

Uncovering Hidden Treatment Effect Heterogeneity Through Optimal Transport

Kyle Schindl
Clemson University
School of Mathematical and Statistical Sciences Seminar

September 23rd, 2026

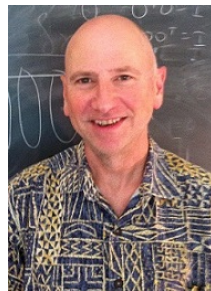


Coauthors



Ashley Naimi

Professor
Department of Epidemiology
Emory University



Larry Wasserman

UPMC Professor of Statistics & Data Science
Department of Statistics & Data Science
Machine Learning Department
Carnegie Mellon University

Warm-Up: Causal Inference

- Suppose we observe $Z_1, \dots, Z_n \stackrel{iid}{\sim} P$ where $Z_i = (X_i, A_i, Y_i)$.
- $X_i \in \mathbb{R}^d$ are covariates, $A_i \in \{0, 1\}$ is treatment, and $Y_i \in \mathbb{R}$ are outcomes.
- This is what we actually **observed**.
- The causal perspective is when we explore what *would have happened*.
- Let $Y_i(a)$ denote the *potential outcome*, such that

$$Y_i(A_i) = \begin{cases} Y_i(0) & \text{if } A_i = 0 \\ Y_i(1) & \text{if } A_i = 1 \end{cases}$$

- Typically we are interested in estimating the *Average Treatment Effect*

$$\tau = \mathbb{E}[Y(1) - Y(0)].$$

Identifying Assumptions

- The counterfactual τ is identified under the following conditions:
 - *Consistency*: $Y = Y(a)$ if $A = a$.
 - *Exchangeability*: $A \perp\!\!\!\perp Y(a) \mid X$ for $a \in \{0, 1\}$.
 - *Positivity*: $0 < \varepsilon \leq \pi(x) = \mathbb{P}(A = 1 \mid X = x) \leq 1 - \varepsilon < 1$.
- Then, it follows that $\tau = \mathbb{E}[\mathbb{E}[Y \mid A = 1, X] - \mathbb{E}[Y \mid A = 0, X]]$.
- **Idea**: These assumptions identify more than just the mean!
- Let $P_{Y(a)|x}$ denote the counterfactual distribution of $Y(a) \mid X = x$.
- Then, the cumulative distribution function of $Y(a) \mid X = x$ is identified as

$$F_a(y|x) = P(Y \leq y \mid X = x, A = a).$$

- We use $Q_a(u|x) = \inf\{y : F_a(y|x) \geq u\}$ for $u \in (0, 1)$ to denote the quantile function.

Conditional Treatment Effects

- **Problem:** Considering τ in isolation might be misleading.
- Large positive effects for one subgroup may cancel out negative effects for others.
- **Solution:** consider instead the Conditional Average Treatment Effect:

$$\tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x].$$

- $\tau(x)$ reveals which groups benefit more or less, on average.
- This can be very useful to inform treatment targeting, more granular analyses, etc.
- In many ways, the CATE has become synonymous with the term “treatment effect heterogeneity.” Over the years, a very large literature has developed.

At a Glance: CATE Papers in the Literature

30 selected journal papers on CATE estimation, inference, and interpretation

Crump et al. (2008)	<i>Rev. Econ. Stat.</i>	Nie and Wager (2021)	<i>Biometrika</i>
Su et al. (2009)	<i>JMLR</i>	Semenova and Chernozhukov (2021) [†]	<i>Econometrics J.</i>
Hill (2011) [†]	<i>JCGS</i>	Fan et al. (2022)	<i>JBES</i>
Imai and Ratkovic (2013)	<i>Ann. Appl. Stat.</i>	Liang and Yu (2022) [†]	<i>JASA</i>
Tian et al. (2014) [†]	<i>JASA</i>	Yadlowsky et al. (2022)	<i>Ann. Statist.</i>
Abrevaya et al. (2015)	<i>JBES</i>	Zhao et al. (2022)	<i>JRSS-B</i>
Athey and Imbens (2016)	<i>PNAS</i>	Cui et al. (2023)	<i>JRSS-B</i>
Chen et al. (2017) [†]	<i>Biometrics</i>	Foster and Syrgkanis (2023) [†]	<i>Ann. Statist.</i>
Huang and Chan (2017)	<i>Biometrika</i>	Kennedy (2023)	<i>Electron. J. Stat.</i>
Lee et al. (2017)	<i>J. Appl. Econom.</i>	Yang et al. (2023)	<i>JRSS-B</i>
Lu et al. (2018)	<i>JCGS</i>	Kennedy et al. (2024)	<i>Ann. Statist.</i>
Wager and Athey (2018)	<i>JASA</i>	Chernozhukov et al. (2025)	<i>Econometrica</i>
Athey et al. (2019) [†]	<i>Ann. Statist.</i>	Hines et al. (2025)	<i>Biometrics</i>
Künzel et al. (2019)	<i>PNAS</i>	Imai and Li (2025)	<i>JBES</i>
Hahn et al. (2020)	<i>Bayesian Anal.</i>	Yadlowsky et al. (2025)	<i>JASA</i>

[†] Broader frameworks or effect scales with CATE-related applications. Selected examples, not an exhaustive review.

Beyond Average Effects

- **Problem:** The CATE is useful, but it only describes effects at the mean level.
- E.g. treatment could leave the mean unchanged, but double the standard deviation:

$$Y(0) \mid X = x \sim N(0, 1),$$

$$Y(1) \mid X = x \sim N(0, 2^2)$$

- Here, $\tau(x) = 0$, but there is clearly a treatment effect!
- One solution is to instead consider quantile treatment effects.
- **Problem:** QTEs are often difficult to analyze and interpret.
- **Solution:** We consider causal effects defined by distributional distances.

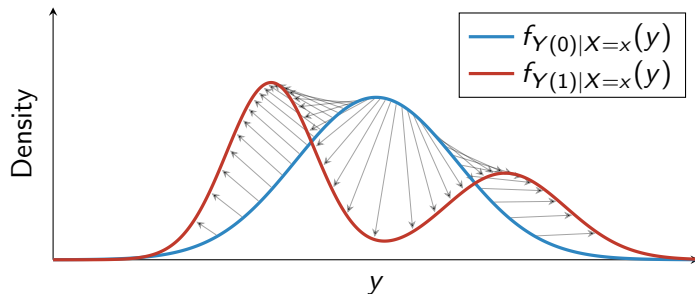
Warm-Up: The Wasserstein Distance

- The Wasserstein distance between any two probability distributions P and Q is

$$W_2^2(P, Q) = \inf_{\gamma \in \Gamma(P, Q)} \int \|x - y\|_2^2 d\gamma(x, y)$$

where $\Gamma(P, Q)$ is the set of all couplings of P and Q .

- This describes the minimal transportation cost of transforming or “morphing” P into Q .



Conditional Wasserstein Treatment Effects

- When $Y \in \mathbb{R}$, the Wasserstein distance can be represented as

$$\psi^2(x) = W_2^2(P_{Y(1)|x}, P_{Y(0)|x}) = \int_0^1 \{Q_{Y(1)|x}(u) - Q_{Y(0)|x}(u)\}^2 du$$

- **Big Idea:** $\psi^2(x)$ and $\Delta(x, u) = Q_{Y(1)|x}(u) - Q_{Y(0)|x}(u)$ are *the* central tools required to understand heterogeneous treatment effects.
- They are more closely related to the CATE than you may expect.
- In fact,

$$\int \Delta(x, u) du = \mathbb{E}[Y(1) - Y(0) \mid X = x] = \tau(x)$$

$$\int \Delta^2(x, u) du = \psi^2(x).$$

- $\Delta(x, u)$ can be used to unify mean and distributional notions of heterogeneity.

Two is Better Than One

- Why are Wasserstein Treatment Effects useful? It can be combined with other effects!

Theorem 1 (Conditional Effect Inequality)

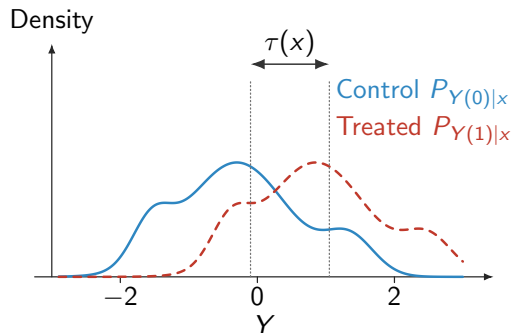
Suppose $P_{Y(a)|x} \in \mathcal{P}_2(\mathbb{R})$ for $a \in \{0, 1\}$. Then, for P_X -almost every x ,

$$|\tau(x)| \leq \psi(x)$$

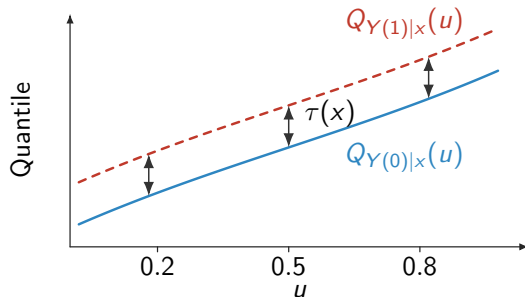
where equality holds if and only if $\Delta(x, u)$ is constant in u .

- That is, if $Q_{Y(1)|x}(u) = Q_{Y(0)|x}(u) + \tau(x)$.
- This provides a measure of *heterogeneity across quantiles*.
- If $|\tau(x)|$ and $\psi(x)$ are very close together, this suggests a homogeneous effect.
- If they are far apart, there must be higher-order effects not picked up by the mean.

Alignment of $\tau(x)$ and $\psi^2(x)$



(a) Same-shape location shift



(b) Constant gap across quantiles

Figure: A homogeneous treatment effect at covariate level x : $Q_{Y(1)|x}(u) = Q_{Y(0)|x}(u) + \tau(x)$.

Heterogeneity Across Quantiles

- More generally, for $U \sim \text{Uniform}(0, 1)$, we show that

$$\psi^2(x) = \tau^2(x) + \mathbb{V}(\Delta(X, U) \mid X = x)$$

- $\psi^2(x)$ measures the location-based shift, and the heterogeneity around that shift.
- This can be used to define the index

$$\gamma(x) = \frac{\mathbb{V}(\Delta(X, U) \mid X = x)}{\psi^2(x)} = 1 - \left(\frac{|\tau(x)|}{\psi(x)} \right)^2 \in [0, 1]$$

- When $\gamma(x) = 0$ then the covariate-specific treatment effect is homogeneous.
- When $\gamma(x) = 1$ then the CATE explains none of the distributional effect.

Individual Treatment Effect Heterogeneity

- Curiously, $\psi^2(x)$ and $\tau(x)$ also describe *individual* treatment effect heterogeneity.

Theorem 2 (Individual Treatment Effect Bounds)

Let $\Gamma_x = \Gamma(P_{Y(1)|x}, P_{Y(0)|x})$ and suppose that $P_{Y(a)|x} \in \mathcal{P}_2(\mathbb{R})$ for $a \in \{0, 1\}$. Then,

$$\inf_{\pi \in \Gamma_x} \left\{ \mathbb{V}_{\pi}(Y(1) - Y(0) \mid X = x) \right\} = \psi^2(x) - \tau^2(x) = \mathbb{V}(\Delta(X, U) \mid X = x),$$

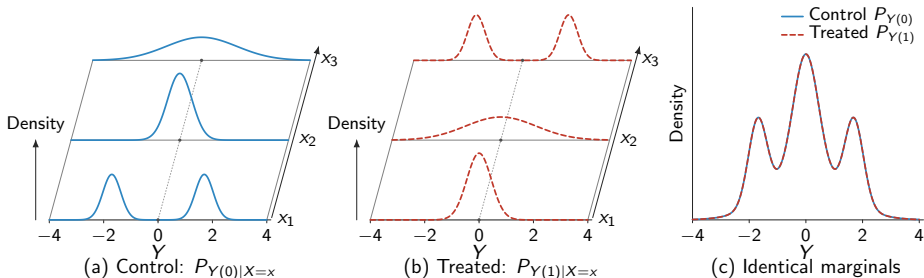
$$\sup_{\pi \in \Gamma_x} \left\{ \mathbb{V}_{\pi}(Y(1) - Y(0) \mid X = x) \right\} = \psi_{\max}^2(x) - \tau^2(x)$$

where $\psi_{\max}^2(x) = \int_0^1 \{Q_{Y(1)|x}(u) - Q_{Y(0)|x}(1 - u)\}^2 du$.

Uncovering Hidden Treatment Effects

- Earlier I suggested that $\tau(x)$ may miss out on treatment effects.
- We aggregate all possible conditional treatment effects with the integrated Wasserstein effect,

$$\Psi^2 = \mathbb{E}[\psi^2(X)] = \int_{\mathcal{X}} W_2^2(P_{Y(1)|x}, P_{Y(0)|x}) dP_X(x).$$



The Aggregation Gap

- In fact, we can actually quantify how much was missed without conditioning.
- It is easily shown that $W_2^2(P_{Y(1)}, P_{Y(0)}) \leq \mathbb{E}[W_2^2(P_{Y(1)|X}, P_{Y(0)|X})] = \Psi^2$.

Definition 3 (Aggregate-Gap Index)

Suppose that $\Psi > 0$. Then, define

$$\mathcal{S} = 1 - \frac{W_2^2(P_{Y(1)}, P_{Y(0)})}{\Psi^2} \in [0, 1]$$

- When $\mathcal{S} = 0$ then the marginal Wasserstein effect is identical to Ψ^2 .
- Therefore, no distributional effects are hidden across sub-populations.
- When $\mathcal{S} = 1$ then the marginal outcome distributions are identical despite there being material distributional effects occurring in particular strata.

Decomposing Treatment Effect Heterogeneity

- **Idea:** We have now shown there are *three* sources of heterogeneity to care about.
- That is, across ranks and across covariates, and their interaction.
- We isolate each by applying a functional ANOVA decomposition to $\Delta(x, u)$.
- Let

$$a(x) = \int \Delta(x, u) du - \tau$$

$$b(u) = \int \Delta(x, u) dP_X(x) - \tau$$

$$c(x, u) = \Delta(x, u) - \tau - a(x) - b(u)$$

- $a(x)$ is the covariate main effect, $b(u)$ is the rank main effect, and $c(x, u)$ is the covariate-rank interaction.

A Taxonomy of Heterogeneity

Theorem 4 (Two-Way Distributional Heterogeneity Decomposition)

Suppose $\Psi < \infty$. Then,

$$\Psi^2 = \tau^2 + \mathbb{E}[a^2(X)] + \mathbb{E}[b^2(U)] + \mathbb{E}[c^2(X, U)]$$

where the functions $(x, u) \mapsto \tau$, $(x, u) \mapsto a(x)$, $(x, u) \mapsto b(u)$, and $(x, u) \mapsto c(x, u)$ are mutually orthogonal in $L^2(P_X \otimes du)$.

- Now we have a way to isolate the exact contributions from each source of heterogeneity.
- Note that $\mathbb{E}[a^2(X)] = \mathbb{V}(\tau(X))$ is simply covariate heterogeneity.
- Let $\bar{\Delta}(u) = \int_{\mathcal{X}} \Delta(x, u) dP_X(x)$ to be the effect profile, then

$$\mathbb{E}[b^2(U)] = \mathbb{V}(\bar{\Delta}(U)).$$

A Testing Framework

Proposition 1 (Heterogeneity Characterizations)

The following equivalences hold:

- (i) $\mathbb{E}[a^2(X)] = 0$ if and only if $\tau(x)$ is constant in x .*
- (ii) $\mathbb{E}[b^2(U)] = 0$ if and only if $\bar{\Delta}(u)$ is constant in u .*
- (iii) $\mathbb{E}[c^2(X, U)] = 0$ if and only if the effect surface is additively separable, i.e.*

$$\Delta(x, u) = \tau(x) + \bar{\Delta}(u) - \tau$$

- Now, we have a way to test for which types of heterogeneity are present.
- In fact, we can easily enumerate all 2^3 possible settings we could observe.

An Abundance of Tests

(i)	(ii)	(iii)	Interpretation
0	0	0	Homogeneous additive effect: $\Delta(x, u) = \tau$.
$\neq 0$	0	0	CATE-only heterogeneity: $\Delta(x, u) = \tau(x)$.
0	$\neq 0$	0	Rank-only heterogeneity: $\Delta(x, u) = \bar{\Delta}(u)$.
0	0	$\neq 0$	Pure covariate-rank interaction: $\Delta(x, u) = \tau + c(x, u)$.
$\neq 0$	$\neq 0$	0	Additively separable covariate and rank heterogeneity: $\Delta(x, u) = \tau(x) + \bar{\Delta}(u) - \tau$.
$\neq 0$	0	$\neq 0$	CATE heterogeneity with subgroup-specific rank effects: $\Delta(x, u) = \tau(x) + c(x, u)$.
0	$\neq 0$	$\neq 0$	Homogeneous CATE with subgroup-specific rank heterogeneity: $\Delta(x, u) = \tau + b(u) + c(x, u)$.
$\neq 0$	$\neq 0$	$\neq 0$	Full heterogeneity: $\Delta(x, u) = \tau + a(x) + b(u) + c(x, u)$.

Table: Taxonomy and interpretation of heterogeneity tests.

A Model Based Perspective

- This taxonomy defines a collection of lower-complexity models for the full surface $\Delta(x, u)$.
- Let $\mathcal{H} = L^2(P_X \otimes du)$ and consider the model classes:

$$\mathcal{M}_{\text{ATE}} = \{m \in \mathcal{H} : m(x, u) = \gamma \text{ for some } \gamma \in \mathbb{R}\}$$

$$\mathcal{M}_{\text{CATE}} = \{m \in \mathcal{H} : m(x, u) = g(x) \text{ for some } g \in L^2(P_X)\}$$

$$\mathcal{M}_{\text{Rank}} = \{m \in \mathcal{H} : m(x, u) = h(u) \text{ for some } h \in L^2(du)\}$$

$$\mathcal{M}_{\text{Add}} = \{m \in \mathcal{H} : m(x, u) = \gamma + g(x) + h(u), \mathbb{E}[g(X)] = 0, \mathbb{E}[h(U)] = 0\}.$$

- Under $\mathcal{M}_{\text{CATE}}$ the CATE is sufficient for describing the full distributional effect $\Delta(x, u)$.
- Each of these components are orthogonal projections of $\Delta(x, u)$ onto a simpler model.

Estimation of Conditional Treatment Effects

- In general, estimation of $\tau(x)$ and $\psi^2(x)$ is difficult.
- When X is continuous, both are function-valued targets, so they are not pathwise differentiable and the usual $\sqrt{n}$ -semiparametric efficiency theory is not available.
- This has partially motivated the large literature surrounding the CATE.
- Thus, for $\psi^2(x)$ we use plug-in estimation,

$$\widehat{\psi}^2(x) = \sum_{m=1}^M \omega_m \{ \widehat{Q}_1(u_m|x) - \widehat{Q}_0(u_m|x) \}^2$$

although we are developing alternative approaches in follow-up work.

- However, *integrated* effects are pathwise differentiable, so we can develop efficiency theory for our testing framework.

Efficiency Theory

- The efficient influence function for Ψ^2 directly builds off of (Firpo, 2007). Let

$$\eta_a(Z; u) = \frac{\mathbb{I}(A = a)}{\pi_a(X)} \left(\frac{u - \mathbb{I}(Y \leq Q_a(u|X))}{f_a(Q_a(u|X)|X)} \right)$$

denote the quantile score.

Theorem 5 (Efficient Influence Function for the Integrated Wasserstein Effect)

Suppose that $0 < f_a(Q_a(u|x)|x) < \infty$ for almost every $u \in (0, 1)$. Then, the efficient influence function of Ψ^2 under a nonparametric model is given by

$$\varphi_{\Psi^2}(Z) = \int_0^1 \Delta^2(X, u) du + 2 \int_0^1 \Delta(X, u) \{ \eta_1(Z; u) - \eta_0(Z; u) \} du - \Psi^2.$$

Difficulties in Estimation

- The drawback of distributional estimands is estimating distributional quantities.
- We leverage the fact that the conditional density at the conditional quantile can be written as the sparsity function:

$$\frac{1}{f_a(Q_a(u|x)|x)} = \frac{\partial}{\partial u} Q_a(u|x) =: s_a(x, u)$$

- We smooth $u \mapsto \hat{Q}_a(u|x)$ and then differentiate to obtain $\hat{s}_a(x, u)$.
- While this can be difficult to estimate in practice, we establish Neyman-orthogonality for the one step estimator.

Limiting Distribution

Theorem 6 (Asymptotic Normality)

- (i) *Positivity and boundedness:* $\pi(x)$, $\hat{\pi}_a(x)$, and $s_a(x, u)$ are bounded away from zero, $\hat{s}_a(x, u)$ and $|\hat{\Delta}(x, u)|$ are uniformly bounded from above.
- (ii) *Local Lipschitzness:* There exists some constant $L < \infty$ and $\delta > 0$ such that for P_X -almost every x and all relevant u , $\sup_{x,u} |\hat{\Delta}(x, u)| \leq \delta$ and for all $|t| < \delta$,

$$|f_a(Q_a(u|x) + t|x) - f_a(Q_a(u|x)|x)| \leq L|t|.$$

- (iii) *Nuisance function consistency:* $\|\varphi(\cdot; \hat{P}) - \varphi(\cdot; P)\|_2^2 = o_P(1)$.
- (iv) *Nuisance errors:* $\sum_{a=0}^1 \left\{ \|\hat{Q}_a - Q_a\|_2 \left(\|\hat{\pi}_a - \pi_a\|_2 + \|\hat{s}_a - s_a\|_2 + \|\hat{Q}_a - Q_a\|_2 \right) \right\} = o_P(n^{-1/2})$.

Then, as $n \rightarrow \infty$ it follows that $\frac{\sqrt{n}}{\sigma_{\Psi^2}} (\hat{\Psi}_n^2 - \Psi^2) \xrightarrow{d} \mathcal{N}(0, 1)$ where $\sigma_{\Psi^2}^2 = \mathbb{V}(\varphi_{\Psi^2}(Z))$.

Extension to Multivariate Outcomes

- Often we observe $Y(a) \in \mathbb{R}^p$, but in $\mathbb{R}^p$ there is no longer a canonical quantile function.
- Fix an absolutely continuous reference distribution $\nu \in \mathcal{P}_2(\mathbb{R}^p)$ with support $\mathcal{U} \subseteq \mathbb{R}^p$.
- E.g. the uniform distribution on $[0, 1]^p$ or the unit ball, a standard Gaussian.
- **Idea:** We map $u \in \mathcal{U}$ to the untreated group, then to the treated group.
- Roughly,

$$\underbrace{u}_{\text{reference}} \xrightarrow{S_{0,x}} \underbrace{y_0 = S_{0,x}(u)}_{\text{untreated outcome}} \xrightarrow{T_x} \underbrace{y_1 = T_x(y_0)}_{\text{treated endpoint}}$$

- We let $S_{0,x} : \mathcal{U} \rightarrow \mathbb{R}^p$ to be the unique Brenier map that satisfies $(S_{0,x})_{\#}\nu = P_{Y(0)|x}$.
- And $T_x = \nabla \phi_x : \mathbb{R}^p \rightarrow \mathbb{R}^p$ such that $(T_x)_{\#}P_{Y(0)|x} = P_{Y(1)|x}$ is the optimal transport map, where ϕ_x is some convex function.
- Now, we may define the multivariate displacement surface to be

$$\Lambda(x, u) = T_x(S_{0,x}(u)) - S_{0,x}(u)$$

Multivariate Integrated Effects

- Consider $\Lambda(x, u) = T_x(S_{0,x}(u)) - S_{0,x}(u) \iff \Delta(x, u) = Q_{Y(1)|x}(u) - Q_{Y(0)|x}(u)$.
- $\Lambda(x, u)$ is the direct analogue of $\Delta(x, u)$.
- If $p = 1$, $\mathcal{U} = (0, 1)$, ν is the Lebesgue measure, then

$$\begin{aligned} S_{0,x}(u) &= Q_0(u|x), \\ T_x(S_{0,x}(u)) &= Q_1(u|x) \end{aligned}$$

and so $\Lambda(x, u) = Q_1(u|x) - Q_0(u|x) = \Delta(x, u)$.

- We can now define the integrated Wasserstein effect as

$$\Psi^2 = \mathbb{E} \left[\int_{\mathcal{U}} \|\Lambda(X, u)\|_2^2 d\nu(u) \right].$$

Multivariate Functional ANOVA Decomposition

- Again, we may define

$$a(x) = \tau(x) - \tau \quad b(u) = \int_{\mathcal{X}} \Lambda(x, u) dP_X(x) - \tau$$

and $c(x, u) = \Lambda(x, u) - a(x) - b(u) - \tau$.

Theorem 7 (Multivariate Functional ANOVA Decomposition)

Suppose $\Psi^2 < \infty$. Then,

$$\Psi^2 = \|\tau\|_2^2 + \mathbb{E} [\|a(X)\|_2^2] + \mathbb{E} [\|b(U)\|_2^2] + \mathbb{E} [\|c(X, U)\|_2^2]$$

where the functions $(x, u) \mapsto \tau$, $(x, u) \mapsto a(x)$, $(x, u) \mapsto b(u)$, and $(x, u) \mapsto c(x, u)$ are mutually orthogonal in $L^2(P_X \otimes \nu; \mathbb{R}^p)$.

Dependence on the Reference Distribution

- Recall that our construction of a multivariate quantile required a reference measure ν .
- However, only $\mathbb{E} [\|b(U)\|_2^2]$ and $\mathbb{E} [\|c(X, U)\|_2^2]$ depend on ν .
- Ψ^2 , $\|\tau\|_2^2$, and $\mathbb{E} [\|a(X)\|_2^2]$ do not depend on ν .
- Thus, we can *still* test for heterogeneity across quantiles without choosing a reference:

$$\mathbb{E} [\|b(U)\|_2^2] + \mathbb{E} [\|c(X, U)\|_2^2] = \left(\mathbb{E} [\psi^2(X) - \|\tau(X)\|_2^2] \right).$$

- We only run into trouble if we want to isolate the rank main effects or interactions.

Matrix-valued Decompositions

- When $Y \in \mathbb{R}^p$, we can consider other geometric perspectives. Let

$$M_\Lambda = \mathbb{E} \left[\int_{\mathcal{U}} \Lambda(X, u) [\Lambda(X, u)]^\top d\nu(u) \right]$$

- The trace of this matrix is actually the integrated Wasserstein effect!

$$\text{tr}(M_\Lambda) = \mathbb{E} \left[\int_{\mathcal{U}} \text{tr}\{\Lambda(X, u) \Lambda(X, u)^\top\} d\nu(u) \right] = \mathbb{E} \left[\int_{\mathcal{U}} \|\Lambda(X, u)\|_2^2 d\nu(u) \right] = \Psi^2.$$

- Let $d_x(y) = T_x(y) - y$. Then,

$$M_\Lambda = \tau\tau^\top + \mathbb{E} \left[\{\tau(X) - \tau\} \{\tau(X) - \tau\}^\top \right] + \mathbb{E} \left[\int \{d_x(y) - \tau(X)\} \{d_x(y) - \tau(X)\}^\top dP_{Y(0)|X}(y) \right]$$

- This opens the door to *spectral analyses* of heterogeneous effects.

Sinkhorn Treatment Effects

- In practice, computation of the multivariate Wasserstein distance is very challenging.
- Instead, we consider entropy-regularized optimal transport. For some $\varepsilon > 0$, let

$$OT_{\varepsilon}(P, Q) = \inf_{\gamma \in \Gamma(P, Q)} \left\{ \int_{\mathbb{R}^p \times \mathbb{R}^p} \|x - y\|_2^2 d\gamma(x, y) + \varepsilon \text{KL}(\gamma \parallel P \otimes Q) \right\}.$$

- Note that $OT_{\varepsilon}(P, Q) > 0$ even when $P = Q$.
- This motivates the *Sinkhorn divergence*, which removes this self-interaction bias,

$$S_{\varepsilon}(P, Q) = OT_{\varepsilon}(P, Q) - \frac{1}{2}OT_{\varepsilon}(P, P) - \frac{1}{2}OT_{\varepsilon}(Q, Q).$$

- Future work will consider $\Psi_{\varepsilon}^2 = \mathbb{E}[S_{\varepsilon}(P_{Y|X, A=1}, P_{Y|X, A=0})]$, including decompositions, inference, and dependence on ε .

Challenges for Statistical Inference

- Interestingly, constructing confidence intervals for L_2 norms is less straightforward than you may expect.
- Consider a toy example where $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ and we are interested in estimating $\psi = \mu^2$. Using the estimator $\hat{\psi} = \bar{X}_n^2$, it follows that

$$\begin{aligned}\sqrt{n}(\hat{\psi} - \psi) &\rightsquigarrow N(0, \eta^2) \text{ for some } \eta^2 \text{ when } \mu \neq 0, \text{ and} \\ n\hat{\psi} &\rightsquigarrow \sigma^2 \chi_1^2 \quad \text{when } \mu = 0.\end{aligned}$$

- When μ is close to zero, its distribution will be neither normal or χ_1^2 , and its rate of convergence will be between $1/n$ and $1/\sqrt{n}$.

Statistical Inference

- For our squared-norm targets, the regularity conditions give

$$\hat{\theta} - \theta = \frac{1}{n} \sum_{i=1}^n \varphi(Z_i) + R_n, \quad R_n = o_P(n^{-1/2}).$$

- Away from the null: $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$, at the null $\varphi = 0$ almost surely.
- Following Verdinelli and Wasserman (2024), we retain an $n^{-1/2}$ -scale interval using

$$\widehat{\text{se}} = \left\{ \frac{\hat{\sigma}^2 + \varsigma^2}{n} \right\}^{1/2}$$

where $\varsigma > 0$ is some constant, e.g. $\varsigma = \mathbb{V}(Y)$.

- Then, a conservative stabilized Wald interval is

$$\mathcal{C}_{\mathcal{T}, 1-\alpha} = \left[(\hat{\theta} - z_{1-\alpha/2} \widehat{\text{se}})_+, (\hat{\theta} + z_{1-\alpha/2} \widehat{\text{se}})_+ \right],$$

where $(t)_+ = \max\{t, 0\}$. At the null, coverage tends to one.

Simulation: one CATE, three stories

- For $i = 1, \dots, n$, independently generate

$$X_i \sim \text{Uniform}(-1, 1), \quad A_i \sim \text{Bernoulli}(1/2), \quad \varepsilon_i \sim N(0, 1),$$

and set

$$Y_i = X_i^2 + A_i\tau(X_i) + \{1 + A_i s_j(X_i)\}\varepsilon_i, \quad \tau(x) = 0.5 + 0.25x.$$

- Three settings.** Fix $\kappa = 0.75$ and vary only $s_j(x)$.

Setting	$s_j(x)$	Heterogeneity beyond the CATE
Location shift	0	None
Common dispersion	κ	Common rank profile
Varying dispersion	κx	Covariate-rank interaction

- This yields the same CATE and untreated distribution in every setting:

$$\Delta_j(x, u) = Q_{1,j}(u | x) - Q_{0,j}(u | x) = \tau(x) + s_j(x)\Phi^{-1}(u).$$

Simulation results: one CATE, different rank effects

One-step estimates: Monte Carlo mean (empirical SD), 500 replications.

Setting	n	$\mathbb{E}[a^2(X)]$	$\mathbb{E}[b^2(U)]$	$\mathbb{E}[c^2(X, U)]$	L_{CATE}
Location shift	<i>Truth</i>	1/48	0	0	0
	500	−0.026 (0.046)	−0.030 (0.036)	−0.083 (0.056)	−0.113 (0.060)
	2,000	0.010 (0.016)	−0.007 (0.008)	−0.031 (0.017)	−0.038 (0.019)
Common dispersion	<i>Truth</i>	1/48	0.5625	0	0.5625
	500	−0.074 (0.090)	0.539 (0.167)	−0.172 (0.114)	0.367 (0.176)
	2,000	−0.002 (0.027)	0.557 (0.068)	−0.063 (0.036)	0.494 (0.072)
Varying dispersion	<i>Truth</i>	1/48	0	0.1875	0.1875
	500	−0.030 (0.058)	−0.035 (0.055)	0.102 (0.098)	0.067 (0.096)
	2,000	0.008 (0.018)	−0.008 (0.010)	0.158 (0.038)	0.150 (0.039)

$L_{\text{CATE}} = \mathbb{E}[b^2(U)] + \mathbb{E}[c^2(X, U)]$: squared error from replacing $\Delta(x, u)$ by $\tau(x)$.

Discussion and Conclusion

- In this paper we showed there are three important sources of heterogeneity:
- Across covariates, quantiles, and the interaction between them.
- By decomposing the integrated Wasserstein effect

$$\Psi^2 = \int_{\mathcal{X}} W_2^2(P_{Y(1)|x}, P_{Y(0)|x}) dP_X(x)$$

we can isolate each component.

- Future work will focus on decompositions using other distributional distances.
- It would also be interesting to consider spectral analyses of heterogeneous effects.
- We are also exploring doubly robust estimation of $\psi^2(x)$.

Thank you!

- Questions?

References

- Abrevaya, J., Hsu, Y.-C., and Lieli, R. P. (2015). Estimating conditional average treatment effects. *Journal of Business & Economic Statistics*, 33(4):485–505.
- Athey, S. and Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360.
- Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178.
- Chen, S., Tian, L., Cai, T., and Yu, M. (2017). A general statistical framework for subgroup identification and comparative treatment scoring. *Biometrics*, 73(4):1199–1209.
- Chernozhukov, V., Demirer, M., Duflo, E., and Fernández-Val, I. (2025). Fisher–Schultz lecture: Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in India. *Econometrica*, 93(4):1121–1164.

References (cont.)

- Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. (2008). Nonparametric tests for treatment effect heterogeneity. *The Review of Economics and Statistics*, 90(3):389–405.
- Cui, Y., Kosorok, M. R., Sverdrup, E., Wager, S., and Zhu, R. (2023). Estimating heterogeneous treatment effects with right-censored data via causal survival forests. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(2):179–211.
- Fan, Q., Hsu, Y.-C., Lieli, R. P., and Zhang, Y. (2022). Estimation of conditional average treatment effects with high-dimensional data. *Journal of Business & Economic Statistics*, 40(1):313–327.
- Firpo, S. (2007). Efficient semiparametric estimation of quantile treatment effects. *Econometrica*, 75(1):259–276.
- Foster, D. J. and Syrgkanis, V. (2023). Orthogonal statistical learning. *The Annals of Statistics*, 51(3):879–908.

References (cont.)

- Hahn, P. R., Murray, J. S., and Carvalho, C. M. (2020). Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects. *Bayesian Analysis*, 15(3):965–1056.
- Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240.
- Hines, O. J., Diaz-Ordaz, K., and Vansteelandt, S. (2025). Variable importance measures for heterogeneous treatment effects. *Biometrics*, 81(4).
- Huang, M.-Y. and Chan, K. C. G. (2017). Joint sufficient dimension reduction and estimation of conditional and average treatment effects. *Biometrika*, 104(3):583–596.
- Imai, K. and Li, M. L. (2025). Statistical inference for heterogeneous treatment effects discovered by generic machine learning in randomized experiments. *Journal of Business & Economic Statistics*, 43(1):256–268.

References (cont.)

- Imai, K. and Ratkovic, M. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1):443–470.
- Kennedy, E. H. (2023). Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, 17(2):3008–3049.
- Kennedy, E. H., Balakrishnan, S., Robins, J. M., and Wasserman, L. (2024). Minimax rates for heterogeneous causal effect estimation. *The Annals of Statistics*, 52(2):793–816.
- Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165.
- Lee, S., Okui, R., and Whang, Y.-J. (2017). Doubly robust uniform confidence band for the conditional average treatment effect function. *Journal of Applied Econometrics*, 32(7):1207–1225.

References (cont.)

- Liang, M. and Yu, M. (2022). A semiparametric approach to model effect modification. *Journal of the American Statistical Association*, 117(538):752–764.
- Lu, M., Sadiq, S., Feaster, D. J., and Ishwaran, H. (2018). Estimating individual treatment effect in observational data using random forest methods. *Journal of Computational and Graphical Statistics*, 27(1):209–219.
- Nie, X. and Wager, S. (2021). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319.
- Semenova, V. and Chernozhukov, V. (2021). Debiased machine learning of conditional average treatment effects and other causal functions. *The Econometrics Journal*, 24(2):264–289.
- Su, X., Tsai, C.-L., Wang, H., Nickerson, D. M., and Li, B. (2009). Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research*, 10(5):141–158.

References (cont.)

- Tian, L., Alizadeh, A. A., Gentles, A. J., and Tibshirani, R. (2014). A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association*, 109(508):1517–1532.
- Verdinelli, I. and Wasserman, L. (2024). Decorrelated variable importance. *Journal of Machine Learning Research*, 25(7):1–27.
- Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.
- Yadlowsky, S., Fleming, S., Shah, N., Brunskill, E., and Wager, S. (2025). Evaluating treatment prioritization rules via rank-weighted average treatment effects. *Journal of the American Statistical Association*, 120(549):38–51.
- Yadlowsky, S., Namkoong, H., Basu, S., Duchi, J., and Tian, L. (2022). Bounds on the conditional and average treatment effect with unobserved confounding factors. *The Annals of Statistics*, 50(5):2587–2615.

References (cont.)

- Yang, S., Gao, C., Zeng, D., and Wang, X. (2023). Elastic integrative analysis of randomised trial and real-world data for treatment heterogeneity estimation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(3):575–596.
- Zhao, Q., Small, D. S., and Ertefaie, A. (2022). Selective inference for effect modification via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(2):382–413.