t) = \mathbb{E}\left[\frac{\partial}{\partial t}\mu(t, \boldsymbol{S}) \Big| T = t\right] = \theta_C(t) \quad \textit{and} \quad \mathbb{E}(Y) = \mathbb{E}[\bar{m}(T) + \eta(\boldsymbol{S})] = \mathbb{E}[m(T)].$$

Consider the additive confounding model, which is commonly assumed in spatial statistics (Paciorek, 2010; Schnell and Papadogeorgou, 2020; Gilbert et al., 2023):

$$Y(t) = \bar{m}(t) + \eta(\boldsymbol{S}) + \epsilon \quad \text{or} \quad Y = \bar{m}(T) + \eta(\boldsymbol{S}) + \epsilon. \tag{1}$$

- $\bar{m} : \mathcal{T} \to \mathbb{R}$ and $\eta : \mathcal{S} \to \mathbb{R}$ are deterministic functions.
- $\epsilon \in \mathbb{R}$ is an independent noise variable with $\mathbb{E}(\epsilon) = 0$ and $\mathrm{Var}(\epsilon) > 0$.
- $m(t) = \mathbb{E}[Y(t)] = \bar{m}(t) + \mathbb{E}[\eta(\boldsymbol{S})]$ and $\theta(t) = m'(t) = \frac{d}{dt}\mathbb{E}[Y(t)] = \bar{m}'(t)$.

Proposition (Proposition 2 in Zhang et al. 2024)

Under the additive confounding model (1), the extrapolation condition holds:

$$\theta(t) = \mathbb{E}\left[\frac{\partial}{\partial t}\mu(t,\boldsymbol{S})\Big| T = t\right] = \theta_C(t) \quad \textit{and} \quad \mathbb{E}(Y) = \mathbb{E}\left[\bar{m}(T) + \eta(\boldsymbol{S})\right] = \mathbb{E}\left[m(T)\right].$$

▶ **Drawback of (1):** The treatment effect is homogeneous for any $\boldsymbol{S} = \boldsymbol{s} \in \mathcal{S}$.

Estimation and Inference

Recall our identification formulae

$$m(t) = \mathbb{E}\left[Y + \int_{\tilde{t}=T}^{\tilde{t}=t} \theta_C(\tilde{t})\, d\tilde{t}\right] \quad \text{and} \quad \theta_C(t) = \mathbb{E}\left[\frac{\partial}{\partial t}\mu(t, \boldsymbol{S}) \middle| T = t\right] = \int \frac{\partial}{\partial t}\mu(t, \boldsymbol{s})\, d\mathrm{P}(\boldsymbol{s}|t).$$

Proposed Integral Estimator of $m(t)$

Recall our identification formulae

$$m(t) = \mathbb{E}\left[Y + \int_{\tilde{t}=T}^{\tilde{t}=t} \theta_C(\tilde{t})\, d\tilde{t}\right] \quad \text{and} \quad \theta_C(t) = \mathbb{E}\left[\frac{\partial}{\partial t}\mu(t, \boldsymbol{S}) \Big| T = t\right] = \int \frac{\partial}{\partial t}\mu(t, \boldsymbol{s})\, d\mathrm{P}(\boldsymbol{s}|t).$$

Our **integral estimator** of $m(t)$ is given by

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

and our **localized derivative** estimator of $\theta(t)$ is

$$\widehat{\theta}_C(t) = \int \widehat{\beta}_2(t, \boldsymbol{s})\, d\widehat{\mathrm{P}}(\boldsymbol{s}|t).$$

Recall our identification formulae

$$m(t) = \mathbb{E}\left[Y + \int_{\tilde{t}=T}^{\tilde{t}=t} \theta_C(\tilde{t})\, d\tilde{t}\right] \quad \text{and} \quad \theta_C(t) = \mathbb{E}\left[\frac{\partial}{\partial t}\mu(t,\mathbf{S}) \Big| T = t\right] = \int \frac{\partial}{\partial t}\mu(t,\mathbf{s})\, d\mathrm{P}(\mathbf{s}|t).$$

Our **integral estimator** of $m(t)$ is given by

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

and our **localized derivative** estimator of $\theta(t)$ is

$$\widehat{\theta}_C(t) = \int \widehat{\beta}_2(t,\mathbf{s})\, d\widehat{\mathrm{P}}(\mathbf{s}|t).$$

- $\beta_2(t,\mathbf{s}) := \frac{\partial}{\partial t}\mu(t,\mathbf{s})$ is fitted by (partial) local polynomial regression.
- $\mathrm{P}(\mathbf{s}|t)$ is estimated by Nadaraya-Watson conditional cumulative distribution function (CDF) estimator.

1. **Order q (Partial) Local Polynomial Regression** (Fan and Gijbels, 1996): Let $\widehat{\boldsymbol{\beta}}(t, \boldsymbol{s}) \in \mathbb{R}^{q+1}$ and $\widehat{\boldsymbol{\alpha}}(t, \boldsymbol{s}) \in \mathbb{R}^d$ be the minimizer of

$$\arg\min_{(\boldsymbol{\beta},\boldsymbol{\alpha})^T \in \mathbb{R}^{q+1+d}} \sum_{i=1}^{n} \left[Y_i - \sum_{j=0}^{q} \beta_j (T_i - t)^q - \sum_{\ell=1}^{d} \alpha_\ell (S_{i,\ell} - s_\ell) \right]^2 K_T\left(\frac{T_i - t}{h}\right) K_S\left(\frac{\boldsymbol{S}_i - \boldsymbol{s}}{b}\right).$$

- $K_T : \mathbb{R} \to [0, \infty), K_S : \mathbb{R}^d \to [0, \infty)$ are two symmetric kernel functions, and $h, b > 0$ are smoothing bandwidth parameters.
- The second component $\widehat{\beta}_2(t, \boldsymbol{s})$ is a consistent estimator of $\beta_2(t, \boldsymbol{s}) = \frac{\partial}{\partial t}\mu(t, \boldsymbol{s})$.

1. **Order q (Partial) Local Polynomial Regression** (Fan and Gijbels, 1996): Let $\widehat{\boldsymbol{\beta}}(t, \boldsymbol{s}) \in \mathbb{R}^{q+1}$ and $\widehat{\boldsymbol{\alpha}}(t, \boldsymbol{s}) \in \mathbb{R}^d$ be the minimizer of

$$\arg\min_{(\boldsymbol{\beta},\boldsymbol{\alpha})^T \in \mathbb{R}^{q+1+d}} \sum_{i=1}^{n} \left[Y_i - \sum_{j=0}^{q} \beta_j (T_i - t)^q - \sum_{\ell=1}^{d} \alpha_\ell (S_{i,\ell} - s_\ell) \right]^2 K_T\left(\frac{T_i - t}{h}\right) K_S\left(\frac{\boldsymbol{S}_i - \boldsymbol{s}}{b}\right).$$

   - $K_T : \mathbb{R} \to [0, \infty), K_S : \mathbb{R}^d \to [0, \infty)$ are two symmetric kernel functions, and $h, b > 0$ are smoothing bandwidth parameters.
   - The second component $\widehat{\beta}_2(t, \boldsymbol{s})$ is a consistent estimator of $\beta_2(t, \boldsymbol{s}) = \frac{\partial}{\partial t}\mu(t, \boldsymbol{s})$.

2. **Nadaraya-Watson conditional CDF Estimator** (Hall et al., 1999):

$$\widehat{P}_\hbar(\boldsymbol{s}|t) = \frac{\sum_{i=1}^n \mathbb{1}_{\{\boldsymbol{S}_i \leq \boldsymbol{s}\}} \cdot \bar{K}_T\left(\frac{T_i - t}{\hbar}\right)}{\sum_{j=1}^n \bar{K}_T\left(\frac{T_j - t}{\hbar}\right)}.$$

   - $\bar{K}_T : \mathbb{R} \to [0, \infty)$ is a kernel function and $\hbar > 0$ is its smoothing bandwidth parameter.

Combining two nuisance function estimators $\widehat{\beta}_2(t, \boldsymbol{s})$ and $\widehat{\mathrm{P}}(\boldsymbol{s}|t)$, we derive our **localized derivative estimator** of $\theta(t)$ as:

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

Our integral estimator takes the form

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

Combining two nuisance function estimators $\widehat{\beta}_2(t, \boldsymbol{s})$ and $\widehat{\mathrm{P}}(\boldsymbol{s}|t)$, we derive our **localized derivative estimator** of $\theta(t)$ as:

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

Our integral estimator takes the form

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

- Other methods can be applied to estimate $\frac{\partial}{\partial t}\mu(t, \boldsymbol{s})$ and $\mathrm{P}(\boldsymbol{s}|t)$.
- $\widehat{m}_\theta(t)$, under our kernel-based estimators, is a *linear smoother*.

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

Combining two nuisance function estimators $\widehat{\beta}_2(t, \boldsymbol{s})$ and $\widehat{\mathrm{P}}(\boldsymbol{s}|t)$, we derive our **localized derivative estimator** of $\theta(t)$ as:

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

Our integral estimator takes the form

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

- Other methods can be applied to estimate $\frac{\partial}{\partial t}\mu(t, \boldsymbol{s})$ and $\mathrm{P}(\boldsymbol{s}|t)$.
- $\widehat{m}_\theta(t)$, under our kernel-based estimators, is a *linear smoother*.

► **Issue:** The integral could be analytically difficult to compute.

Our integral estimator takes the form

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

▶ **Riemann Sum Approximation:** 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 each $j = 1, ..., n$ as:

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

Our integral estimator takes the form

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

▶ **Riemann Sum Approximation:** 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 each $j = 1, ..., n$ as:

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

- 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 $\widehat{m}_\theta(T_{(j+1)})$.

- The approximation error is at most $O_P\left(\frac{1}{n}\right)$, which is *asymptotically negligible*.

Nonparametric Bootstrap Inference

1. Compute $\widehat{m}_\theta(t)$ on the original data $\{(Y_i, T_i, \boldsymbol{S}_i)\}_{i=1}^n$.

Nonparametric Bootstrap Inference

1. Compute $\widehat{m}_\theta(t)$ on the original data $\{(Y_i, T_i, \boldsymbol{S}_i)\}_{i=1}^n$.
2. Generate B bootstrap samples $\left\{\left(Y_i^{*(b)}, T_i^{*(b)}, \boldsymbol{S}_i^{*(b)}\right)\right\}_{i=1}^n$ by sampling with replacement and compute $\widehat{m}_\theta^{*(b)}(t)$ for each $b = 1, ..., B$.

1. Compute $\widehat{m}_\theta(t)$ on the original data $\{(Y_i, T_i, \boldsymbol{S}_i)\}_{i=1}^n$.
2. Generate B bootstrap samples $\left\{\left(Y_i^{*(b)}, T_i^{*(b)}, \boldsymbol{S}_i^{*(b)}\right)\right\}_{i=1}^n$ by sampling with replacement and compute $\widehat{m}_\theta^{*(b)}(t)$ for each $b = 1, ..., B$.
3. Let $\alpha \in (0,1)$ be a pre-specified significance level.
   - For pointwise inference at $t_0 \in \mathcal{T}$, calculate the $1-\alpha$ quantile $\zeta_{1-\alpha}^*(t_0)$ of $\{D_1(t_0), ..., D_B(t_0)\}$, where $D_b(t_0) = \left|\widehat{m}_\theta^{*(b)}(t_0) - \widehat{m}_\theta(t_0)\right|$ for $b = 1, ..., B$.
   - For uniform inference on $m(t)$, compute the $1-\alpha$ quantile $\xi_{1-\alpha}^*$ of $\{D_{\mathrm{sup},1}, ..., D_{\mathrm{sup},B}\}$, where $D_{\mathrm{sup},b} = \sup_{t\in\mathcal{T}} \left|\widehat{m}_\theta^{*(b)}(t) - \widehat{m}_\theta(t)\right|$ for $b = 1, ..., B$.

Nonparametric Bootstrap Inference

1. Compute $\widehat{m}_\theta(t)$ on the original data $\{(Y_i, T_i, \boldsymbol{S}_i)\}_{i=1}^n$.
2. Generate B bootstrap samples $\left\{\left(Y_i^{*(b)}, T_i^{*(b)}, \boldsymbol{S}_i^{*(b)}\right)\right\}_{i=1}^n$ by sampling with replacement and compute $\widehat{m}_\theta^{*(b)}(t)$ for each $b = 1, ..., B$.
3. Let $\alpha \in (0,1)$ be a pre-specified significance level.
   - For pointwise inference at $t_0 \in \mathcal{T}$, calculate the $1-\alpha$ quantile $\zeta_{1-\alpha}^*(t_0)$ of $\{D_1(t_0), ..., D_B(t_0)\}$, where $D_b(t_0) = \left|\widehat{m}_\theta^{*(b)}(t_0) - \widehat{m}_\theta(t_0)\right|$ for $b = 1, ..., B$.
   - For uniform inference on $m(t)$, compute the $1-\alpha$ quantile $\xi_{1-\alpha}^*$ of $\{D_{\sup,1}, ..., D_{\sup,B}\}$, where $D_{\sup,b} = \sup_{t\in\mathcal{T}}\left|\widehat{m}_\theta^{*(b)}(t) - \widehat{m}_\theta(t)\right|$ for $b = 1, ..., B$.
4. Define the $1-\alpha$ confidence interval for $m(t_0)$ as:

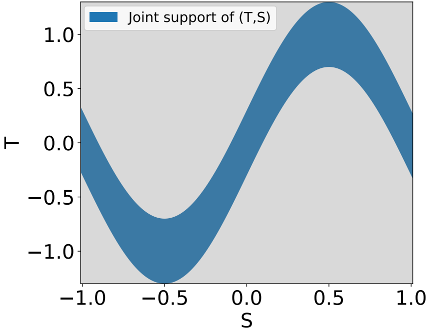
$$\left[\widehat{m}_\theta(t_0) - \zeta_{1-\alpha}^*(t_0),\ \widehat{m}_\theta(t_0) + \zeta_{1-\alpha}^*(t_0)\right]$$

and the simultaneous $1-\alpha$ confidence band for every $t \in \mathcal{T}$ as:

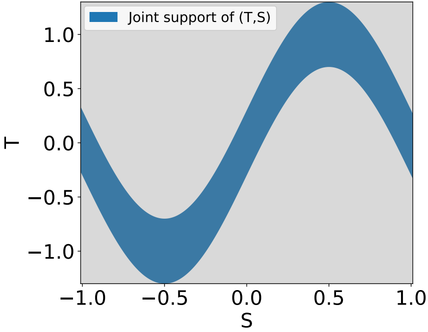
$$\left[\widehat{m}_\theta(t) - \xi_{1-\alpha}^*,\ \widehat{m}_\theta(t) + \xi_{1-\alpha}^*\right].$$

Asymptotic Theory

- The support $\mathcal{E}$ of (T, S) may not cover $\mathcal{T} \times \mathcal{S}$ without positivity.



- The support $\mathcal{E}$ of $(T, \boldsymbol{S})$ may not cover $\mathcal{T} \times \mathcal{S}$ without positivity.



- $\widehat{\theta}_C(t) = \int \widehat{\beta}_2(t, \boldsymbol{s})\, d\widehat{\mathrm{P}}_{\hbar}(\boldsymbol{s}|t)$ only requires $\widehat{\beta}_2(t, \boldsymbol{s})$ to be consistent in $\mathcal{E}$.

Lemma (Lemma 3 in Zhang et al. 2024)

Under some regularity conditions, as $h, b, \frac{\max\{h,b\}^4}{h} \to 0$ and $\frac{|\log(hb^d)|}{nh^3b^d} \to \infty$,

$$\sup_{(t,\boldsymbol{s})\in\mathcal{E}} \left| \widehat{\beta}_2(t, \boldsymbol{s}) - \frac{\partial}{\partial t}\mu(t, \boldsymbol{s}) \right| = O\left(h^q + b^2 + \frac{\max\{h, b\}^4}{h} \right) + O_P\left(\sqrt{\frac{|\log(hb^d)|}{nh^3b^d}} \right).$$

Combining with the consistency of $\widehat{\mathrm{P}}_{\hbar}(\boldsymbol{s}|t)$ via the technique in Fan et al. (1998), we have the following results.

Theorem (Theorem 4 in Zhang et al. 2024)

Let $\mathcal{T}' \subset \mathcal{T}$ be a compact set so that $p_T(t) \geq p_{T,\min} > 0$ for all $t \in \mathcal{T}'$. When $q = 2$ and $h, b, \hbar, \frac{\max\{h,b\}^4}{h} \to 0$ and $\frac{n \max\{h,\hbar\} b^d}{\log n}, \frac{n\hbar}{\log n} \to \infty$,

$$\sup_{t\in\mathcal{T}'} \left|\widehat{\theta}_C(t) - \theta_C(t)\right| = \underbrace{O\left(h^2 + b^2 + \frac{\max\{b,h\}^4}{h}\right)}_{\textit{Bias term}} + \underbrace{O_P\left(\sqrt{\frac{\log n}{nh^3}} + \hbar^2 + \sqrt{\frac{\log n}{n\hbar}}\right)}_{\textit{Stochastic variation}},$$

$$\sup_{t\in\mathcal{T}'} \left|\widehat{m}_\theta(t) - m(t)\right| = O\left(h^2 + b^2 + \frac{\max\{b,h\}^4}{h}\right) + O_P\left(\frac{1}{\sqrt{n}} + \sqrt{\frac{\log n}{nh^3}} + \hbar^2 + \sqrt{\frac{\log n}{n\hbar}}\right).$$

Uniform Rate of Convergence For the Integral Estimator

$$\sup_{t\in\mathcal{T}'} |\widehat{m}_\theta(t) - m(t)| = O\left(h^2 + b^2 + \frac{\max\{b,h\}^4}{h}\right) + O_P\left(\frac{1}{\sqrt{n}} + \sqrt{\frac{\log n}{nh^3}} + \hbar^2 + \sqrt{\frac{\log n}{n\hbar}}\right).$$

- Blue term: the estimation bias of local polynomial estimator $\widehat{\beta}_2(t, \boldsymbol{s})$.
- Orange term: additional bias of $\widehat{\beta}_2(t, \boldsymbol{s})$ at the boundary $\partial\mathcal{E}$.

Uniform Rate of Convergence For the Integral Estimator

$$\sup_{t\in\mathcal{T}'} \left|\widehat{m}_\theta(t) - m(t)\right| = O\left(h^2 + b^2 + \frac{\max\{b,h\}^4}{h}\right) + O_P\left(\frac{1}{\sqrt{n}} + \sqrt{\frac{\log n}{nh^3}} + \hbar^2 + \sqrt{\frac{\log n}{n\hbar}}\right).$$

- Blue term: the estimation bias of local polynomial estimator $\widehat{\beta}_2(t, \boldsymbol{s})$.
- Orange term: additional bias of $\widehat{\beta}_2(t, \boldsymbol{s})$ at the boundary $\partial\mathcal{E}$.
- Teal term: asymptotic rate from $\bar{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i$.
- Red term: stochastic variation of $\widehat{\beta}_2(t, \boldsymbol{s})$.

Uniform Rate of Convergence For the Integral Estimator

$$\sup_{t\in\mathcal{T}'} |\widehat{m}_\theta(t) - m(t)| = O\left(h^2 + b^2 + \frac{\max\{b,h\}^4}{h}\right) + O_P\left(\frac{1}{\sqrt{n}} + \sqrt{\frac{\log n}{nh^3}} + \hbar^2 + \sqrt{\frac{\log n}{n\hbar}}\right).$$

- Blue term: the estimation bias of local polynomial estimator $\widehat{\beta}_2(t, \boldsymbol{s})$.
- Orange term: additional bias of $\widehat{\beta}_2(t, \boldsymbol{s})$ at the boundary $\partial\mathcal{E}$.
- Teal term: asymptotic rate from $\bar{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i$.
- Red term: stochastic variation of $\widehat{\beta}_2(t, \boldsymbol{s})$.
- Cyan term: asymptotic rate from the Nadaraya-Watson conditional CDF estimator $\widehat{\mathrm{P}}_\hbar(\boldsymbol{s}|t)$.

Lemma (Lemma 5 in Zhang et al. 2024)

Under the same regularity conditions, if $h \asymp n^{-\frac{1}{\gamma}}$ *and* $\hbar \asymp n^{-\frac{1}{\varpi}}$ *for some* $\gamma \geq \varpi > 0$ *such that* $\frac{nh^5}{\log n} \to c_1$ *and* $\frac{n\hbar^5}{\log n} \to c_2$ *for some* $c_1, c_2 \geq 0$ *and* $\frac{n \max\{h,\hbar\} b^d}{\log n}, \frac{n\hbar}{\log n}, \frac{h^3 \log n}{\hbar}, \frac{nh^3\hbar^4}{\log n} \to \infty$ *as* $n \to \infty$*, then for any* $t \in \mathcal{T}'$*,*

$$\sqrt{nh^3}\left[\widehat{\theta}_C(t) - \theta(t)\right] = \mathbb{G}_n \bar{\varphi}_t + o_P(1), \quad \textit{and} \quad \sqrt{nh^3}\left[\widehat{m}_\theta(t) - m(t)\right] = \mathbb{G}_n \varphi_t + o_P(1),$$

where

$$\bar{\varphi}_t(Y, T, \boldsymbol{S}) = \frac{C_{K_T}\left[Y - \mu(T, \boldsymbol{S})\right]}{\sqrt{h} \cdot p_T(t)} \left(\frac{T-t}{h}\right) K_T\left(\frac{T-t}{h}\right)$$

and $\varphi_t\left(Y, T, \boldsymbol{S}\right) = \mathbb{E}_{T_1}\left[\int_{T_1}^{t} \bar{\varphi}_{\tilde{t}}(Y, T, \boldsymbol{S})\, d\tilde{t}\right]$ *with* $\mathbb{G}_n = \sqrt{n}\left(\mathbb{P}_n - \mathrm{P}\right)$*, where* $C_{K_T} > 0$ *is a constant that only depends on* K_T*.*

▶ **Note:** $\bar{\varphi}_t$ and φ_t are the IPW components of the *approximated* efficient influence functions.

Theorem (Theorems 6 and 7 in Zhang et al. 2024)

Under the same regularity conditions, if $h \asymp n^{-\frac{1}{\gamma}}$ *and* $b \lesssim \hbar \asymp n^{-\frac{1}{\varpi}}$ *for some* $\gamma \geq \varpi > 0$ *such that* $\frac{nh^{d+5}}{\log n} \to c_1$ *and* $\frac{n\hbar^5}{\log n} \to c_2$ *for some* $c_1, c_2 \geq 0$ *and* $\frac{\hbar}{h^3 \log n}, \hbar n^{\frac{1}{3}} \log n, \frac{\sqrt{n\hbar}}{\log n}, \frac{n \max\{h,\hbar\} b^d}{\log n} \to \infty$ *as* $n \to \infty$,

1.
$$\left|\sqrt{nh^3} \sup_{t\in\mathcal{T}} \left|\widehat{m}_\theta(t) - m(t)\right| - \sup_{t\in\mathcal{T}} \left|\mathbb{G}_n \varphi_t\right|\right| = O_P\left(\sqrt{nh^3 \max\{h,\hbar\}^4} + \sqrt{\frac{h^3 \log n}{\hbar}} + \frac{\log n}{\sqrt{n\hbar}} + \sqrt{\frac{\log n}{nb^d \hbar}}\right).$$

Theorem (Theorems 6 and 7 in Zhang et al. 2024)

Under the same regularity conditions, if $h \asymp n^{-\frac{1}{\gamma}}$ *and* $b \lesssim \hbar \asymp n^{-\frac{1}{\varpi}}$ *for some* $\gamma \geq \varpi > 0$ *such that* $\frac{nh^{d+5}}{\log n} \to c_1$ *and* $\frac{n\hbar^5}{\log n} \to c_2$ *for some* $c_1, c_2 \geq 0$ *and* $\frac{\hbar}{h^3 \log n}, \hbar n^{\frac{1}{3}} \log n, \frac{\sqrt{n\hbar}}{\log n}, \frac{n \max\{h,\hbar\} b^d}{\log n} \to \infty$ *as* $n \to \infty$,

1. $$\left| \sqrt{nh^3} \sup_{t \in \mathcal{T}} \left|\widehat{m}_\theta(t) - m(t)\right| - \sup_{t \in \mathcal{T}} \left|\mathbb{G}_n \varphi_t\right| \right| = O_P\left(\sqrt{nh^3 \max\{h,\hbar\}^4} + \sqrt{\frac{h^3 \log n}{\hbar}} + \frac{\log n}{\sqrt{n\hbar}} + \sqrt{\frac{\log n}{nb^d \hbar}} \right).$$

2. *there exists a mean-zero Gaussian process* $\mathbb{B}$ *such that*

$$\sup_{u \geq 0} \left| \mathrm{P}\left(\sqrt{nh^3} \sup_{t \in \mathcal{T}} \left|\widehat{m}_\theta(t) - m(t)\right| \leq u \right) - \mathrm{P}\left(\sup_{f \in \mathcal{F}} |\mathbb{B}(f)| \leq u \right) \right| = O\left(\left(\frac{\log^5 n}{nh^3}\right)^{\frac{1}{8}} + \left(\frac{\log^2 n}{nb^d \hbar}\right)^{\frac{3}{8}} \right).$$

Theorem (Theorems 6 and 7 in Zhang et al. 2024)

Under the same regularity conditions, if $h \asymp n^{-\frac{1}{\gamma}}$ *and* $b \lesssim \hbar \asymp n^{-\frac{1}{\varpi}}$ *for some* $\gamma \geq \varpi > 0$ *such that* $\frac{nh^{d+5}}{\log n} \to c_1$ *and* $\frac{n\hbar^5}{\log n} \to c_2$ *for some* $c_1, c_2 \geq 0$ *and* $\frac{\hbar}{h^3 \log n}, \hbar n^{\frac{1}{3}} \log n, \frac{\sqrt{n\hbar}}{\log n}, \frac{n \max\{h,\hbar\} b^d}{\log n} \to \infty$ *as* $n \to \infty$,

1. $$\left| \sqrt{nh^3} \sup_{t \in \mathcal{T}} \left| \widehat{m}_\theta(t) - m(t) \right| - \sup_{t \in \mathcal{T}} \left| \mathbb{G}_n \varphi_t \right| \right| = O_P\left(\sqrt{nh^3 \max\{h, \hbar\}^4} + \sqrt{\frac{h^3 \log n}{\hbar}} + \frac{\log n}{\sqrt{n\hbar}} + \sqrt{\frac{\log n}{nb^d \hbar}} \right).$$

2. *there exists a mean-zero Gaussian process* $\mathbb{B}$ *such that*

$$\sup_{u \geq 0} \left| \mathrm{P}\left(\sqrt{nh^3} \sup_{t \in \mathcal{T}} \left| \widehat{m}_\theta(t) - m(t) \right| \leq u \right) - \mathrm{P}\left(\sup_{f \in \mathcal{F}} \left| \mathbb{B}(f) \right| \leq u \right) \right| = O\left(\left(\frac{\log^5 n}{nh^3} \right)^{\frac{1}{8}} + \left(\frac{\log^2 n}{nb^d \hbar} \right)^{\frac{3}{8}} \right).$$

3. $$\sup_{u \geq 0} \left| \mathrm{P}\left(\sqrt{nh^3} \sup_{t \in \mathcal{T}} \left| \widehat{m}_\theta^*(t) - \widehat{m}_\theta(t) \right| \leq u \,\middle|\, \mathbb{U}_n \right) - \mathrm{P}\left(\sup_{f \in \mathcal{F}} \left| \mathbb{B}(f) \right| \leq u \right) \right| = O_P\left(\left(\frac{\log^5 n}{nh^3} \right)^{\frac{1}{8}} + \left(\frac{\log^2 n}{nb^d \hbar} \right)^{\frac{3}{8}} \right),$$

   where $\mathcal{F} = \{(v, x, z) \mapsto \varphi_t(v, x, z) : t \in \mathcal{T}\}$.

Remarks on Our Asymptotic Results

1. $\mathcal{F}$ is not Donsker because φ_t is not uniformly bounded as $h \to 0$.
    - However, $\widetilde{\mathcal{F}} = \left\{(v,x,z) \mapsto \sqrt{h^3} \cdot \varphi_t(v,x,z) : t \in \mathcal{T}'\right\}$ is of VC-type.
    - Gaussian approximation in Chernozhukov et al. (2014) can be applied to bound the difference between $\sup_{f\in\mathcal{F}} |\mathbb{G}_n(f)|$ and $\sup_{f\in\mathcal{F}} |\mathbb{B}(f)|$.

Remarks on Our Asymptotic Results

1. $\mathcal{F}$ is not Donsker because φ_t is not uniformly bounded as $h \to 0$.
   - However, $\widetilde{\mathcal{F}} = \left\{(v, x, \boldsymbol{z}) \mapsto \sqrt{h^3} \cdot \varphi_t(v, x, \boldsymbol{z}) : t \in \mathcal{T}'\right\}$ is of VC-type.
   - Gaussian approximation in Chernozhukov et al. (2014) can be applied to bound the difference between $\sup_{f\in\mathcal{F}} |\mathbb{G}_n(f)|$ and $\sup_{f\in\mathcal{F}} |\mathbb{B}(f)|$.

2. As long as $\mathrm{Var}(Y|T = t, \boldsymbol{S} = \boldsymbol{s}) \geq \sigma^2 > 0$, $\mathrm{Var}\left[\varphi_t(Y, T, \boldsymbol{S})\right]$ is a positive finite number.
   - The asymptotic linearity (or V-statistic) is non-degenerate.
   - Pointwise bootstrap confidence intervals are asymptotically valid.

Remarks on Our Asymptotic Results

1. $\mathcal{F}$ is not Donsker because φ_t is not uniformly bounded as $h \to 0$.
   - However, $\widetilde{\mathcal{F}} = \left\{(v, x, z) \mapsto \sqrt{h^3} \cdot \varphi_t(v, x, z) : t \in \mathcal{T}'\right\}$ is of VC-type.
   - Gaussian approximation in Chernozhukov et al. (2014) can be applied to bound the difference between $\sup_{f \in \mathcal{F}} |\mathbb{G}_n(f)|$ and $\sup_{f \in \mathcal{F}} |\mathbb{B}(f)|$.
2. As long as $\text{Var}(Y|T = t, \boldsymbol{S} = \boldsymbol{s}) \geq \sigma^2 > 0$, $\text{Var}\left[\varphi_t(Y, T, \boldsymbol{S})\right]$ is a positive finite number.
   - The asymptotic linearity (or V-statistic) is non-degenerate.
   - Pointwise bootstrap confidence intervals are asymptotically valid.
3. For the validity of uniform bootstrap confidence band, one can choose the bandwidths $h \asymp \hbar = O\left(n^{-\frac{1}{5}}\right)$ and $\left(\frac{\log n}{n}\right)^{\frac{4}{5d}} \lesssim b \lesssim n^{-\frac{1}{5}}$.
   - These orders align with the outputs from the usual bandwidth selection methods (Bashtannyk and Hyndman, 2001; Li and Racine, 2004).
   - No explicit undersmoothing is required!!

Case Study: $PM_{2.5}$ on CMR

$PM_{2.5}$ and CMRs Data Recap

FIPS	County name	Longitude	Latitude	PM2.5	CMR
1025	Clarke	-87.830772	31.676955	6.766443	379.421713
1061	Geneva	-85.839330	31.094869	8.254272	378.524698
1073	Jefferson	-86.896571	33.554343	10.825441	352.790427
1077	Lauderdale	-87.654117	34.901500	9.208783	332.594557
5085	Lonoke	-91.887917	34.754412	8.213144	365.061085
8045	Garfield	-107.903621	39.599420	2.601772	250.781477

1. The dataset (Wyatt et al., 2020a,b) contains the average annual CMRs (Y) and $PM_{2.5}$ levels (T) across $n = 2132$ U.S. counties over 1990-2010.

$PM_{2.5}$ and CMRs Data Recap

FIPS	County name	Longitude	Latitude	PM2.5	CMR
1025	Clarke	-87.830772	31.676955	6.766443	379.421713
1061	Geneva	-85.839330	31.094869	8.254272	378.524698
1073	Jefferson	-86.896571	33.554343	10.825441	352.790427
1077	Lauderdale	-87.654117	34.901500	9.208783	332.594557
5085	Lonoke	-91.887917	34.754412	8.213144	365.061085
8045	Garfield	-107.903621	39.599420	2.601772	250.781477

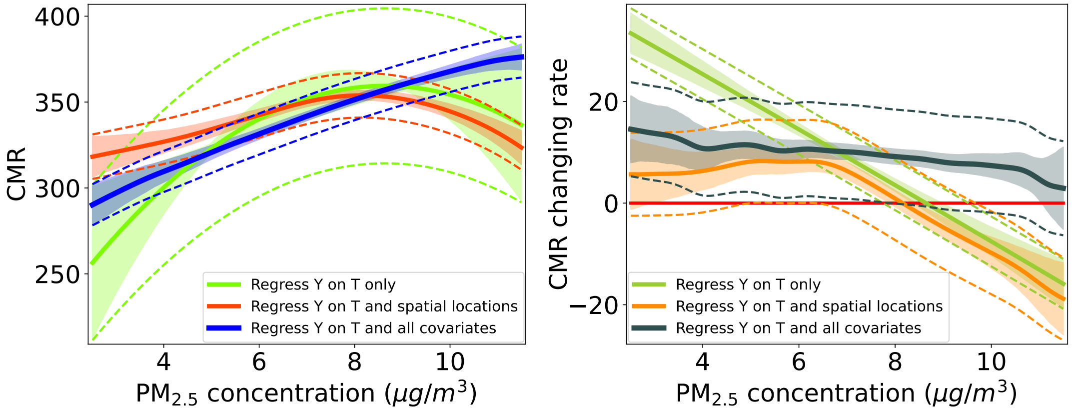
1. The dataset (Wyatt et al., 2020a,b) contains the average annual CMRs (Y) and $PM_{2.5}$ levels (T) across $n = 2132$ U.S. counties over 1990-2010.
2. The covariate vector $\boldsymbol{S} \in \mathbb{R}^{10}$ consists of two parts:
   - 2 spatial confounders: latitude and longitude of each county.
   - 8 county-level socioeconomic factors acquired from the US census.

$PM_{2.5}$ and CMRs Data Recap

FIPS	County name	Longitude	Latitude	PM2.5	CMR
1025	Clarke	-87.830772	31.676955	6.766443	379.421713
1061	Geneva	-85.839330	31.094869	8.254272	378.524698
1073	Jefferson	-86.896571	33.554343	10.825441	352.790427
1077	Lauderdale	-87.654117	34.901500	9.208783	332.594557
5085	Lonoke	-91.887917	34.754412	8.213144	365.061085
8045	Garfield	-107.903621	39.599420	2.601772	250.781477

1. The dataset (Wyatt et al., 2020a,b) contains the average annual CMRs (Y) and $PM_{2.5}$ levels (T) across $n = 2132$ U.S. counties over 1990-2010.
2. The covariate vector $\boldsymbol{S} \in \mathbb{R}^{10}$ consists of two parts:
   - 2 spatial confounders: latitude and longitude of each county.
   - 8 county-level socioeconomic factors acquired from the US census.
3. Focus on the values of $PM_{2.5}$ between 2.5 $\mu g/m^3$ and 11.5 $\mu g/m^3$ to avoid boundary effects (Takatsu and Westling, 2022).

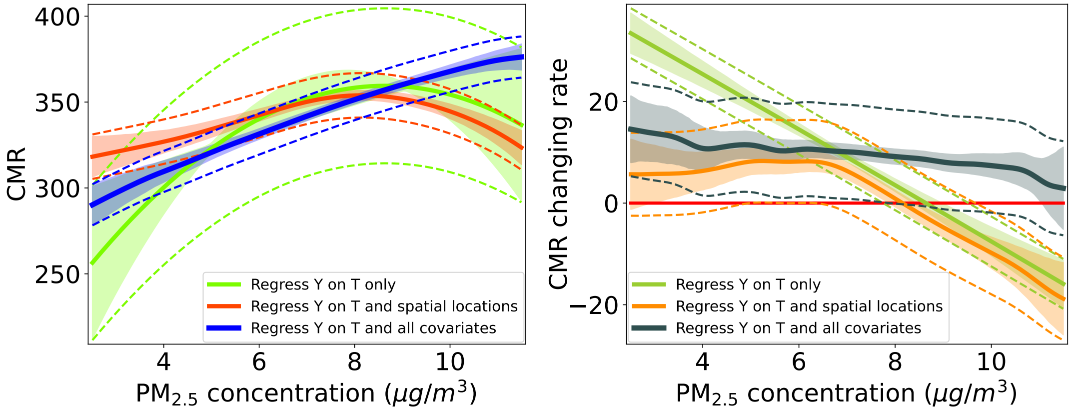
Effect of $PM_{2.5}$ on the Cardiovascular Mortality Rate (CMR)



Shaded areas: 95% pointwise confidence intervals; **Regions between dashed lines:** 95% uniform confidence bands.

- We compare three models:
  1. Regress Y on T alone via local quadratic regression.
  2. Regress Y on T with spatial locations.
  3. Regress Y on T with both spatial and socioeconomic covariates.

Effect of $PM_{2.5}$ on the Cardiovascular Mortality Rate (CMR)



Shaded areas: 95% pointwise confidence intervals; **Regions between dashed lines:** 95% uniform confidence bands.

- We compare three models:
  1. Regress Y on T alone via local quadratic regression.
  2. Regress Y on T with spatial locations.
  3. Regress Y on T with both spatial and socioeconomic covariates.
- For model 3, the increasing trends are **significant** when $PM_{2.5} < 8\,\mu g/m^3$.

Discussion

We study nonparametric inference on the dose-response curve $m(t) = \mathbb{E}[Y(t)]$ and its derivative $\theta(t) = \frac{d}{dt}\mathbb{E}[Y(t)]$ without the **positivity** condition.

- Our key technique relies on two pillars in calculus:

$$\underbrace{\theta(t) = \mathbb{E}\left[\frac{\partial}{\partial t}\mu(t, \mathbf{S}) \Big| T = t\right]}_{\textbf{Differentiation}} \quad \text{and} \quad \underbrace{m(t) = \mathbb{E}\left[Y + \int_{u=T}^{u=t} \theta(u)\, du\right]}_{\textbf{Integration}}.$$

We study nonparametric inference on the dose-response curve $m(t) = \mathbb{E}[Y(t)]$ and its derivative $\theta(t) = \frac{d}{dt}\mathbb{E}[Y(t)]$ without the **positivity** condition.

- Our key technique relies on two pillars in calculus:

$$\underbrace{\theta(t) = \mathbb{E}\left[\frac{\partial}{\partial t}\mu(t, \mathbf{S}) \Big| T = t\right]}_{\textbf{Differentiation}} \quad \text{and} \quad \underbrace{m(t) = \mathbb{E}\left[Y + \int_{u=T}^{u=t} \theta(u)\, du\right]}_{\textbf{Integration}}.$$

- Both estimators are consistent without the positivity condition.

Summary of Today's Talk

We study nonparametric inference on the dose-response curve $m(t) = \mathbb{E}[Y(t)]$ and its derivative $\theta(t) = \frac{d}{dt}\mathbb{E}[Y(t)]$ without the **positivity** condition.

- Our key technique relies on two pillars in calculus:

$$\underbrace{\theta(t) = \mathbb{E}\left[\frac{\partial}{\partial t}\mu(t, \boldsymbol{S}) \Big| T = t\right]}_{\textbf{Differentiation}} \quad \text{and} \quad \underbrace{m(t) = \mathbb{E}\left[Y + \int_{u=T}^{u=t} \theta(u)\, du\right]}_{\textbf{Integration}}.$$

- Both estimators are consistent without the positivity condition.
- Our integration idea opens a new direction for causal inference with continuous treatments under violations of positivity!

1. **IPW and Doubly Robust Estimation:** Our estimators are based on regression adjustment forms. Can we generalize to IPW and doubly robust forms (Zhang and Chen, 2025)?

1. **IPW and Doubly Robust Estimation:** Our estimators are based on regression adjustment forms. Can we generalize to IPW and doubly robust forms (Zhang and Chen, 2025)?

2. **Violation of Ignorability:** Can we conduct sensitivity analysis on unmeasured confounding (Chernozhukov et al., 2022)?

3. **High-Dimensional Confounders:** $\left(\frac{\log n}{n}\right)^{\frac{4}{5d}} \lesssim b \lesssim n^{-\frac{1}{5}}$ only works when $d < 5$. Can we use additive models (Guo et al., 2019) to resolve the dimensionality issue?

$$\mathbb{E}\left(Y|T = t, \boldsymbol{S} = \boldsymbol{s}, \boldsymbol{Z} = \boldsymbol{z}\right) = m(t) + \eta(\boldsymbol{s}) + \sum_{j=1}^{d'} g_j(\boldsymbol{z}_j) \quad \text{with} \quad \boldsymbol{z} \in \mathbb{R}^{d'}, d' \gg d.$$

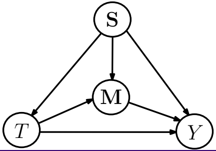
1. **IPW and Doubly Robust Estimation:** Our estimators are based on regression adjustment forms. Can we generalize to IPW and doubly robust forms (Zhang and Chen, 2025)?

2. **Violation of Ignorability:** Can we conduct sensitivity analysis on unmeasured confounding (Chernozhukov et al., 2022)?

3. **High-Dimensional Confounders:** $\left(\frac{\log n}{n}\right)^{\frac{4}{5d}} \lesssim b \lesssim n^{-\frac{1}{5}}$ only works when $d < 5$. Can we use additive models (Guo et al., 2019) to resolve the dimensionality issue?

$$\mathbb{E}\left(Y|T=t, \boldsymbol{S}=\boldsymbol{s}, \boldsymbol{Z}=\boldsymbol{z}\right) = m(t) + \eta(\boldsymbol{s}) + \sum_{j=1}^{d'} g_j(\boldsymbol{z}_j) \quad \text{with} \quad \boldsymbol{z} \in \mathbb{R}^{d'}, d' \gg d.$$

4. **Mediation Analysis:** Can we generalize our strategies for the estimation of direct and indirect causal effects (Huber et al., 2020; Xu et al., 2021) without positivity?



Thank you!

More details can be found in

[1] Y. Zhang, Y.-C. Chen, and A. Giessing. Nonparametric Inference on Dose-Response Curves Without the Positivity Condition. *arXiv preprint*, 2024. https://arxiv.org/abs/2405.09003.

All the code and data are available at
https://github.com/zhangyk8/npDoseResponse/tree/main.

Python Package: npDoseResponse and R Package: npDoseResponse.

[2] Y. Zhang and Y.-C. Chen. Doubly Robust Inference on Causal Derivative Effects for Continuous Treatments. *arXiv preprint*, 2025. https://arxiv.org/abs/2501.06969.

D. M. Bashtannyk and R. J. Hyndman. Bandwidth selection for kernel conditional density estimation. *Computational Statistics & Data Analysis*, 36(3):279–298, 2001.

S. Basith, B. Manavalan, T. H. Shin, C. B. Park, W.-S. Lee, J. Kim, and G. Lee. The impact of fine particulate matter 2.5 on the cardiovascular system: a review of the invisible killer. *Nanomaterials*, 12(15):2656, 2022.

J. E. Chacón, T. Duong, and M. Wand. Asymptotics for general multivariate kernel density derivative estimators. *Statistica Sinica*, pages 807–840, 2011.

Y.-C. Chen, C. R. Genovese, and L. Wasserman. A comprehensive approach to mode clustering. *Electronic Journal of Statistics*, 10(1):210 – 241, 2016.

V. Chernozhukov, D. Chetverikov, and K. Kato. Gaussian approximation of suprema of empirical processes. *The Annals of Statistics*, 42(4):1564–1597, 2014.

V. Chernozhukov, C. Cinelli, W. Newey, A. Sharma, and V. Syrgkanis. Long story short: Omitted variable bias in causal machine learning. Technical report, National Bureau of Economic Research, 2022.

K. Colangelo and Y.-Y. Lee. Double debiased machine learning nonparametric inference with continuous treatments. *arXiv preprint arXiv:2004.03036*, 2020.

J. Fan and I. Gijbels. *Local polynomial modelling and its applications*, volume 66. Chapman & Hall/CRC, 1996.

J. Fan, W. Härdle, and E. Mammen. Direct estimation of low-dimensional components in additive models. *The Annals of Statistics*, 26(3):943–971, 1998.

T. Gasser and H.-G. Müller. Estimating regression functions and their derivatives by the kernel method. *Scandinavian Journal of Statistics*, pages 171–185, 1984.

B. Gilbert, A. Datta, J. A. Casey, and E. L. Ogburn. A causal inference framework for spatial confounding. *arXiv preprint arXiv:2112.14946*, 2023.

Reference

R. D. Gill and J. M. Robins. Causal inference for complex longitudinal data: the continuous case. *Annals of Statistics*, 29(6):1785–1811, 2001.

Z. Guo, W. Yuan, and C.-H. Zhang. Decorrelated local linear estimator: Inference for non-linear effects in high-dimensional additive models. *arXiv preprint arXiv:1907.12732*, 2019.

P. Hall, R. C. Wolff, and Q. Yao. Methods for estimating a conditional distribution function. *Journal of the American Statistical Association*, 94(445):154–163, 1999.

K. Hirano and G. W. Imbens. *The Propensity Score with Continuous Treatments*, chapter 7, pages 73–84. John Wiley & Sons, Ltd, 2004.

M. Huber, Y.-C. Hsu, Y.-Y. Lee, and L. Lettry. Direct and indirect effects of continuous treatments based on generalized propensity score weighting. *Journal of Applied Econometrics*, 35(7):814–840, 2020.

E. Kammann and M. P. Wand. Geoadditive models. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 52 (1):1–18, 2003.

E. H. Kennedy, Z. Ma, M. D. McHugh, and D. S. Small. Nonparametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(4):1229–1245, 2017.

Q. Li and J. Racine. Cross-validated local linear nonparametric regression. *Statistica Sinica*, pages 485–512, 2004.

M. R. Miller and D. E. Newby. Air pollution and cardiovascular disease: car sick. *Cardiovascular Research*, 116(2): 279–294, 2020.

C. J. Paciorek. The importance of scale for spatial-confounding bias and precision of spatial regression estimators. *Statistical Science*, 25(1):107–125, 2010.

Reference

J. Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling*, 7(9-12):1393–1512, 1986.

P. Schnell and G. Papadogeorgou. Mitigating unobserved spatial confounding when estimating the effect of supermarket access on cardiovascular disease deaths. *Annals of Applied Statistics*, 14:2069–2095, 12 2020.

K. Takatsu and T. Westling. Debiased inference for a covariate-adjusted regression function. *arXiv preprint arXiv:2210.06448*, 2022.

H. Thaden and T. Kneib. Structural equation models for dealing with spatial confounding. *The American Statistician*, 72(3):239–252, 2018.

L. H. Wyatt, G. C. Peterson, T. J. Wade, L. M. Neas, and A. G. Rappold. The contribution of improved air quality to reduced cardiovascular mortality: Declines in socioeconomic differences over time. *Environment international*, 136:105430, 2020a.

L. H. Wyatt, G. C. L. Peterson, T. J. Wade, L. M. Neas, and A. G. Rappold. Annual pm2.5 and cardiovascular mortality rate data: Trends modified by county socioeconomic status in 2,132 us counties. *Data in Brief*, 30:105318, 2020b.

Y. Xu, N. Sani, A. Ghassami, and I. Shpitser. Multiply robust causal mediation analysis with continuous treatments. *arXiv preprint arXiv:2105.09254*, 2021.

L. Yang and R. Tschernig. Multivariate bandwidth selection for local linear regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 61(4):793–815, 1999.

Y. Zhang and Y.-C. Chen. Doubly robust inference on causal derivative effects for continuous treatments. *arXiv preprint arXiv:2501.06969*, 2025.

Y. Zhang, Y.-C. Chen, and A. Giessing. Nonparametric inference on dose-response curves without the positivity condition. *arXiv preprint arXiv:2405.09003*, 2024.

Let $\mathcal{E} \subset \mathcal{T} \times \mathcal{S}$ be the support of $p(t, \boldsymbol{s})$, $\mathcal{E}^\circ$ be the interior of $\mathcal{E}$, and $\partial\mathcal{E}$ be the boundary of $\mathcal{E}$.

1. For any $(t, \boldsymbol{s}) \in \mathcal{E}^\circ$, $\mu(t, \boldsymbol{s})$ is at least $(q+1)$ times continuously differentiable with respect to t and at least four times continuously differentiable with respect to $\boldsymbol{s}$. All these partial derivatives of $\mu(t, \boldsymbol{s})$ are continuous up to the boundary $\partial\mathcal{E}$. Furthermore, $\mu(t, \boldsymbol{s})$ and the partial derivatives are uniformly bounded on $\mathcal{E}$. Finally, there exist absolute constants $\sigma, A_0 > 0$ such that $\mathrm{Var}(Y|T = t, \boldsymbol{S} = \boldsymbol{s}) = \sigma^2$ and $\mathbb{E}|Y|^4 < A_0 < \infty$ uniformly in $\mathcal{E}$.
2. $p(t, \boldsymbol{s})$ is bounded and at least twice continuously differentiable with bounded partial derivatives up to the second order on $\mathcal{E}^\circ$. All these partial derivatives of $p(t, \boldsymbol{s})$ are continuous up to the boundary $\partial\mathcal{E}$. Furthermore, $\mathcal{E}$ is compact and $p(t, \boldsymbol{s})$ is uniformly bounded away from 0 on $\mathcal{E}$. Finally, the marginal density $p_T(t)$ of T is non-degenerate, *i.e.*, its support $\mathcal{T}$ has a nonempty interior.

Regularity Assumptions (Boundary Conditions)

3. There exists some constants $r_1, r_2 \in (0,1)$ such that for any $(t, \boldsymbol{s}) \in \mathcal{E}$ and all $\delta \in (0, r_1]$, there is a point $(t', \boldsymbol{s}') \in \mathcal{E}$ satisfying

$$\mathcal{B}\left((t', \boldsymbol{s}'),\, r_2\delta\right) \subset \mathcal{B}\left((t, \boldsymbol{s}),\, \delta\right) \cap \mathcal{E},$$

   where

$$\mathcal{B}((t, \boldsymbol{s}),\, r) = \left\{(t_1, \boldsymbol{s}_1) \in \mathbb{R}^{d+1} : ||(t_1 - t, \boldsymbol{s}_1 - \boldsymbol{s})||_2 \leq r\right\}$$

   with $||\cdot||_2$ being the standard Euclidean norm.

4. For any $(t, \boldsymbol{s}) \in \partial\mathcal{E}$, the boundary of $\mathcal{E}$, it satisfies that $\frac{\partial}{\partial t} p(t, \boldsymbol{s}) = \frac{\partial}{\partial s_j} p(t, \boldsymbol{s}) = 0$ and $\frac{\partial^2}{\partial s_j^2} \mu(t, \boldsymbol{s}) = 0$ for all $j = 1, ..., d$.

5. For any $\delta > 0$, the Lebesgue measure of the set $\partial\mathcal{E} \oplus \delta$ satisfies $|\partial\mathcal{E} \oplus \delta| \leq A_1 \cdot \delta$ for some absolute constant $A_1 > 0$, where

$$\partial\mathcal{E} \oplus \delta = \left\{\boldsymbol{z} \in \mathbb{R}^{d+1} : \inf_{\boldsymbol{x} \in \partial\mathcal{E}} ||\boldsymbol{z} - \boldsymbol{x}||_2 \leq \delta\right\}.$$

Regularity Assumptions (Kernel Conditions)

6. $K_T : \mathbb{R} \to [0, \infty)$ and $K_S : \mathbb{R}^d \to [0, \infty)$ are compactly supported and Lispchitz continuous kernels such that $\int_{\mathbb{R}} K_T(t)\, dt = \int_{\mathbb{R}^d} K_S(\boldsymbol{s})\, d\boldsymbol{s} = 1$, $K_T(t) = K_T(-t)$, and K_S is radially symmetric with $\int \boldsymbol{s} \cdot K_S(\boldsymbol{s}) d\boldsymbol{s} = \boldsymbol{0}$. In addition, for all $j = 1, 2, ...,$ and $\ell = 1, ..., d$,

$$\kappa_j^{(T)} := \int_{\mathbb{R}} u^j K_T(u)\, du < \infty, \quad \nu_j^{(T)} := \int_{\mathbb{R}} u^j K_T^2(u)\, du < \infty,$$
$$\kappa_{j,\ell}^{(S)} := \int_{\mathbb{R}^d} u_\ell^j K_S(\boldsymbol{u})\, d\boldsymbol{u} < \infty, \quad \text{and} \quad \nu_{j,k}^{(S)} := \int_{\mathbb{R}^d} u_\ell^j K_S^2(\boldsymbol{u})\, d\boldsymbol{u} < \infty.$$

Finally, both K_T and K_S are second-order kernels, *i.e.*, $\kappa_2^{(T)} > 0$ and $\kappa_{2,\ell}^{(S)} > 0$ for all $\ell = 1, ..., d$.

7. Let $\mathcal{K}_{q,d} = \left\{ (y, \boldsymbol{z}) \mapsto \left(\frac{y-t}{h}\right)^\ell \left(\frac{z_i - s_i}{b}\right)^{k_1} \left(\frac{z_j - s_j}{b}\right)^{k_2} K_T\left(\frac{y-t}{h}\right) K_S\left(\frac{\boldsymbol{z}-\boldsymbol{s}}{b}\right) : (t, \boldsymbol{s}) \in \mathcal{T} \times \mathcal{S}; i, j = 1, ..., d; \ell = 0, ..., 2q; k_1, k_2 = 0, 1; h, b > 0 \right\}$. It holds that $\mathcal{K}_{q,d}$ is a bounded VC-type class of measurable functions on $\mathbb{R}^{d+1}$.

Regularity Assumptions (Kernel Conditions)

8. The function $\bar{K}_T : \mathbb{R} \to [0, \infty)$ is a second-order, Lipschitz continuous, and symmetric kernel with a compact support, *i.e.*, $\int_{\mathbb{R}} \bar{K}_T(t)\, dt = 1$, $\bar{K}_T(t) = \bar{K}_T(-t)$, and $\int_{\mathbb{R}} t^2 \bar{K}_T(t)\, dt \in (0, \infty)$.

9. Let $\bar{\mathcal{K}} = \left\{ y \mapsto \bar{K}_T\left(\frac{y-t}{\hbar}\right) : t \in \mathcal{T}, \hbar > 0 \right\}$. It holds that $\bar{\mathcal{K}}$ is a bounded VC-type class of measurable functions on $\mathbb{R}$.

Recall that the class $\mathcal{G}$ of measurable functions on $\mathbb{R}^{d+1}$ is VC-type if there exist constants $A_2, \upsilon_2 > 0$ such that for any $0 < \epsilon < 1$,

$$\sup_Q N\left(\mathcal{G}, L_2(Q), \epsilon \,||G||_{L_2(Q)}\right) \leq \left(\frac{A_2}{\epsilon}\right)^{\upsilon_2},$$

where $N\left(\mathcal{G}, L_2(Q), \epsilon \,||G||_{L_2(Q)}\right)$ is the $\epsilon \,||G||_{L_2(Q)}$-covering number of the (semi-)metric space $\left(\mathcal{G}, ||\cdot||_{L_2(Q)}\right)$, Q is any probability measure on $\mathbb{R}^{d+1}$, G is an envelope function of $\mathcal{G}$, and $||G||_{L_2(Q)}$ is defined as $\left[\int_{\mathbb{R}^{d+1}} [G(x)]^2\, dQ(x)\right]^{\frac{1}{2}}$.

Simulation Setup for Estimating $m(t)$ and $\theta(t)$ Without Positivity

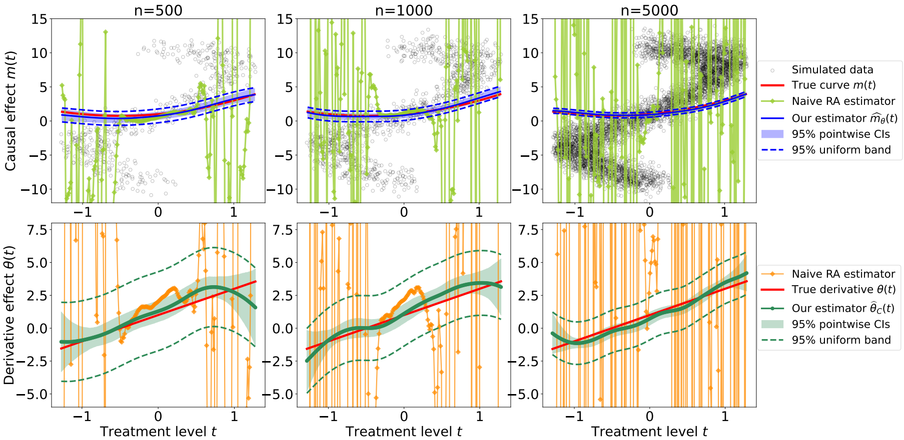
- Use the Epanechnikov kernel for K_T and K_S (with the product kernel technique) and Gaussian kernel for $\bar{K}_T$.
- Select the bandwidth parameters $h, b > 0$ by modifying the rule-of-thumb method in Yang and Tschernig (1999).
- Set the bandwidth parameter $\hbar > 0$ to the normal reference rule in Chacón et al. (2011); Chen et al. (2016).
- Set the bootstrap resampling time $B = 1000$ and the nominal level for confidence intervals or bands to 95%.
- Compare our proposed estimators with the regression adjustment estimators under the same choices of bandwidth parameters:

$$\widehat{m}_{\mathrm{RA}}(t) = \frac{1}{n}\sum_{i=1}^{n} \widehat{\mu}(t, \boldsymbol{S}_i) \quad \text{and} \quad \widehat{\theta}_{\mathrm{RA}}(t) = \frac{1}{n}\sum_{i=1}^{n} \widehat{\beta}_2(t, \boldsymbol{S}_i).$$

Generate i.i.d. observations $\{(Y_i, T_i, \boldsymbol{S}_i)\}_{i=1}^{2000}$ from

$$Y = T^2 + T + 1 + 10S + \epsilon, \quad T = \sin(\pi S) + E, \quad \text{and} \quad S \sim \text{Uniform}[-1, 1].$$

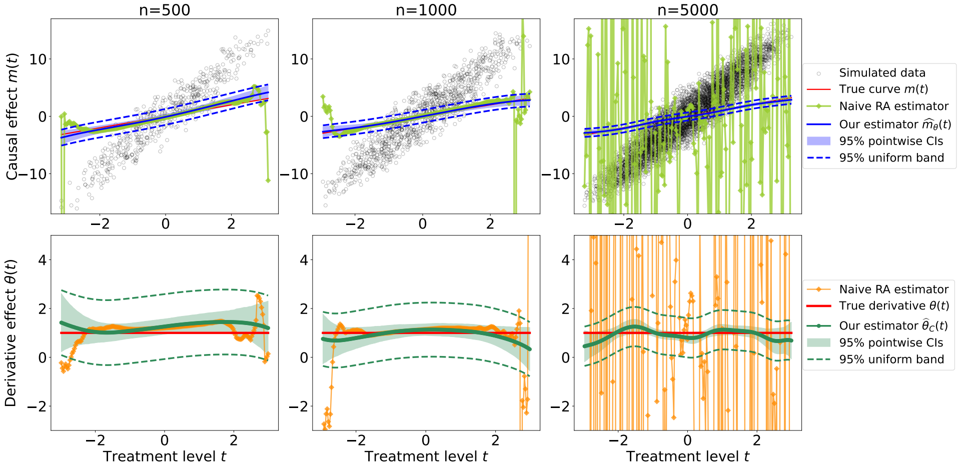
- $E \sim \text{Uniform}[-0.3, 0.3]$ is an independent treatment variation,
- $\epsilon \sim \mathcal{N}(0, 1)$ is an exogenous normal noise.



Generate i.i.d. observations $\{(Y_i, T_i, \mathbf{S}_i)\}_{i=1}^{2000}$ from

$$Y = T + 6S_1 + 6S_2 + \epsilon, \quad T = 2S_1 + S_2 + E, \quad \text{and} \quad (S_1, S_2) \sim \text{Uniform}[-1,1]^2,$$

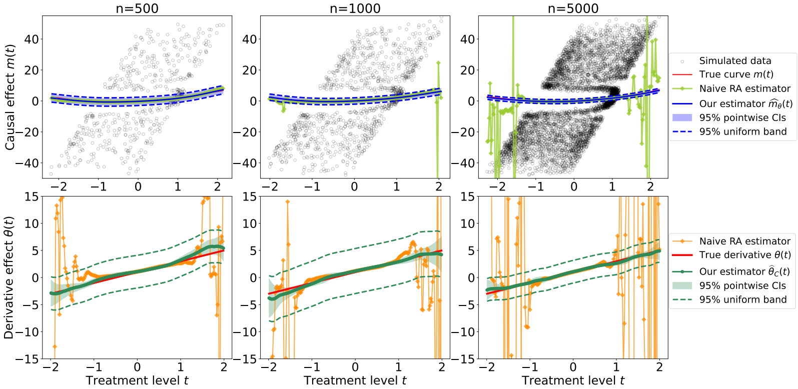
- $E \sim \text{Uniform}[-0.5, 0.5]$ and $\epsilon \sim \mathcal{N}(0,1)$.



Generate i.i.d. observations $\{(Y_i, T_i, \mathbf{S}_i)\}_{i=1}^{2000}$ from

$$Y = T^2 + T + 10Z + \epsilon, \quad T = \cos\left(\pi Z^3\right) + \frac{Z}{4} + E, \quad \text{and} \quad Z = 4S_1 + S_2,$$

- $(S_1, S_2) \sim \text{Uniform}[-1,1]^2$, $E \sim \text{Uniform}[-0.1, 0.1]$, and $\epsilon \sim \mathcal{N}(0,1)$.
- Those doubly robust methods based on pseudo-outcomes (Kennedy et al., 2017; Takatsu and Westling, 2022) do not work in this example.



For simplicity, we assume the additive confounding model

$$Y = \bar{m}(T) + \eta(\mathbf{S}) + \epsilon, \quad T = f(\mathbf{S}) + E \quad \text{with} \quad \mathbb{E}\left[\eta(\mathbf{S})\right] = 0 \quad \text{and} \quad \mathbb{E}(E) = 0.$$

When $\mathrm{Var}(E) = 0$,

- $\mu(t, \mathbf{s})$ can be identified only on a lower-dimensional surface $\{(t, \mathbf{s}) \in \mathcal{T} \times \mathcal{S} : t = f(\mathbf{s})\}$ so that

$$\mu\left(f(\mathbf{s}), \mathbf{s}\right) = \bar{m}\left(f(\mathbf{s})\right) + \eta(\mathbf{s}) = m\left(f(\mathbf{s})\right) + \eta(\mathbf{s}). \tag{2}$$

- The relation $T = f(\mathbf{S})$ can be recovered from the data $\{(T_i, \mathbf{S}_i)\}_{i=1}^n$.

Assumption (Bounded random effect)

Let $L_f(t) = \{\mathbf{s} \in \mathcal{S} : f(\mathbf{s}) = t\}$ be a level set of the function $f : \mathcal{S} \to \mathbb{R}$ at $t \in \mathcal{T}$. There exists a constant $\rho_1 > 0$ such that

$$\rho_1 \geq \max\left\{\sup_{t\in\mathcal{T}} \sup_{\mathbf{s}\in L_f(t)} |\eta(\mathbf{s})|, \; \frac{\sup_{t\in\mathcal{T}} \sup_{\mathbf{s}\in L_f(t)} \mu(f(\mathbf{s}), \mathbf{s}) - \inf_{t\in\mathcal{T}} \inf_{\mathbf{s}\in L_f(t)} \mu(f(\mathbf{s}), \mathbf{s})}{2}\right\}.$$

By (2) and the first lower bound on $\rho_1 \geq \sup_{t\in\mathcal{T}} \sup_{\boldsymbol{s}\in L_f(t)} |\eta(\boldsymbol{s})|$ in the previous assumption, we know that

$$|\mu(f(\boldsymbol{s}), \boldsymbol{s}) - m(t)| = |\eta(\boldsymbol{s})| \leq \rho_1$$

for any $\boldsymbol{s} \in L_f(t)$. It also implies that

$$\begin{aligned} m(t) &\in \bigcap_{\boldsymbol{s}\in L_f(t)} \left[\mu(f(\boldsymbol{s}), \boldsymbol{s}) - \rho_1,\ \mu(f(\boldsymbol{s}), \boldsymbol{s}) + \rho_1\right] \\ &= \left[\sup_{\boldsymbol{s}\in L_f(t)} \mu(f(\boldsymbol{s}), \boldsymbol{s}) - \rho_1,\ \inf_{\boldsymbol{s}\in L_f(t)} \mu(f(\boldsymbol{s}), \boldsymbol{s}) + \rho_1\right], \end{aligned}$$

which is the nonparametric bound on $m(t)$ that contains all the possible values of $m(t)$ for any fixed $t \in \mathcal{T}$ when $\mathrm{Var}(E) = 0$.

- This bound is well-defined and nonempty under the second lower bound on ρ_1 in the previous assumption.