Nonparametric Inference on Dose-Response Curves Without the Positivity Condition

Yen-Chi Chen

Department of Statistics

University of Washington

- Joint work with Yikun Zhang and Alex Giessing
- Supported by NSF DMS-1952781, 2112907, 2141808, and NIH U24-AG07212

Introduction

Prelude: causal inference

- In a typical causal problem, our data consists of IID random vectors
$$(Y_1, T_1, S_1), \cdots, (Y_n, T_n, S_n).$$
- $Y \in \mathbb{R}$: outcome of interest.
- $T \in \mathbb{R}$: treatment.
- $S \in \mathbb{R}^d$: covariates.

Prelude: causal inference

- In a typical causal problem, our data consists of IID random vectors

$$(Y_1, T_1, S_1), \cdots, (Y_n, T_n, S_n).$$

- $Y \in \mathbb{R}$: outcome of interest.
- $T \in \mathbb{R}$: treatment.
- $S \in \mathbb{R}^d$: covariates.
- We want to investigate the causal effect of T on the outcome of interest Y.

Prelude: binary treatment - 1

- The simplest causal problem is the binary treatment problem.
- In this case, the treatment $T \in \{0, 1\}$ is a binary variable.
- $T = 1$ indicates that the individual is treated ($T = 0$ means not treated/received placebo).

Prelude: binary treatment - 1

- The simplest causal problem is the binary treatment problem.
- In this case, the treatment $T \in \{0, 1\}$ is a binary variable.
- $T = 1$ indicates that the individual is treated ($T = 0$ means not treated/received placebo).
- A simple way to investigate the causal effect is the potential outcome model: we denote $Y(0), Y(1)$ to be the potential outcomes.
- $Y(0)$ is the outcome if the individual is NOT treated; $Y(1)$ is the outcome if the individual IS treated.

Prelude: binary treatment - 2

- A common causal effect of interest is the average treatment effect (ATE):

$$\mathbb{E}(Y(1)) - \mathbb{E}(Y(0)).$$

Prelude: binary treatment - 2

- A common causal effect of interest is the average treatment effect (ATE):

$$\mathbb{E}(Y(1)) - \mathbb{E}(Y(0)).$$

- In potential outcome framework, the observed outcome $Y = TY(1) + (1 - T)Y(0)$, or equivalently,

$$Y = Y(t)$$

  conditioned on $T = t$. This is also known as the **consistency**.

Prelude: binary treatment - 2

- A common causal effect of interest is the average treatment effect (ATE):
$$\mathbb{E}(Y(1)) - \mathbb{E}(Y(0)).$$
- In potential outcome framework, the observed outcome $Y = TY(1) + (1 - T)Y(0)$, or equivalently,
$$Y = Y(t)$$
conditioned on $T = t$. This is also known as the **consistency**.
- The challenge is: we only observe one of the two potential outcomes!

Prelude: binary treatment - 3

- To resolve this problem, we often use the ignorability assumption.
- **Ignorability assumption:** $(Y(1), Y(0)) \perp T | S$.
- Under this assumption, the covariates S are called the *confounders*.

Prelude: binary treatment - 3

- To resolve this problem, we often use the ignorability assumption.
- **Ignorability assumption:** $(Y(1), Y(0)) \perp T|S$.
- Under this assumption, the covariates S are called the *confounders*.
- With the ignorability, we can rewrite

$$\mathbb{E}(Y(1)) = \mathbb{E}\left(\frac{YT}{P(T=1|S)}\right), \qquad \mathbb{E}(Y(0)) = \mathbb{E}\left(\frac{Y(1-T)}{P(T=0|S)}\right)$$

and estimate the ATE accordingly.

- To resolve this problem, we often use the ignorability assumption.
- **Ignorability assumption:** $(Y(1), Y(0)) \perp T|S$.
- Under this assumption, the covariates S are called the *confounders*.
- With the ignorability, we can rewrite

$$\mathbb{E}(Y(1)) = \mathbb{E}\left(\frac{YT}{P(T=1|S)}\right), \qquad \mathbb{E}(Y(0)) = \mathbb{E}\left(\frac{Y(1-T)}{P(T=0|S)}\right)$$

  and estimate the ATE accordingly.
- Here, we need an additional **positivity assumption**: $P(T = t|s) > 0$ for $s \in \mathbb{S} \equiv \mathsf{supp}(S)$ and $t = 0, 1$.

Continuous treatment: PM2.5 Example

	fips	name	lng	lat	PM2.5	CMR
1	1059	Franklin	-87.84328	34.44238	8.045251	452.8492
3	19109	Kossuth	-94.20690	43.20414	6.857354	294.3387
4	40115	Ottawa	-94.81059	36.83588	8.073921	424.5076
5	42115	Susquehanna	-75.80090	41.82128	7.955338	383.5730
8	29213	Taney	-93.04128	36.65474	7.026484	348.6023
9	32510	Carson City	-119.74735	39.15108	4.063737	347.6080

Figure: An example of PM2.5 data on cardiovascular mortality rate (CMR) at county-level.

- We want to investigate the effect of PM2.5 on the CMR[1].
- The treatment variable T is the amount of PM2.5 at a county, which is *not binary but a continuous number*!

[1]Data from US National Center for Health Statistics and Community Multiscale Air Quality modeling system

Continuous treatment: PM2.5 Example

	fips	name	lng	lat	PM2.5	CMR
1	1059	Franklin	-87.84328	34.44238	8.045251	452.8492
3	19109	Kossuth	-94.20690	43.20414	6.857354	294.3387
4	40115	Ottawa	-94.81059	36.83588	8.073921	424.5076
5	42115	Susquehanna	-75.80090	41.82128	7.955338	383.5730
8	29213	Taney	-93.04128	36.65474	7.026484	348.6023
9	32510	Carson City	-119.74735	39.15108	4.063737	347.6080

Figure: An example of PM2.5 data on cardiovascular mortality rate (CMR) at county-level.

- We then encounter the problem of *continuous* treatment.
- We will work with the same assumptions:
  1. **Consistency:** $Y = Y(t)$ conditioned on $T = t$.
  2. **Ignorability:** $\{Y(t) : t \in \mathsf{supp}(T)\} \perp T | S$.

Continuous treatment: potential outcome models

- The effect of continuous treatment is characterized by the *dose-response curve*

$$m(t) = \mathbb{E}(Y(t)).$$

Continuous treatment: potential outcome models

- The effect of continuous treatment is characterized by the *dose-response curve*

$$m(t) = \mathbb{E}(Y(t)).$$

- A popular parametric model is the Marginal Structural Models (MSMs; [RHB2000]):

$$\mathbb{E}(Y(t)) = f(t;\theta),$$

  where $f(t;\theta)$ belongs to a given family parameterized by θ such as $f(t;\theta) = \theta_0 + \theta_1 t$.

- The MSMs is a fully parametric model, which may not capture the structure of $m(t)$.

Continuous treatment: do-calculus

- An alternative way of framing a causal problem is the graphical model approach and so-called *do-calculus*.
- In this case, the dose-response curve can be written as

$$m(t) = \mathbb{E}(Y|\text{do}(T = t)) = \mathbb{E}[\mathbb{E}(Y|T = t, S)].$$

Continuous treatment: do-calculus

- An alternative way of framing a causal problem is the graphical model approach and so-called *do-calculus*.
- In this case, the dose-response curve can be written as

$$m(t) = \mathbb{E}(Y|\text{do}(T = t)) = \mathbb{E}[\mathbb{E}(Y|T = t, S)].$$

- The above implies a simple estimation procedure. We first estimate $\mu(t, s) = \mathbb{E}(Y|T = t, S = s)$ with $\widehat{\mu}(t, s)$. Then we estimate $m(t)$ via a naive estimator

$$\widetilde{m}(t) = \frac{1}{n} \sum_{i=1}^{n} \widehat{\mu}(t, S_i).$$

- Alternatively, we can use an inverse probability weighting (IPW; [CL2020, HHLL2020]) estimator for this problem:

$$\widetilde{m}_{IPW}(t) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{T_i - t}{h}\right) \frac{Y_i}{\widehat{p}(t|S_i)},$$

  where $\widehat{p}(t|s)$ is an estimator of the conditional PDF $p(t|s)$ and $K(\cdot)$ is a smoothing kernel such as a Gaussian.
- There is also a doubly-robust version of this idea via pseudo-outcome [KMMS2017].

Continuous treatment: the positivity condition

- Both the naive and IPW estimators require the positivity condition, i.e.,

$$\text{(PS)} \qquad p(t|s) > 0 \qquad \forall s \in \mathbb{S},$$

  where $\mathbb{S}$ is the support of S.

Continuous treatment: the positivity condition

- Both the naive and IPW estimators require the positivity condition, i.e.,

$$(\text{PS}) \qquad p(t|s) > 0 \qquad \forall s \in \mathbb{S},$$

  where $\mathbb{S}$ is the support of S.

- To see why (PS) is needed, recall the naive estimator is

$$\widetilde{m}(t) = \frac{1}{n} \sum_{i=1}^{n} \widehat{\mu}(t, S_i).$$

- Without (PS), we cannot have a consistent estimator of $\widehat{\mu}(t, s)$ evaluating on $s = S_i$!

Identification

Additive confounding model

- In this work, we will focus on additive confounding model.
- Recall that we have a triplet of observations (Y, T, S), where $Y \in \mathbb{R}$ is the outcome, $T \in \mathbb{R}$ is the treatment, and $S \in \mathbb{R}^d$ is the confounder.

Additive confounding model

- In this work, we will focus on additive confounding model.
- Recall that we have a triplet of observations (Y, T, S), where $Y \in \mathbb{R}$ is the outcome, $T \in \mathbb{R}$ is the treatment, and $S \in \mathbb{R}^d$ is the confounder.
- We assume that

$$\begin{aligned} Y &= m(T) + \eta(S) + \epsilon, \\ T &= f(S) + E, \end{aligned} \tag{1}$$

where (ϵ, E) are independent mean 0 noises and $\mathbb{E}(\eta(S)) = 0$.

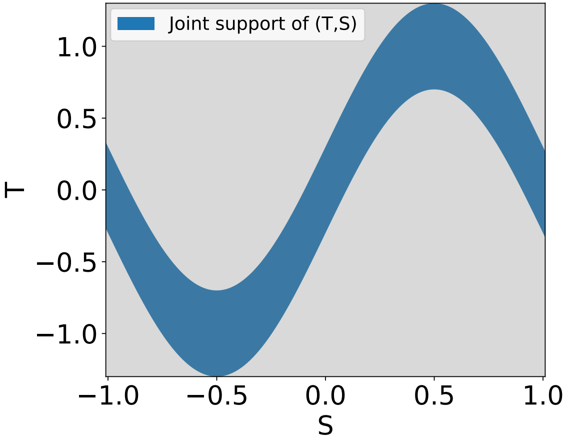
Additive confounding model

- In this work, we will focus on additive confounding model.
- Recall that we have a triplet of observations (Y, T, S), where $Y \in \mathbb{R}$ is the outcome, $T \in \mathbb{R}$ is the treatment, and $S \in \mathbb{R}^d$ is the confounder.
- We assume that

$$\begin{aligned} Y &= m(T) + \eta(S) + \epsilon, \\ T &= f(S) + E, \end{aligned} \tag{1}$$

  where (ϵ, E) are independent mean 0 noises and $\mathbb{E}(\eta(S)) = 0$.
- In spatial confounding problem (such as PM2.5 studies), the above model is often assumed and is known as *spatial additive confounding model* [KW2003, WD2024].

The support of (T,S)



A very common scenario is that the noise E is bounded, leading to a violation of the positivity condition.

Properties of the additive model

Theorem (ZCG2024)

Assume the additive confounding model and $\mathbb{E}(\eta(S)) = 0$. *Then*

1. $\mathbb{E}(Y|T = t) = m(t) + \mathbb{E}(\eta(S)|T = t) \neq m(t)$.
2. *Let* $\theta(t) = \frac{\partial}{\partial t} m(t)$. *Then*

$$\theta(t) = \theta_C(t)$$
$$\theta_C(t) = \mathbb{E}\left(\frac{\partial}{\partial t}\mu(t, S)|T = t\right)$$

The first result shows that naively using conditional mean suffers from a spatial confounding bias. The second result is a key to our identification.

Properties of the derivative

- Without positivity, $p(t|s)$ can be 0 so we do not have a consistent estimator of $\mu(t, s)$.
- Our integral estimator is based on the following fact:

$$\theta(t) = m'(t) = \theta_C(t) = \mathbb{E}\left(\frac{\partial}{\partial t}\mu(t, S)|T = t\right).$$

Properties of the derivative

- Without positivity, $p(t|s)$ can be 0 so we do not have a consistent estimator of $\mu(t,s)$.
- Our integral estimator is based on the following fact:

$$\theta(t) = m'(t) = \theta_C(t) = \mathbb{E}\left(\frac{\partial}{\partial t}\mu(t,S)|T=t\right).$$

- The quantity $\theta_C(t)$ **can be estimated consistently** because it is conditioned on $T=t$.

Properties of the derivative

- Without positivity, $p(t|s)$ can be 0 so we do not have a consistent estimator of $\mu(t,s)$.
- Our integral estimator is based on the following fact:

$$\theta(t) = m'(t) = \theta_C(t) = \mathbb{E}\left(\frac{\partial}{\partial t}\mu(t,S)|T=t\right).$$

- The quantity $\theta_C(t)$ **can be estimated consistently** because it is conditioned on $T=t$.
- We then use the relation

$$m(t) - m(\tau) = \int_{s=\tau}^{s=t} m'(s)ds = \int_{s=\tau}^{s=t} \theta_C(s)ds$$

to estimate $m(t)$.

The integral estimator

The integral estimator - 1

- Recall that we have

$$m(t) - m(\tau) = \int_{s=\tau}^{s=t} m'(s)ds = \int_{s=\tau}^{s=t} \theta_C(s)ds$$

for any τ.

The integral estimator - 1

- Recall that we have

$$m(t) - m(\tau) = \int_{s=\tau}^{s=t} m'(s)ds = \int_{s=\tau}^{s=t} \theta_C(s)ds$$

  for any τ.
- Thus, $m(t) = m(T) + \int_{s=T}^{s=t} \theta_C(s)ds$, which implies

$$\begin{aligned} m(t) &= \mathbb{E}\left(m(T) + \int_{s=T}^{s=t} \theta_C(s)ds\right) \\ &= \mathbb{E}\left(m(T) + \eta(S) + \epsilon + \int_{s=T}^{s=t} \theta_C(s)ds\right) \\ &= \mathbb{E}\left(Y + \int_{s=T}^{s=t} \theta_C(s)ds\right). \end{aligned}$$

The integral estimator - 2

- Let $\widehat{\theta}_C(t)$ be an estimator of $\theta_C(t)$.
- The **integral estimator** is

$$\widehat{m}(t) = \frac{1}{n}\sum_{i=1}^{n} Y_i + \int_{s=T_i}^{s=t} \widehat{\theta}_C(s)ds.$$

- Thus, the key is to construct a good estimator of $\theta_C(t) = \mathbb{E}\left(\frac{\partial}{\partial t}\mu(t,S)|T=t\right)$.

The derivative estimator - 1

- We recommend to use the local polynomial regression to estimate $\frac{\partial}{\partial t}\mu(t,s)$.

The derivative estimator - 1

- We recommend to use the local polynomial regression to estimate $\frac{\partial}{\partial t}\mu(t,s)$.
- Let $\widehat{\beta}(t,s) \in \mathbb{R}^3, \widehat{\alpha}(t,s) \in \mathbb{R}^d$ be the minimizer of

$$\sum_{i=1}^{n}\left[Y_i - \sum_{j=1}^{3}\beta_j(T_i - t)^{j-1} - \sum_{\ell=1}^{d}\alpha_\ell(S_{i,\ell} - s_\ell)\right]^2 K_T\left(\frac{T_i - t}{h}\right)K_S\left(\frac{\|S_i - s\|}{b}\right),$$

where K_T and K_S are smoothing kernel and $h, b > 0$ are smoothing bandwidth.

The derivative estimator - 1

- We recommend to use the local polynomial regression to estimate $\frac{\partial}{\partial t}\mu(t,s)$.
- Let $\widehat{\beta}(t,s) \in \mathbb{R}^3, \widehat{\alpha}(t,s) \in \mathbb{R}^d$ be the minimizer of

$$\sum_{i=1}^{n}\left[Y_i - \sum_{j=1}^{3}\beta_j(T_i - t)^{j-1} - \sum_{\ell=1}^{d}\alpha_\ell(S_{i,\ell} - s_\ell)\right]^2 K_T\left(\frac{T_i - t}{h}\right)K_S\left(\frac{\|S_i - s\|}{b}\right),$$

where K_T and K_S are smoothing kernel and $h, b > 0$ are smoothing bandwidth.

- It is known that the second component $\widehat{\beta}_2(t,s)$ is a consistent estimator of $\frac{\partial}{\partial t}\mu(t,s)$; see, e.g., [F2018].

The derivative estimator - 2

- Note that

$$\theta_C(t) = \mathbb{E}\left(\frac{\partial}{\partial t}\mu(t,S)|T=t\right) = \int \frac{\partial}{\partial t}\mu(t,s)dP(s|t).$$

The derivative estimator - 2

- Note that

$$\theta_C(t) = \mathbb{E}\left(\frac{\partial}{\partial t}\mu(t, S)|T = t\right) = \int \frac{\partial}{\partial t}\mu(t, s)dP(s|t).$$

- Thus, we also need an estimator of $P(s|t)$. Here we simply use a kernel CDF estimator

$$\widehat{P}(s|t) = \frac{\sum_{i=1}^{n} I(S_i \leq s)\bar{K}_T\left(\frac{T_i - t}{\hbar}\right)}{\sum_{j=1}^{n} \bar{K}_T\left(\frac{T_j - t}{\hbar}\right)}.$$

- Note: other estimators are applicable–kernel CDF is just a simple and reliable estimator.

The derivative estimator - 3

- Combining the above two estimators, our estimator $\theta_C(t)$ can be written as

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

The derivative estimator - 3

- Combining the above two estimators, our estimator $\theta_C(t)$ can be written as

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

- Thus, the integral estimator is

$$\widehat{m}(t) = \frac{1}{n}\sum_{i=1}^{n} Y_i + \int_{s=T_i}^{s=t} \widehat{\theta}_C(s) ds.$$

- Note: the above integral estimator is also a *linear smoother*.

The numerical method - 1

- The integral estimator

$$\widehat{m}(t) = \frac{1}{n}\sum_{i=1}^{n} Y_i + \int_{s=T_i}^{s=t} \widehat{\theta}_C(s)ds$$

  require the evaluation of integration $\int_{s=T_i}^{s=t}$, which could be computationally expansive.
- Here we propose a simple numerical method for approximating this.

The numerical method - 1

- The integral estimator
$$\widehat{m}(t) = \frac{1}{n}\sum_{i=1}^{n} Y_i + \int_{s=T_i}^{s=t} \widehat{\theta}_C(s)ds$$
require the evaluation of integration $\int_{s=T_i}^{s=t}$, which could be computationally expansive.
- Here we propose a simple numerical method for approximating this.
- Let $T_{(1)} \leq T_{(2)} \leq \cdots \leq T_{(n)}$ be the ordered values of the observed treatment.
- We then have
$$\frac{1}{n}\sum_{i=1}^{n}\int_{s=T_i}^{s=t} \widehat{\theta}_C(s)ds = \frac{1}{n}\sum_{i=1}^{n}\int_{s=T_{(i)}}^{s=t} \widehat{\theta}_C(s)ds.$$

The numerical method - 2

- The above result implies

$$\widehat{m}(T_{(j)}) = \bar{Y}_n + \frac{1}{n}\sum_{i=1}^{n}\int_{s=T_{(i)}}^{s=T_{(j)}} \widehat{\theta}_C(s)ds.$$

- Let $\Delta_j = T_{(j+1)} - T_{(j)}$.

The numerical method - 2

- The above result implies

$$\widehat{m}(T_{(j)}) = \bar{Y}_n + \frac{1}{n}\sum_{i=1}^{n}\int_{s=T_{(i)}}^{s=T_{(j)}} \widehat{\theta}_C(s)ds.$$

- Let $\Delta_j = T_{(j+1)} - T_{(j)}$.
- When $i < j$, we use Riemann sum,

$$\int_{s=T_{(i)}}^{s=T_{(j)}} \widehat{\theta}_C(s)ds \approx \sum_{\ell=i}^{\ell=j-1} \widehat{\theta}_C(T_{(\ell)})\Delta_\ell.$$

The numerical method - 2

- The above result implies

$$\widehat{m}(T_{(j)}) = \bar{Y}_n + \frac{1}{n}\sum_{i=1}^{n}\int_{s=T_{(i)}}^{s=T_{(j)}} \widehat{\theta}_C(s)ds.$$

- Let $\Delta_j = T_{(j+1)} - T_{(j)}$.
- When $i < j$, we use Riemann sum,

$$\int_{s=T_{(i)}}^{s=T_{(j)}} \widehat{\theta}_C(s)ds \approx \sum_{\ell=i}^{\ell=j-1} \widehat{\theta}_C(T_{(\ell)})\Delta_\ell.$$

- When $i > j$, we use Riemann sum,

$$\int_{s=T_{(i)}}^{s=T_{(j)}} \widehat{\theta}_C(s)ds \approx -\sum_{\ell=j}^{\ell=i-1} \widehat{\theta}_C(T_{(\ell+1)})\Delta_\ell.$$

The numerical method - 3

- When we include $\sum_{i=1}^{n}$, some $\widehat{\theta}_C(T_{(\ell)})$ will be used multiple times, which eventually leads to the following result:

$$\frac{1}{n}\sum_{i=1}^{n}\int_{s=T_{(i)}}^{s=T_{(j)}}\widehat{\theta}_C(s)ds$$

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

The numerical method - 3

- When we include $\sum_{i=1}^{n}$, some $\widehat{\theta}_C(T_{(\ell)})$ will be used multiple times, which eventually leads to the following result:

$$\frac{1}{n}\sum_{i=1}^{n}\int_{s=T_{(i)}}^{s=T_{(j)}}\widehat{\theta}_C(s)ds$$
$$\approx \frac{1}{n}\sum_{i=1}^{n-1}\Delta_i\left[i\cdot\widehat{\theta}_C(T_{(i)})I(i<j)-(n-i)\cdot\widehat{\theta}_C(T_{(i+1)})I(i\geq j)\right].$$

- The above result only requires evaluating $\widehat{\theta}_C(t)$ at the observed $T_1,\cdots,T_n$ once!

The numerical method - 3

- When we include $\sum_{i=1}^{n}$, some $\widehat{\theta}_C(T_{(\ell)})$ will be used multiple times, which eventually leads to the following result:

$$\frac{1}{n}\sum_{i=1}^{n}\int_{s=T_{(i)}}^{s=T_{(j)}}\widehat{\theta}_C(s)ds$$
$$\approx \frac{1}{n}\sum_{i=1}^{n-1}\Delta_i\left[i\cdot\widehat{\theta}_C(T_{(i)})I(i<j)-(n-i)\cdot\widehat{\theta}_C(T_{(i+1)})I(i\geq j)\right].$$

- The above result only requires evaluating $\widehat{\theta}_C(t)$ at the observed $T_1,\cdots,T_n$ once!
- As a result, we can quickly approximate

$$\widehat{m}(T_{(j)})\approx \bar{Y}_n+\frac{1}{n}\sum_{i=1}^{n-1}\Delta_i\left[i\cdot\widehat{\theta}_C(T_{(i)})I(i<j)-(n-i)\cdot\widehat{\theta}_C(T_{(i+1)})I(i\geq j)\right]$$

The numerical method - 4

- Finally, to approximate $\widehat{m}(t)$, we first find the interval $[T_{(j^*)}, T_{(j^*+1)}]$ such that

$$t \in [T_{(j^*)}, T_{(j^*+1)}].$$

- We then use a linear interpolation between $\widehat{m}(T_{(j^*)})$ and $\widehat{m}(T_{(j^*+1)})$ to approximate $\widehat{m}(t)$.

Confidence bands via the bootstrap

- We may construct a simultaneous confidence band of $m(t)$ via the bootstrap.
- Let $(Y_1^*, T_1^*, S_1^*), \cdots, (Y_n^*, T_n^*, S_n^*)$ be a bootstrap sample (sampling with replacement of the original data).
- We compute the bootstrap estimator $\widehat{m}^*(t)$.
- Let $\widehat{\xi}^*_{1-\alpha}$ be the $1-\alpha$ quantile of

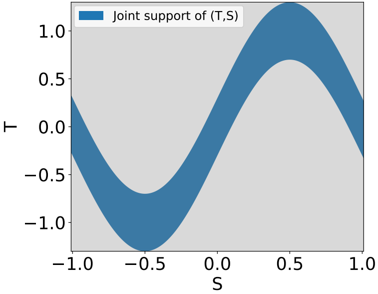
$$\sup_t |\widehat{m}^*(t) - \widehat{m}(t)|.$$

- A $1-\alpha$ simultaneous confidence band is

$$[\widehat{m}(t) - \widehat{\xi}^*_{1-\alpha}, \quad \widehat{m}(t) + \widehat{\xi}^*_{1-\alpha}]$$

Asymptotic theory

The support of (T,S) revisited



- Let $\mathscr{E}$ be the support of (T, S).
- In the above figure, the support is the blue area, which shows a clear violation of (PS).

Bypassing the positivity condition

- $\theta_C(t) = \int \frac{\partial}{\partial t}\mu(t,s)dP(s|t)$ only require $\widehat{\beta}_2(t,s)$ to be consistent on $\mathcal{E}$!
- Feature of the local polynomial estimator: $\widehat{\beta}_2(t,s)$ is consistent estimator in $\mathcal{E}$.

Uniform convergence of derivative estimator

Lemma (ZCG2024)

Under regularity conditions (A3-A5, A6-1, A6-2),

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

This shows that the local polynomial estimator is uniformly consistent in $\mathscr{E}$. Note that the convergence rate differs a little on the boundary of $\mathscr{E}$ versus its interior.

Uniform convergence of integral estimator - 1

Combining with the convergence of kernel CDF, we immediately have the following result:

Theorem (ZCG2024)

Let $\mathcal{T}' \subset \mathcal{T} \equiv \mathsf{supp}(T)$ be a compact set. Under regularity conditions (A1-A6),

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

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

Uniform convergence of integral estimator - 2

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

- Blue term: the bias in local polynomial estimator.
- Red term: additional bias from boundary of $\mathcal{E}$.
- Orange term: rate from $\bar{Y}_n$.
- Brown term: stochastic variation of local polynomial estimator.
- Cyan term: rate from kernel CDF.

- Clearly, $\hbar^* \asymp n^{-1/5}$ is the optimal rate, which is similar to the conventional problem.
- If we choose $h \asymp b$, then the optimal rate is

$$h^* \asymp b^* \asymp n^{-1/(d+7)},$$

which is slightly slower than the conventional rate $n^{-1/(d+5)}$.
- The slightly slowness of the rate is due to estimating the derivative.

- To show the bootstrap validity, we first need to derive an asymptotic linear form of $\widehat{m}(t)$.
- For simplicity, we assume that $h \asymp b$, so the convergence rate becomes

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

- We let $\hbar \asymp \left(\frac{\log n}{n}\right)^{-1/5}$ be the optimal choice so the kernel CDF converges faster. Thus, we only need to focus on the primary term

$$O\left(h^2\right) + O_P\left(\sqrt{\frac{|\log(h^{d+1})}{nh^{d+3}}}\right).$$

Bootstrap Validity - 2

- We consider an undersmoothing h so that $nh^{d+7} \to 0$. Under this choice, the bias converges faster than the variance, and the rate is

$$\sup_{t \in \mathcal{T}'} |\widehat{m}(t) - m(t)| = O_P\left(\sqrt{\frac{|\log(h^{d+1})}{nh^{d+3}}}\right).$$

Bootstrap Validity - 3

Lemma (Asymptotic linearity)

Under regularity conditions (A1-A6), $h \asymp b$, $\hbar \asymp \left(\frac{\log n}{n}\right)^{-1/5}$, *and* $nh^{d+7} \to 0$. *There exists a function* $\psi_t : \mathbb{Y} \times \mathbb{T} \times \mathbb{S} \to \mathbb{R}$ *such that*

$$\left|\sqrt{nh^{d+3}}(\widehat{m}(t) - m(t)) - \mathbb{G}_n\psi_t\right| = O_P\left(\sqrt{nh^{d+7}} + \sqrt{\frac{\log n}{n\hbar^2}} + \sqrt{\frac{h^{d+3}\log n}{\hbar}} + \sqrt{\frac{h^{d+3}}{\hbar^2}}\right),$$

where $\mathbb{G}_n f = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}[f(Y_i, T_i, S_i) - \mathbb{E}(f(Y, T, S))$.

Note that $\mathbb{Y}, \mathbb{T}, \mathbb{S}$ are the support of Y, T, S, respectively.

Bootstrap Validity - 4

- With the above asymptotic linearity, we are able to approximate the distribution of $\sup_t |\widehat{m}(t) - m(t)|$ by a maximum of a Gaussian process, leading to the validity of the bootstrap.
- Namely, we have

$$\sqrt{nh^{d+3}} \sup_{t \in \mathcal{T}'} |\widehat{m}(t) - m(t)| \approx \sup_{t \in \mathcal{T}'} |\mathbb{G}_n \psi_t| \approx \sup_{t \in \mathcal{T}'} |\mathbb{B}_n \psi_t|,$$

  where $\mathbb{B}_n f_t$ is a Gaussian process on the function class f_t indexed by t.
- The bootstrap maximum approximates the above maximum, leading to the consistency of the bootstrap confidence band [CCK2014, G2023].

Corollary (Bootstrap validity)

Under regularity conditions (A1-A6), $h \asymp b$, $\hbar \asymp \left(\frac{\log n}{n}\right)^{-1/5}$, and $nh^{d+7} \to 0$. Let $\xi^*_{1-\alpha}$ be the bootstrap quantile. Then

$$P\left(m(t) \in [\widehat{m}(t) - \widehat{\xi}^*_{1-\alpha}, \widehat{m}(t) + \widehat{\xi}^*_{1-\alpha}] \quad \forall t \in \mathcal{T}'\right) = 1 - \alpha + O_P\left(\left(\frac{\log^5 n}{nh^{d+3}}\right)^{1/8}\right)$$

Case study: PM2.5 effect

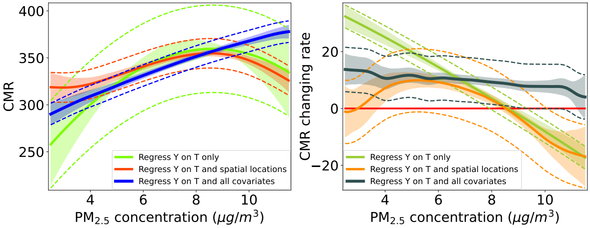
PM2.5 data

	fips	name	lng	lat	PM2.5	CMR
1	1059	Franklin	-87.84328	34.44238	8.045251	452.8492
3	19109	Kossuth	-94.20690	43.20414	6.857354	294.3387
4	40115	Ottawa	-94.81059	36.83588	8.073921	424.5076
5	42115	Susquehanna	-75.80090	41.82128	7.955338	383.5730
8	29213	Taney	-93.04128	36.65474	7.026484	348.6023
9	32510	Carson City	-119.74735	39.15108	4.063737	347.6080

Figure: An example of PM2.5 data on cardiovascular mortality rate (CMR) at county-level.

- The above data table shows the average PM2.5 and CMR over 1990-2010 of each county.
- We also have other 8 county-level informations such as population, unemployment rates, household income, ...etc.
- We want to investigate how PM2.5 would impact the CMR.

PM2.5 data



- We consider three model: naive method, adjusting for spatial confounding, adjusting for all covariates.
- The confidence bands are pointwise.
- A clear increasing effect after adjusting for all covariates.

Discussion

Summary

- Our integral estimator allows us to bypass the positivity condition.
- We have a fast algorithm, nice asymptotic theory, and methods for making inferences.
- This idea opens a new direction for investigating continuous treatments because the violation of positivity is very common!

Open problems and future work

- **Inverse probability weighting.** Our method is essentially a regression adjustment (g-computation) method. Can we generalize it to the inverse probability weighting approach?
- **Doubly-robustness.** Following the previous result, are we able to construct a doubly-robust estimator? We may need to use a cross-fitting (double machine learning) approach in this case.
- **High-dimensional confounders.** In addition to 2D spatial confounders, we may have high-dimensional confounders with a sparse linear effect. Will our method work?
- **Unmeasured confounders.** We assume all confounders are observed. Can we handle unmeasured confounders? Perhaps with some known instruments?

Thank You!

All codes and data are available:

`https://github.com/zhangyk8/npDoseResponse/tree/main`

Paper reference: `https://arxiv.org/abs/2405.09003`.

References

1. **[CCK2014]** Chernozhukov, V., Chetverikov, D., & Kato, K. (2014). Gaussian approximation of suprema of empirical processes.
2. **[CL2020]** Colangelo, K., & Lee, Y. Y. (2020). Double debiased machine learning nonparametric inference with continuous treatments. arXiv preprint arXiv:2004.03036.
3. **[F2018]** Fan, J. (2018). Local polynomial modelling and its applications: monographs on statistics and applied probability 66. Routledge.
4. **[G2023]** Giessing, A. (2023). Gaussian and Bootstrap Approximations for Suprema of Empirical Processes. arXiv preprint arXiv:2309.01307.
5. **[HHLL2020]** Huber, M., Hsu, Y. C., Lee, Y. Y., & Lettry, L. (2020). Direct and indirect effects of continuous treatments based on generalized propensity score weighting. Journal of Applied Econometrics, 35(7), 814-840.
6. **[KW2003]** Kammann, E. E., & Wand, M. P. (2003). Geoadditive models. Journal of the Royal Statistical Society Series C: Applied Statistics, 52(1), 1-18.
7. **[KMMS2017]** Kennedy, E. H., Ma, Z., McHugh, M. D., & Small, D. S. (2017). Non-parametric methods for doubly robust estimation of continuous treatment effects. Journal of the Royal Statistical Society Series B: Statistical Methodology, 79(4), 1229-1245.
8. **[RHB2000]** Robins, J. M., Hernan, M. A., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. Epidemiology, 11(5), 550-560.
9. **[WR2024]** Wiecha, N., & Reich, B. J. (2024). Two-stage Spatial Regression Models for Spatial Confounding. arXiv preprint arXiv:2404.09358.
10. **[ZCG2024]** Zhang, Y., Chen, Y. C., & Giessing, A. (2024). Nonparametric Inference on Dose-Response Curves Without the Positivity Condition. arXiv preprint arXiv:2405.09003.

Assumptions: causal assumptions

- **A1-1: Consistency.** Given $T = t$, $Y = Y(t)$.
- **A1-2: Ignorability.** $\{Y(t) : t \in \mathbb{T}\} \perp T|S$.
- **A1-3: Treatment variation.** The variance $\mathsf{Var}(E) > 0$ in the equation $T = f(S) + E$.

- **A2: Derivative identification.** $\theta(t) = \theta_C(t) = \mathbb{E}\left[\frac{\partial}{\partial t}\mu(t, S)|T = t\right]$ and $\mathbb{E}(\mu(T, S)) = \mathbb{E}(m(T))$.

Assumptions: nuisance parameters

- **A3: Conditional mean.** $\mu(t, s)$ is at least 3-times continuously differentiable with respect to t and at least 4-times continuously differentiable with respect to s.

- **A4: Joint density.** $p(t, s)$ is at least twice continuously differentiable with bounded partial derivatives up to 2nd order in the interior of $\mathscr{E}$. All partial derivative are continuous up to, $\partial\mathscr{E}$, the boundary of $\mathscr{E}$. $\mathscr{E}$ is compact and $\sup_{(t,s)\in\mathscr{E}} p(t, s) > 0$.

Assumptions: boundary condition

- **A5-1: Smooth boundary.** There are constants $r_1, r_2 \in (0,1)$ such that for any $(t,s) \in \mathcal{E}$ and all $\delta \in (0, r_1]$, there is another point $(t', s') \in \mathcal{E}$ such that

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

- **A5-2: Boundary derivative.** For any $(t,s) \in \mathcal{E}$, $\frac{\partial}{\partial t} p(t,s) = \frac{\partial}{\partial s_j} p(t,s) = 0$ and $\frac{\partial^2}{\partial s_j^2} \mu(t,s) = 0$ for all $j = 1, \cdots, d$.
- **A5-3: Stable volume.** The Lebesgue measure of the set $\partial\mathcal{E} \oplus \delta$ satisfies

$$\mathsf{Leb}(\partial\mathcal{E} \oplus \delta) \leq A_1 \cdot \delta$$

for some constant A_1, where $\mathbb{A} \oplus \delta = \{z : \inf_{x \in \mathbb{A}} \|x - z\| \leq \delta\}$.

Assumptions: kernels in local polynomials

- **A6-1: Regular.** K_T, K_S are compactly supported and Lipchitz kernel with K_T being symmetric and K_S is radially symmetric and are second-order kernels.
- **A6-2: VC-type kernels.** Let

$$\mathscr{K}_{3,d} = \Big\{(y,z) \mapsto \left(\frac{y-t}{h}\right)^{\ell} \left(\frac{z_i - s_i}{b}\right)^{k_1} \left(\frac{z_j - s_j}{b}\right)^{k_2} \times K_T\left(\frac{y-t}{h}\right) K_S\left(\frac{z-s}{b}\right) : (t,s) \in \mathscr{E}; \; i,j = 1,\cdots,d; \ell = 0,\cdots,6; k_1,k_2 = 0,1; h,b > 0\Big\}$$

The class $\mathscr{K}_{3,d}$ is VC-type class.

Assumptions: kernel in the kernel CDF

- **A6-3: Regular of kernel CDF.** $\bar{K}_T$ is a compactly supported, Lipchitz, symmetric, and second-order kernel.
- **A6-4: VC-type kernel CDF.** Let

$$\bar{\mathcal{K}} = \left\{ y \mapsto \bar{K}_T \left(\frac{y - t}{\hbar} \right) : t \in \mathbb{T}, \hbar > 0 \right\}$$

The class $\bar{\mathcal{K}}$ is VC-type class.

Asymptotic linearity - 1

- In the asymptotic linearity, we have

$$\sqrt{nh^{d+3}}(\widehat{m}(t) - m(t)) \approx \mathbb{G}_n \psi_t.$$

- ψ_t is the following function

$$\psi_t(Y, T, S) = \mathbb{E}_{T_2}\left[\int_{\widetilde{t}=T_2}^{t} \widetilde{\psi}_{\widetilde{t}}(Y, T, S) d\widetilde{t}\right]$$

with

$$\widetilde{\psi}_{\widetilde{t}}(Y, T, S) = \mathbb{E}_{T_3, S_3}\left[\frac{e_2^T M_3^{-1} \Psi_{\widetilde{t}, S_3}(Y, T, S)}{\sqrt{hb^d} p(\widetilde{t}, S_3) p_T(\widetilde{t})} \cdot \frac{1}{\hbar} \bar{K}_T\left(\frac{\widetilde{t} - T_3}{\hbar}\right)\right],$$

where $e_2 = (0, 1, 0, \cdots, 0) \in \mathbb{R}^{3+d}$ and $M_2 \in \mathbb{R}^{(3+d)\times(3+d)}$ is a block diagonal matrix of constants.

Asymptotic linearity - 2

- $\Psi_{t,s}(y,z,v) \in \mathbb{R}^{3+d}$ is the following function

$$\Psi_{t,s}(y,z,v) = y \cdot \begin{bmatrix} \left(\frac{z-t}{h}\right)^{j-1} K_T\left(\frac{z-t}{h}\right) K_S\left(\frac{v-s}{b}\right)_{1\le j\le 3} \\ \left(\frac{v_{j-3}-s_{j-3}}{b}\right) K_T\left(\frac{z-t}{h}\right) K_S\left(\frac{v-s}{b}\right)_{4\le j\le 3+d} \end{bmatrix}$$