

Causal Inference Meets Optimal Transport

COPSS Lecture

Larry Wasserman
Carnegie Mellon University

Collaborators

- ▶ Siva Balakrishnan
- ▶ Florian Gunsilius
- ▶ Ed Kennedy
- ▶ Arun Kuchibhotla
- ▶ Alex Levis
- ▶ Tudor Manole
- ▶ Alec McClean
- ▶ Gonzalo Mena
- ▶ Jon Niles-Weed
- ▶ Tristan Saidi
- ▶ Kyle Schindl

Other People Working in this Area

- ▶ Medha Agarwal
- ▶ Victor Chernozhukov
- ▶ Alfred Galichon
- ▶ Elizabeth Dunipace
- ▶ Alex Luedtke
- ▶ Hans Mueller
- ▶ Phillip Rigollet
- ▶ William Torous
- ▶ + others

Apologies for anyone I omitted.

Outline

- ▶ Part I: Causal Inference (in one slide)
- ▶ Part II: Optimal Transport
- ▶ Part III: Causal Inference + Optimal Transport
 - ▶ Marginal and conditional effects
 - ▶ Structural Nested Models
 - ▶ Differences-in-Differences (Parallel trends)
 - ▶ Regression Discontinuities
 - ▶ Bounds
 - ▶ Causal Geodesy
- ▶ Part IV: Limitations

Main Takeaway

Main Takeaway

- ▶ Much of causal inference concerns means and differences of means.

Main Takeaway

- ▶ Much of causal inference concerns means and differences of means.
- ▶ Optimal transport provides a convenient language for studying distributional effect. (But not the only one.)

Main Takeaway

- ▶ Much of causal inference concerns means and differences of means.
- ▶ Optimal transport provides a convenient language for studying distributional effect. (But not the only one.)
- ▶ Often permits easy generalization of existing methods.

Main Takeaway

- ▶ Much of causal inference concerns means and differences of means.
- ▶ Optimal transport provides a convenient language for studying distributional effect. (But not the only one.)
- ▶ Often permits easy generalization of existing methods.
- ▶ I will illustrate this with a series of short vignettes.

Part I: Causal Inference

(in one slide)

Causal Inference

- ▶ We have $Z = (X, A, Y)$.

Causal Inference

- ▶ We have $Z = (X, A, Y)$.
- ▶ Outcome Y , treatment/exposure A and measured confounders X .

Causal Inference

- ▶ We have $Z = (X, A, Y)$.
- ▶ Outcome Y , treatment/exposure A and measured confounders X .
- ▶ Counterfactuals $Y(a)$: the value Y would have if $A = a$.

Causal Inference

- ▶ We have $Z = (X, A, Y)$.
- ▶ Outcome Y , treatment/exposure A and measured confounders X .
- ▶ Counterfactuals $Y(a)$: the value Y would have if $A = a$.
- ▶ The usual (tentative) assumptions:

Causal Inference

- ▶ We have $Z = (X, A, Y)$.
- ▶ Outcome Y , treatment/exposure A and measured confounders X .
- ▶ Counterfactuals $Y(a)$: the value Y would have if $A = a$.
- ▶ The usual (tentative) assumptions:
 1. Consistency: $Y = Y(A)$.
 2. Positivity: $\pi(a|x) \geq \epsilon > 0$.
 3. No unmeasured confounding: $Y(a)$ is independent of A given X .

Causal Inference

- ▶ We have $Z = (X, A, Y)$.
- ▶ Outcome Y , treatment/exposure A and measured confounders X .
- ▶ Counterfactuals $Y(a)$: the value Y would have if $A = a$.
- ▶ The usual (tentative) assumptions:
 1. Consistency: $Y = Y(A)$.
 2. Positivity: $\pi(a|x) \geq \epsilon > 0$.
 3. No unmeasured confounding: $Y(a)$ is independent of A given X .
- ▶ Then we have identification by way of the g -formula:

$$P(Y(a) \leq y) = \int P(Y \leq y | A = a, X = x) dP(x)$$

Causal Inference

- ▶ We have $Z = (X, A, Y)$.
- ▶ Outcome Y , treatment/exposure A and measured confounders X .
- ▶ Counterfactuals $Y(a)$: the value Y would have if $A = a$.
- ▶ The usual (tentative) assumptions:
 1. Consistency: $Y = Y(A)$.
 2. Positivity: $\pi(a|x) \geq \epsilon > 0$.
 3. No unmeasured confounding: $Y(a)$ is independent of A given X .
- ▶ Then we have identification by way of the g -formula:

$$P(Y(a) \leq y) = \int P(Y \leq y | A = a, X = x) dP(x)$$

- ▶ Hence

$$\mathbb{E}[Y(a)] = \int \mu(x, a) dP(x)$$

where $\mu(x, a) = \mathbb{E}[Y | X = x, A = a]$.

Causal Inference

- ▶ We have $Z = (X, A, Y)$.
- ▶ Outcome Y , treatment/exposure A and measured confounders X .
- ▶ Counterfactuals $Y(a)$: the value Y would have if $A = a$.
- ▶ The usual (tentative) assumptions:
 1. Consistency: $Y = Y(A)$.
 2. Positivity: $\pi(a|x) \geq \epsilon > 0$.
 3. No unmeasured confounding: $Y(a)$ is independent of A given X .
- ▶ Then we have identification by way of the g -formula:

$$P(Y(a) \leq y) = \int P(Y \leq y | A = a, X = x) dP(x)$$

- ▶ Hence

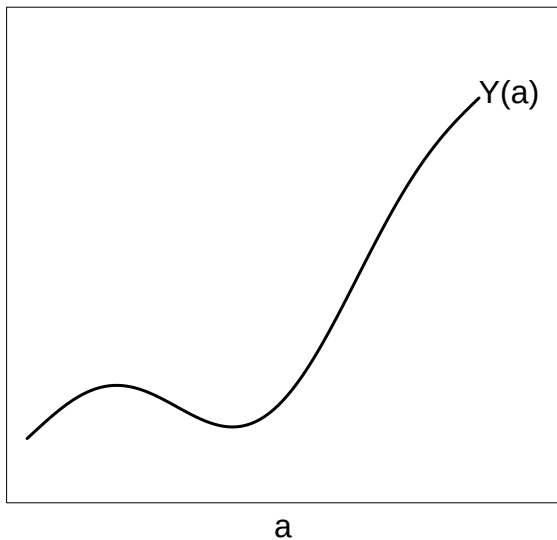
$$\mathbb{E}[Y(a)] = \int \mu(x, a) dP(x)$$

where $\mu(x, a) = \mathbb{E}[Y | X = x, A = a]$.

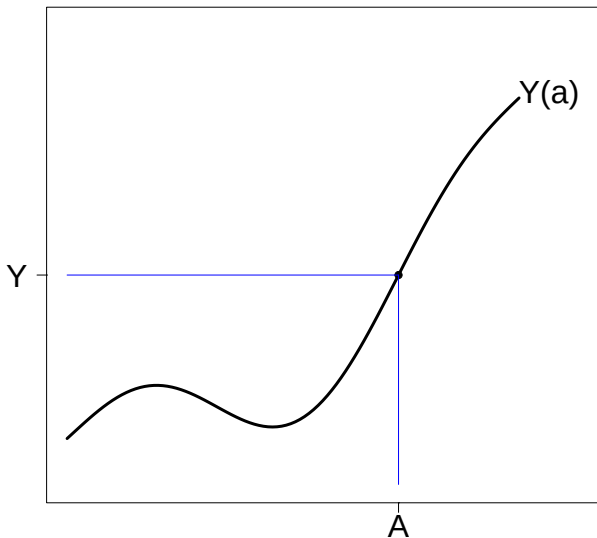
- ▶ Of course,

$$\mathbb{E}[Y(a)] = \int \mu(x, a) dP(x) \neq \int \mu(x, a) dP(x|a) = \mathbb{E}[Y | A = a].$$

The Consistency Assumption



The Consistency Assumption

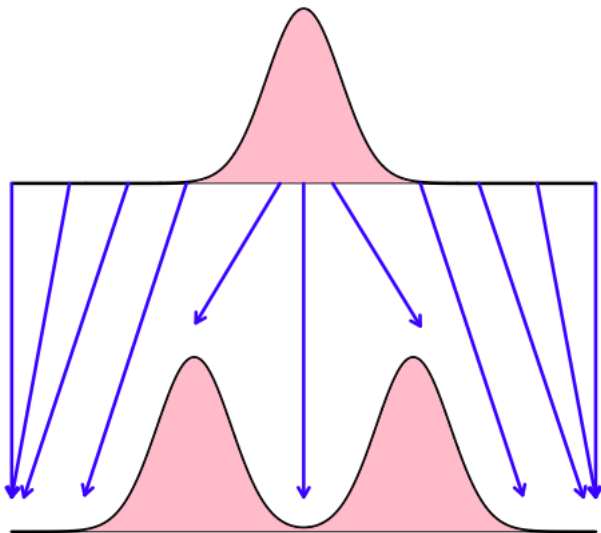


Part II: Optimal Transport

Reading Material

- Chewi, Niles-Weed and Rigollet (2025). *Statistical Optimal Transport*.
- Panaretos and Zemel (2020). *An Invitation to Statistics in Wasserstein Space*.
- Gunsilius (2025). *A Primer on Optimal Transport for Causal Inference with Observational Data*.
- Balakrishnan, Manole and Wasserman (2025). *Statistical inference for optimal transport maps: Recent advances and perspectives*.

Optimal Transport: Monge 1781



Optimal Transport

Optimal Transport

- ▶ Given μ_0 and μ_1 find $T(x)$ such that

$$X \sim \mu_0 \implies T(X) \sim \mu_1.$$

Write $\mu_1 = T_{\#}\mu_0$.

Optimal Transport

- ▶ Given μ_0 and μ_1 find $T(x)$ such that

$$X \sim \mu_0 \implies T(X) \sim \mu_1.$$

Write $\mu_1 = T_{\#}\mu_0$.

- ▶ There are infinitely many such **transport** maps.

Optimal Transport

- ▶ Given μ_0 and μ_1 find $T(x)$ such that

$$X \sim \mu_0 \implies T(X) \sim \mu_1.$$

Write $\mu_1 = T_{\#}\mu_0$.

- ▶ There are infinitely many such **transport** maps.
- ▶ The **optimal transport map** (or Monge map) minimizes

$$\mathbb{E}[\|T(X) - X\|^2]$$

subject to $\mu_1 = T_{\#}\mu_0$. (We will only use the L_2 loss.)

Optimal Transport

- ▶ Given μ_0 and μ_1 find $T(x)$ such that

$$X \sim \mu_0 \implies T(X) \sim \mu_1.$$

Write $\mu_1 = T_{\#}\mu_0$.

- ▶ There are infinitely many such **transport** maps.
- ▶ The **optimal transport map** (or Monge map) minimizes

$$\mathbb{E}[\|T(X) - X\|^2]$$

subject to $\mu_1 = T_{\#}\mu_0$. (We will only use the L_2 loss.)

- ▶ The minimum value is the **Wasserstein distance**

$$W^2(\mu_0, \mu_1) = \inf_{T: \mu_1 = T_{\#}\mu_0} \mathbb{E}[\|T(X) - X\|^2].$$

Optimal Transport

Optimal Transport

- ▶ One dimensional case: $X_0 \sim F_0$ and $X_1 \sim F_1$ then

$$T(x) = F_1^{-1}(F_0(x))$$

and

$$W^2(\mu_0, \mu_1) = \int_0^1 |F_0^{-1}(u) - F_1^{-1}(u)|^2 du.$$

Optimal Transport

- ▶ One dimensional case: $X_0 \sim F_0$ and $X_1 \sim F_1$ then

$$T(x) = F_1^{-1}(F_0(x))$$

and

$$W^2(\mu_0, \mu_1) = \int_0^1 |F_0^{-1}(u) - F_1^{-1}(u)|^2 du.$$

- ▶ Gaussian case: $X_0 \sim N(\mu_0, \Sigma_0)$ and $X_1 \sim N(\mu_1, \Sigma_1)$ then

$$T(x_0) = \mu_1 + A(x_0 - \mu_0)$$

where

$$A = \Sigma_0^{-1/2}(\Sigma_0^{1/2}\Sigma_1\Sigma_0^{1/2})^{1/2}\Sigma_0^{-1/2}$$

and

$$W^2(\mu_0, \mu_1) = \|\mu_0 - \mu_1\|^2 + \text{tr}(\Sigma_0 + \Sigma_1 - 2(\Sigma_0^{1/2}\Sigma_1\Sigma_0^{1/2})^{1/2}).$$

Optimal Transport

Optimal Transport

- ▶ Two empirical distributions:

$$\mu_0 = \text{uniform on } X_1, \dots, X_n$$

and

$$\mu_1 = \text{uniform on } Y_1, \dots, Y_n$$

Optimal Transport

- ▶ Two empirical distributions:

$$\mu_0 = \text{uniform on } X_1, \dots, X_n$$

and

$$\mu_1 = \text{uniform on } Y_1, \dots, Y_n$$

- ▶ Then

$$W^2(\mu_0, \mu_1) = \min_{\pi} \sum_i \|X_i - Y_{\pi(i)}\|^2$$

where the minimum is over all permutations

Optimal Transport

- ▶ Two empirical distributions:

$$\mu_0 = \text{uniform on } X_1, \dots, X_n$$

and

$$\mu_1 = \text{uniform on } Y_1, \dots, Y_n$$

- ▶ Then

$$W^2(\mu_0, \mu_1) = \min_{\pi} \sum_i \|X_i - Y_{\pi(i)}\|^2$$

where the minimum is over all permutations

- ▶ Can be computed in $O(n^3)$ times

Optimal Transport

- ▶ Two empirical distributions:

$$\mu_0 = \text{uniform on } X_1, \dots, X_n$$

and

$$\mu_1 = \text{uniform on } Y_1, \dots, Y_n$$

- ▶ Then

$$W^2(\mu_0, \mu_1) = \min_{\pi} \sum_i \|X_i - Y_{\pi(i)}\|^2$$

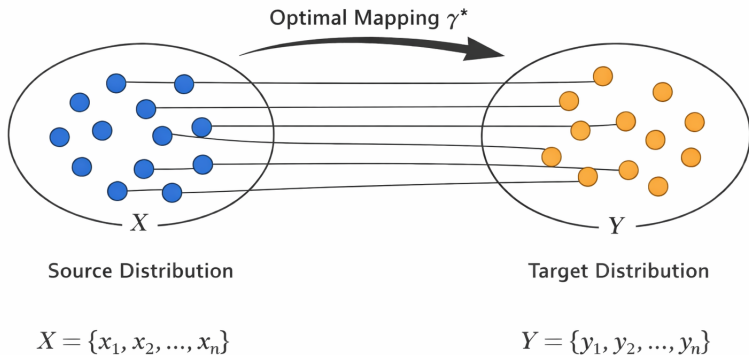
where the minimum is over all permutations

- ▶ Can be computed in $O(n^3)$ times
- ▶ And

$$T(X_i) = Y_{\pi(i)}.$$

Empirical Transport

Optimal Transport Between Datasets



Optimal Transport

Optimal Transport

- ▶ In general, there is no simple formula for T .

Optimal Transport

- ▶ In general, there is no simple formula for T .
- ▶ This makes implementation (and inference) difficult.

Optimal Transport

- ▶ In general, there is no simple formula for T .
- ▶ This makes implementation (and inference) difficult.
- ▶ Brenier's theorem: $T(x) = \nabla\phi$ where $\phi(x)$ minimizes

$$\int \phi(x) d\mu_0(x) + \int \phi^*(y) d\mu_1(y)$$

over all convex functions, and where $\phi^*(y) = \sup_x \{x^T y - \phi(x)\}$ is the convex conjugate.

Optimal Transport

- ▶ In general, there is no simple formula for T .
- ▶ This makes implementation (and inference) difficult.
- ▶ Brenier's theorem: $T(x) = \nabla\phi$ where $\phi(x)$ minimizes

$$\int \phi(x) d\mu_0(x) + \int \phi^*(y) d\mu_1(y)$$

over all convex functions, and where $\phi^*(y) = \sup_x \{x^T y - \phi(x)\}$ is the convex conjugate.

- ▶ In principle, one can solve the sample version (M-estimation) but this is still difficult.

Wasserstein Distance

Wasserstein Distance

- ▶ The Wasserstein distance W has very nice properties.

Wasserstein Distance

- ▶ The Wasserstein distance W has very nice properties.
- ▶ It is on the same scale as the random variables X_0 and X_1 .

Wasserstein Distance

- ▶ The Wasserstein distance W has very nice properties.
- ▶ It is on the same scale as the random variables X_0 and X_1 .
- ▶ It respects the geometry of the sample space. For example, if δ_a is a point mass at a and δ_b is a point mass at b then
$$W(\delta_a, \delta_b) = \|a - b\|.$$

Wasserstein Distance

- ▶ The Wasserstein distance W has very nice properties.
- ▶ It is on the same scale as the random variables X_0 and X_1 .
- ▶ It respects the geometry of the sample space. For example, if δ_a is a point mass at a and δ_b is a point mass at b then $W(\delta_a, \delta_b) = ||a - b||$.



Wasserstein Distance

- ▶ The Wasserstein distance W has very nice properties.
- ▶ It is on the same scale as the random variables X_0 and X_1 .
- ▶ It respects the geometry of the sample space. For example, if δ_a is a point mass at a and δ_b is a point mass at b then $W(\delta_a, \delta_b) = ||a - b||$.



- ▶ In contrast, the total variation distance is 1 no matter how close/far a and b are. Similarly, for Hellinger, χ^2 , L_2 etc. (Horizontal vs vertical distance).

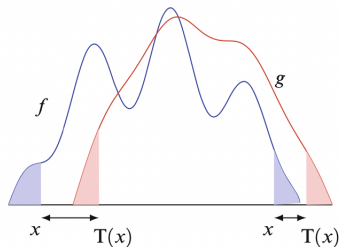
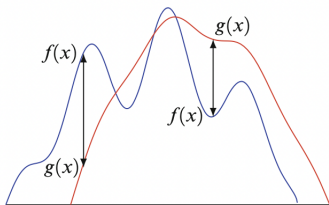
Wasserstein Distance

- ▶ The Wasserstein distance W has very nice properties.
- ▶ It is on the same scale as the random variables X_0 and X_1 .
- ▶ It respects the geometry of the sample space. For example, if δ_a is a point mass at a and δ_b is a point mass at b then $W(\delta_a, \delta_b) = ||a - b||$.



- ▶ In contrast, the total variation distance is 1 no matter how close/far a and b are. Similarly, for Hellinger, χ^2 , L_2 etc. (Horizontal vs vertical distance).
- ▶ Can compare discrete and continuous distributions.

Wasserstein vs Hellinger/Total Variation



From Santambrogio (2015).

Estimation

Estimation

- In one dimension, estimation is easy: $T(x) = F_1^{-1}(F_0(x))$ and

$$\hat{T}(x) = \hat{F}_1^{-1}(\hat{F}_0(x)).$$

Estimation

- ▶ In one dimension, estimation is easy: $T(x) = F_1^{-1}(F_0(x))$ and

$$\hat{T}(x) = \hat{F}_1^{-1}(\hat{F}_0(x)).$$

- ▶ In the multivariate case, estimating the map T and the distance W is difficult (computationally and statistically).

Estimation

- ▶ In one dimension, estimation is easy: $T(x) = F_1^{-1}(F_0(x))$ and

$$\hat{T}(x) = \hat{F}_1^{-1}(\hat{F}_0(x)).$$

- ▶ In the multivariate case, estimating the map T and the distance W is difficult (computationally and statistically).
- ▶ $X_1, \dots, X_n \sim \mu_0$ and $Y_1, \dots, Y_n \sim \mu_1$ then we can estimate $W(\mu_0, \mu_1)$ by $W(\hat{\mu}_0, \hat{\mu}_1)$ where

$$\hat{\mu}_0 = \frac{1}{n} \sum_i \delta_{X_i}, \quad \hat{\mu}_1 = \frac{1}{n} \sum_i \delta_{Y_i}.$$

Estimation

- ▶ In one dimension, estimation is easy: $T(x) = F_1^{-1}(F_0(x))$ and

$$\hat{T}(x) = \hat{F}_1^{-1}(\hat{F}_0(x)).$$

- ▶ In the multivariate case, estimating the map T and the distance W is difficult (computationally and statistically).
- ▶ $X_1, \dots, X_n \sim \mu_0$ and $Y_1, \dots, Y_n \sim \mu_1$ then we can estimate $W(\mu_0, \mu_1)$ by $W(\hat{\mu}_0, \hat{\mu}_1)$ where

$$\hat{\mu}_0 = \frac{1}{n} \sum_i \delta_{X_i}, \quad \hat{\mu}_1 = \frac{1}{n} \sum_i \delta_{Y_i}.$$

- ▶ This is a poor estimator. Very biased, slow rate (Weed and Bach 2019, Hutter and Rigollet 2021, Chewi, Weed, Rigollet 2025).

Estimation

- ▶ In one dimension, estimation is easy: $T(x) = F_1^{-1}(F_0(x))$ and

$$\hat{T}(x) = \hat{F}_1^{-1}(\hat{F}_0(x)).$$

- ▶ In the multivariate case, estimating the map T and the distance W is difficult (computationally and statistically).
- ▶ $X_1, \dots, X_n \sim \mu_0$ and $Y_1, \dots, Y_n \sim \mu_1$ then we can estimate $W(\mu_0, \mu_1)$ by $W(\hat{\mu}_0, \hat{\mu}_1)$ where

$$\hat{\mu}_0 = \frac{1}{n} \sum_i \delta_{X_i}, \quad \hat{\mu}_1 = \frac{1}{n} \sum_i \delta_{Y_i}.$$

- ▶ This is a poor estimator. Very biased, slow rate (Weed and Bach 2019, Hutter and Rigollet 2021, Chewi, Weed, Rigollet 2025).
- ▶ We would like to leverage smoothness, but doing so is not easy. See Hutter and Rigollet (2019), Manole et al (2021, 2023), Chewi, Niles-Weed and Rigollet (2024), Balakrishnan, Manole, Wasserman (2025), Divol, Niles-Weed, Pooladian (2022), + others

Estimation

- ▶ In one dimension, estimation is easy: $T(x) = F_1^{-1}(F_0(x))$ and

$$\hat{T}(x) = \hat{F}_1^{-1}(\hat{F}_0(x)).$$

- ▶ In the multivariate case, estimating the map T and the distance W is difficult (computationally and statistically).
- ▶ $X_1, \dots, X_n \sim \mu_0$ and $Y_1, \dots, Y_n \sim \mu_1$ then we can estimate $W(\mu_0, \mu_1)$ by $W(\hat{\mu}_0, \hat{\mu}_1)$ where

$$\hat{\mu}_0 = \frac{1}{n} \sum_i \delta_{X_i}, \quad \hat{\mu}_1 = \frac{1}{n} \sum_i \delta_{Y_i}.$$

- ▶ This is a poor estimator. Very biased, slow rate (Weed and Bach 2019, Hutter and Rigollet 2021, Chewi, Weed, Rigollet 2025).
- ▶ We would like to leverage smoothness, but doing so is not easy. See Hutter and Rigollet (2019), Manole et al (2021, 2023), Chewi, Niles-Weed and Rigollet (2024), Balakrishnan, Manole, Wasserman (2025), Divol, Niles-Weed, Pooladian (2022), + others
- ▶ Approximate versions: entropic regularized transport (Cuturi 2013, Rigollet and Weed 2018 +) and smoothed transport (Goldfeld, Kato, Nietert, Rioux 2022 +)

L -Moments

- For interpretation, we can decompose the Wasserstein distance using L -moments (Schindl and Wasserman 2026).

L -Moments

- ▶ For interpretation, we can decompose the Wasserstein distance using L -moments (Schindl and Wasserman 2026).
- ▶ Hosking (1990) introduced the L moments:

$$\lambda_1 = \mathbb{E}[Z]$$

$$\lambda_2 = \mathbb{E}[Z_{2:2} - Z_{1:2}]$$

$$\lambda_3 = \mathbb{E}[(Z_{3:3} - Z_{2:3}) - (Z_{2:3} - Z_{1:3})]$$

$$\lambda_4 = \mathbb{E}[(Z_{4:4} - Z_{3:4}) - 2(Z_{3:4} - Z_{2:4}) + (Z_{2:4} - Z_{1:4})]$$

etc.

L-Moments

- ▶ For interpretation, we can decompose the Wasserstein distance using L -moments (Schindl and Wasserman 2026).
- ▶ Hosking (1990) introduced the L moments:

$$\lambda_1 = \mathbb{E}[Z]$$

$$\lambda_2 = \mathbb{E}[Z_{2:2} - Z_{1:2}]$$

$$\lambda_3 = \mathbb{E}[(Z_{3:3} - Z_{2:3}) - (Z_{2:3} - Z_{1:3})]$$

$$\lambda_4 = \mathbb{E}[(Z_{4:4} - Z_{3:4}) - 2(Z_{3:4} - Z_{2:4}) + (Z_{2:4} - 1 : 4)]$$

etc.

- ▶ In general

$$\lambda_k = \int_0^1 Q(u) P_{k-1}(u) du$$

where Q is the quantile function and P_k are (shifted) Legendre polynomials so $Q(u) = \sum_{k=1}^{\infty} \lambda_k P_{k-1}(u)$.

L -Moments

L -Moments

- It follows that

$$W^2(P_0, P_1) = \sum_{k=0}^{\infty} (2k-1)(\lambda_k^{(1)} - \lambda_k^{(0)})^2.$$

L-Moments

- It follows that

$$W^2(P_0, P_1) = \sum_{k=0}^{\infty} (2k+1)(\lambda_k^{(1)} - \lambda_k^{(0)})^2.$$

- We can see how much of the Wasserstein distance is due to location, scale, skewness etc.

L-Moments

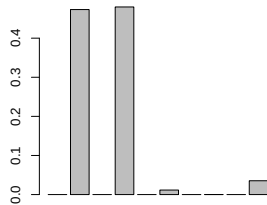
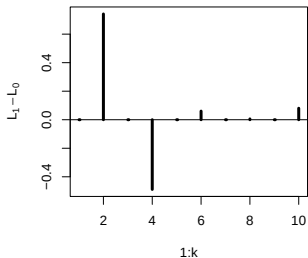
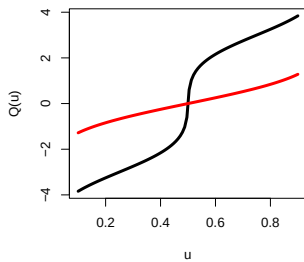
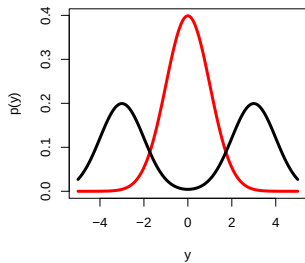
- It follows that

$$W^2(P_0, P_1) = \sum_{k=0}^{\infty} (2k-1)(\lambda_k^{(1)} - \lambda_k^{(0)})^2.$$

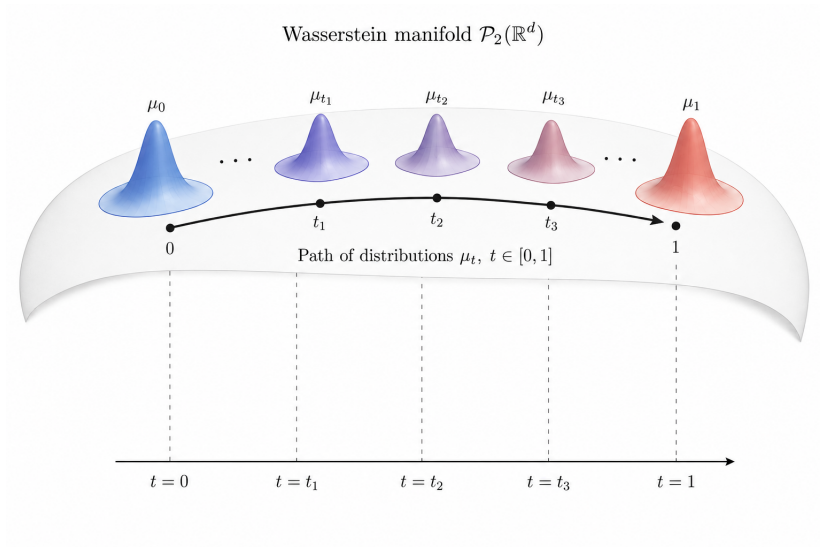
- We can see how much of the Wasserstein distance is due to location, scale, skewness etc.
- Distributional R^2 :

$$R_k^2 = \frac{(2k-1)(\lambda_k^{(1)} - \lambda_k^{(0)})^2}{W^2(P_0, P_1)}.$$

L-Moments



Paths



Paths

Paths

- ▶ Eulerian description: given a path of distributions $(\mu_t : 0 \leq t \leq 1)$ there is a (minimal) vector field v_t that drives the path:

$$\partial_t \mu_t(x) = -\operatorname{div}_x(\mu_t(x)v_t(x)).$$

Paths

- ▶ Eulerian description: given a path of distributions $(\mu_t : 0 \leq t \leq 1)$ there is a (minimal) vector field v_t that drives the path:

$$\partial_t \mu_t(x) = -\operatorname{div}_x(\mu_t(x)v_t(x)).$$

- ▶ Lagrangian description: if

$$\dot{X}_t = v_t(X_t), \quad X_0 \sim \mu_0$$

then $X_t \sim \mu_t$.

Paths

- ▶ Eulerian description: given a path of distributions $(\mu_t : 0 \leq t \leq 1)$ there is a (minimal) vector field v_t that drives the path:

$$\partial_t \mu_t(x) = -\operatorname{div}_x(\mu_t(x) v_t(x)).$$

- ▶ Lagrangian description: if

$$\dot{X}_t = v_t(X_t), \quad X_0 \sim \mu_0$$

then $X_t \sim \mu_t$.

- ▶ v_t is the derivative of the local transport map:

$$v_t(y) = \lim_{\epsilon \rightarrow 0} \frac{T_{t,t+\epsilon}(y) - y}{\epsilon}$$

where $T_{t,t+\epsilon}$ is the optimal transport map from μ_t to $\mu_{t+\epsilon}$.

Paths

- ▶ Eulerian description: given a path of distributions $(\mu_t : 0 \leq t \leq 1)$ there is a (minimal) vector field v_t that drives the path:

$$\partial_t \mu_t(x) = -\operatorname{div}_x(\mu_t(x) v_t(x)).$$

- ▶ Lagrangian description: if

$$\dot{X}_t = v_t(X_t), \quad X_0 \sim \mu_0$$

then $X_t \sim \mu_t$.

- ▶ v_t is the derivative of the local transport map:

$$v_t(y) = \lim_{\epsilon \rightarrow 0} \frac{T_{t,t+\epsilon}(y) - y}{\epsilon}$$

where $T_{t,t+\epsilon}$ is the optimal transport map from μ_t to $\mu_{t+\epsilon}$.

- ▶ Given μ_0 and μ_1 there is a shortest path (geodesic) that connects them $(\mu_t : 0 \leq t \leq 1)$.

Why the Monge Map?

Why the Monge Map?

- ▶ There are many transport maps. Why privilege the Monge map?

Why the Monge Map?

- ▶ There are many transport maps. Why privilege the Monge map?
- ▶ Minimizes a cost.

Why the Monge Map?

- ▶ There are many transport maps. Why privilege the Monge map?
- ▶ Minimizes a cost.
- ▶ Polar factorization theorem (Brenier 1991): if S is a transport map then $S = T \circ M$ where T is the Monge (OT) map and M is measure preserving (doesn't change the distribution).

Why the Monge Map?

- ▶ There are many transport maps. Why privilege the Monge map?
- ▶ Minimizes a cost.
- ▶ Polar factorization theorem (Brenier 1991): if S is a transport map then $S = T \circ M$ where T is the Monge (OT) map and M is measure preserving (doesn't change the distribution).
- ▶ Balakrishnan (2026): the minimax risk for estimating any transport map is at least as large as the minimax risk for estimating the Monge map.

Why the Monge Map?

- ▶ There are many transport maps. Why privilege the Monge map?
- ▶ Minimizes a cost.
- ▶ Polar factorization theorem (Brenier 1991): if S is a transport map then $S = T \circ M$ where T is the Monge (OT) map and M is measure preserving (doesn't change the distribution).
- ▶ Balakrishnan (2026): the minimax risk for estimating any transport map is at least as large as the minimax risk for estimating the Monge map.
- ▶ It gives the Wasserstein distance.

Part III

Causal Inference

+

Optimal Transport

Basic Causal Effects

Basic Causal Effects

- ▶ Suppose $A \in \{0, 1\}$. The counterfactuals are $Y(0)$ and $Y(1)$.

Basic Causal Effects

- ▶ Suppose $A \in \{0, 1\}$. The counterfactuals are $Y(0)$ and $Y(1)$.
- ▶ The mean causal effect is

$$\mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] = \int (\mathbb{E}[Y|X = x, A = 1] - \mathbb{E}[Y|X = x, A = 0]) dP_X(x).$$

Basic Causal Effects

- ▶ Suppose $A \in \{0, 1\}$. The counterfactuals are $Y(0)$ and $Y(1)$.
- ▶ The mean causal effect is

$$\mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] = \int (\mathbb{E}[Y|X = x, A = 1] - \mathbb{E}[Y|X = x, A = 0]) dP_X(x).$$

- ▶ Distributional effect (Wasserstein distance):

$$\psi^2 = W^2(P_{Y(0)}, P_{Y(1)}) = \int |F_{Y(1)}^{-1}(u) - F_{Y(0)}^{-1}(u)|^2 du.$$

where

$$F_{Y(a)}(y) = \int F_{Y|X,A}(y|X = x, A = a) dP_X(x) \quad a = 0, 1.$$

Basic Causal Effects

- ▶ Suppose $A \in \{0, 1\}$. The counterfactuals are $Y(0)$ and $Y(1)$.
- ▶ The mean causal effect is

$$\mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] = \int (\mathbb{E}[Y|X = x, A = 1] - \mathbb{E}[Y|X = x, A = 0]) dP_X(x).$$

- ▶ Distributional effect (Wasserstein distance):

$$\psi^2 = W^2(P_{Y(0)}, P_{Y(1)}) = \int |F_{Y(1)}^{-1}(u) - F_{Y(0)}^{-1}(u)|^2 du.$$

where

$$F_{Y(a)}(y) = \int F_{Y|X,A}(y|X = x, A = a) dP_X(x) \quad a = 0, 1.$$

- ▶ The estimate:

$$\hat{\psi}^2 = \int |\hat{F}_1^{-1}(u) - \hat{F}_0^{-1}(u)|^2 du$$

Basic Causal Effects

- ▶ Suppose $A \in \{0, 1\}$. The counterfactuals are $Y(0)$ and $Y(1)$.
- ▶ The mean causal effect is

$$\mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] = \int (\mathbb{E}[Y|X = x, A = 1] - \mathbb{E}[Y|X = x, A = 0]) dP_X(x).$$

- ▶ Distributional effect (Wasserstein distance):

$$\psi^2 = W^2(P_{Y(0)}, P_{Y(1)}) = \int |F_{Y(1)}^{-1}(u) - F_{Y(0)}^{-1}(u)|^2 du.$$

where

$$F_{Y(a)}(y) = \int F_{Y|X,A}(y|X = x, A = a) dP_X(x) \quad a = 0, 1.$$

- ▶ The estimate:

$$\hat{\psi}^2 = \int |\hat{F}_1^{-1}(u) - \hat{F}_0^{-1}(u)|^2 du$$

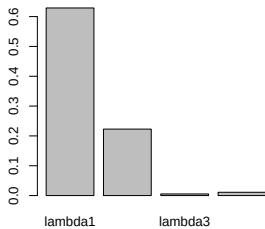
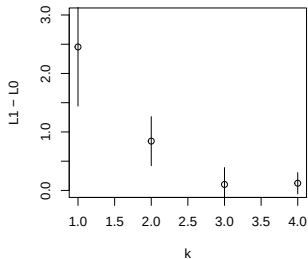
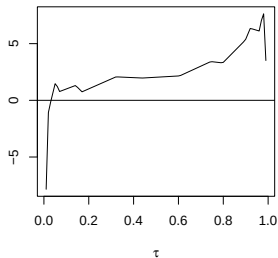
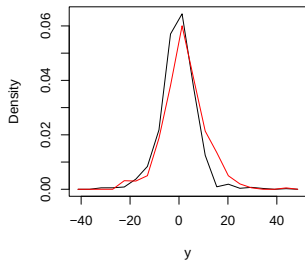
- ▶ Conditional effect:

$$\psi(x) = W(P_{Y(1)|X=x}, P_{Y(0)|X=x})$$

where

$$F_{Y(a)|X=x}(y) = F_{Y|X,A}(y|x, a).$$

Example: Weight Gain from Quitting Smoking



Conditional Distributional Effect

Conditional Distributional Effect

- ▶ We want the conditional effect

$$\psi^2(x) = W^2(P_{Y(1)|X=x}, P_{Y(0)|X=x})$$

or

$$\psi^2(v) = W^2(P_{Y(1)|V=v}, P_{Y(0)|V=v})$$

where V is a subset of X .

Conditional Distributional Effect

- ▶ We want the conditional effect

$$\psi^2(x) = W^2(P_{Y(1)|X=x}, P_{Y(0)|X=x})$$

or

$$\psi^2(v) = W^2(P_{Y(1)|V=v}, P_{Y(0)|V=v})$$

where V is a subset of X .

- ▶ We also want the summary effect

$$\Psi^2 = \int \psi^2(x) dP_X(x).$$

Conditional Distributional Effect

- ▶ We want the conditional effect

$$\psi^2(x) = W^2(P_{Y(1)|X=x}, P_{Y(0)|X=x})$$

or

$$\psi^2(v) = W^2(P_{Y(1)|V=v}, P_{Y(0)|V=v})$$

where V is a subset of X .

- ▶ We also want the summary effect

$$\Psi^2 = \int \psi^2(x) dP_X(x).$$

- ▶ Note that

$$W^2(P_{Y(1)}, P_{Y(0)}) \leq \Psi^2$$

due to treatment heterogeneity.

Conditional Distributional Effect

- ▶ We want the conditional effect

$$\psi^2(x) = W^2(P_{Y(1)|X=x}, P_{Y(0)|X=x})$$

or

$$\psi^2(v) = W^2(P_{Y(1)|V=v}, P_{Y(0)|V=v})$$

where V is a subset of X .

- ▶ We also want the summary effect

$$\Psi^2 = \int \psi^2(x) dP_X(x).$$

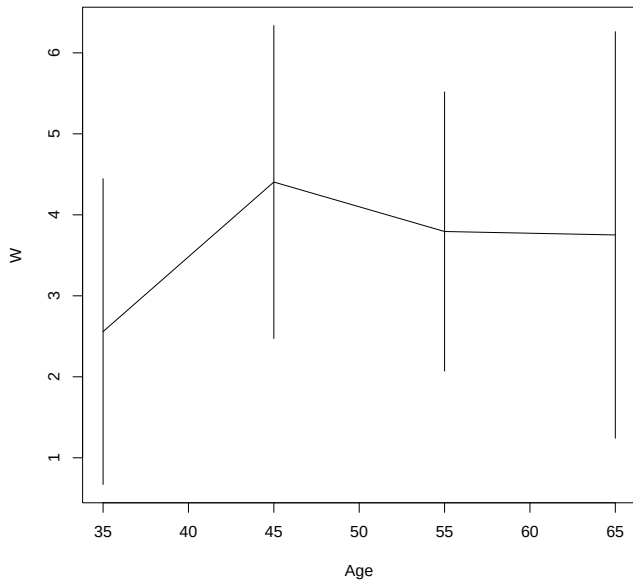
- ▶ Note that

$$W^2(P_{Y(1)}, P_{Y(0)}) \leq \Psi^2$$

due to treatment heterogeneity.

- ▶ An RKHS version Jain and Luedtke (2026).

Example



Implicit OT: Structural Nested Models

Implicit OT: Structural Nested Models

- ▶ Structural Nested Models are semiparametric model for causal contrasts due to Robins (1994).

Implicit OT: Structural Nested Models

- ▶ Structural Nested Models are semiparametric model for causal contrasts due to Robins (1994).
- ▶ Mean model:

$$\mathbb{E}[Y(a)|X = x, A = a] - \mathbb{E}[Y(0)|X = x, A = a] = m(x, a, y; \psi).$$

Implicit OT: Structural Nested Models

- ▶ Structural Nested Models are semiparametric model for causal contrasts due to Robins (1994).
- ▶ Mean model:

$$\mathbb{E}[Y(a)|X = x, A = a] - \mathbb{E}[Y(0)|X = x, A = a] = m(x, a, y; \psi).$$

- ▶ Distributional model:

$$F_{Y(0)|X=x,A=a}^{-1}(F_{Y(a)|X=x,A=a}(y)) = m(x, a, y; \psi).$$

Implicit OT: Structural Nested Models

- ▶ Structural Nested Models are semiparametric model for causal contrasts due to Robins (1994).
- ▶ Mean model:

$$\mathbb{E}[Y(a)|X = x, A = a] - \mathbb{E}[Y(0)|X = x, A = a] = m(x, a, y; \psi).$$

- ▶ Distributional model:

$$F_{Y(0)|X=x, A=a}^{-1}(F_{Y(a)|X=x, A=a}(y)) = m(x, a, y; \psi).$$

- ▶ This is precisely the (conditional) optimal transport map from $Y(a)$ to $Y(0)$.

Implicit OT: Structural Nested Models

- ▶ Structural Nested Models are semiparametric model for causal contrasts due to Robins (1994).
- ▶ Mean model:

$$\mathbb{E}[Y(a)|X = x, A = a] - \mathbb{E}[Y(0)|X = x, A = a] = m(x, a, y; \psi).$$

- ▶ Distributional model:

$$F_{Y(0)|X=x, A=a}^{-1}(F_{Y(a)|X=x, A=a}(y)) = m(x, a, y; \psi).$$

- ▶ This is precisely the (conditional) optimal transport map from $Y(a)$ to $Y(0)$.
- ▶ Quantiles don't generalize easily to multivariate Y but optimal transport maps do.

Implicit OT: Structural Nested Models

- ▶ Structural Nested Models are semiparametric model for causal contrasts due to Robins (1994).
- ▶ Mean model:

$$\mathbb{E}[Y(a)|X = x, A = a] - \mathbb{E}[Y(0)|X = x, A = a] = m(x, a, y; \psi).$$

- ▶ Distributional model:

$$F_{Y(0)|X=x, A=a}^{-1}(F_{Y(a)|X=x, A=a}(y)) = m(x, a, y; \psi).$$

- ▶ This is precisely the (conditional) optimal transport map from $Y(a)$ to $Y(0)$.
- ▶ Quantiles don't generalize easily to multivariate Y but optimal transport maps do.
- ▶ So we can define a structural nested model for multivariate Y by defining

$$T_{Y(a) \rightarrow Y(0)}(y|x, a) = m(x, a, y; \psi).$$

Inference for ψ proceeds in much the same way as in Robins (1994).

Continuous Time SNM

- ▶ Lok (2008, 2107) introduced continuous time SNM's.

Continuous Time SNM

- ▶ Lok (2008, 2107) introduced continuous time SNM's.
- ▶ Let $Y(t)$ be the counterfactual outcome of treatment were stopped at time t .

Continuous Time SNM

- ▶ Lok (2008, 2107) introduced continuous time SNM's.
- ▶ Let $Y(t)$ be the counterfactual outcome of treatment were stopped at time t .
- ▶ Given a history $\bar{Z}_t$, she defines

$$D(y, t; \bar{Z}_t) = \left. \frac{\partial}{\partial h} \right|_{h=0} (F_{Y(t+h)|\bar{Z}_t}^{-1} \circ F_{Y(t)|\bar{Z}_t})(y).$$

Continuous Time SNM

- ▶ Lok (2008, 2107) introduced continuous time SNM's.
- ▶ Let $Y(t)$ be the counterfactual outcome of treatment were stopped at time t .
- ▶ Given a history $\bar{Z}_t$, she defines

$$D(y, t; \bar{Z}_t) = \left. \frac{\partial}{\partial h} \right|_{h=0} (F_{Y(t+h)|\bar{Z}_t}^{-1} \circ F_{Y(t)|\bar{Z}_t})(y).$$

- ▶ She shows that, given a parametric model for D (a continuous time SNM), one can estimate the causal effect at a $n^{-1/2}$ rate using appropriate estimating equations.

Continuous Time SNM

- ▶ Lok (2008, 2107) introduced continuous time SNM's.
- ▶ Let $Y(t)$ be the counterfactual outcome of treatment were stopped at time t .
- ▶ Given a history $\bar{Z}_t$, she defines

$$D(y, t; \bar{Z}_t) = \left. \frac{\partial}{\partial h} \right|_{h=0} (F_{Y(t+h)|\bar{Z}_t}^{-1} \circ F_{Y(t)|\bar{Z}_t})(y).$$

- ▶ She shows that, given a parametric model for D (a continuous time SNM), one can estimate the causal effect at a $n^{-1/2}$ rate using appropriate estimating equations.
- ▶ But now we notice that D is precisely the Wasserstein vector field v_t corresponding to the path of distributions $F_{Y(t)|\bar{Z}_t}$.

Continuous Time SNM

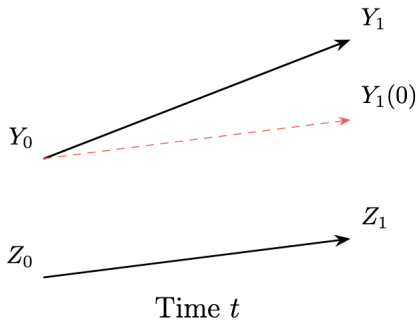
- ▶ Lok (2008, 2107) introduced continuous time SNM's.
- ▶ Let $Y(t)$ be the counterfactual outcome of treatment were stopped at time t .
- ▶ Given a history $\bar{Z}_t$, she defines

$$D(y, t; \bar{Z}_t) = \left. \frac{\partial}{\partial h} \right|_{h=0} (F_{Y(t+h)|\bar{Z}_t}^{-1} \circ F_{Y(t)|\bar{Z}_t})(y).$$

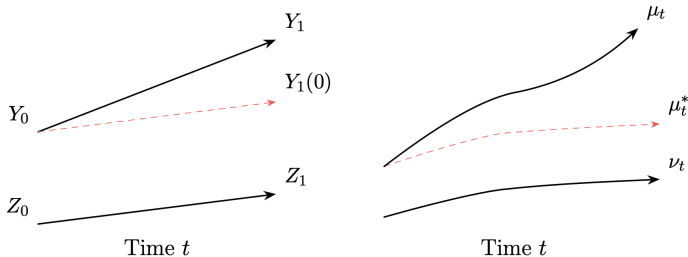
- ▶ She shows that, given a parametric model for D (a continuous time SNM), one can estimate the causal effect at a $n^{-1/2}$ rate using appropriate estimating equations.
- ▶ But now we notice that D is precisely the Wasserstein vector field v_t corresponding to the path of distributions $F_{Y(t)|\bar{Z}_t}$.
- ▶ This perspective suggests some generalizations of SNM's to: arbitrary continuous treatments, continuous time and multivariate outcomes.

Wasserstein Differences-in-Differences

The usual **parallel trends assumption**

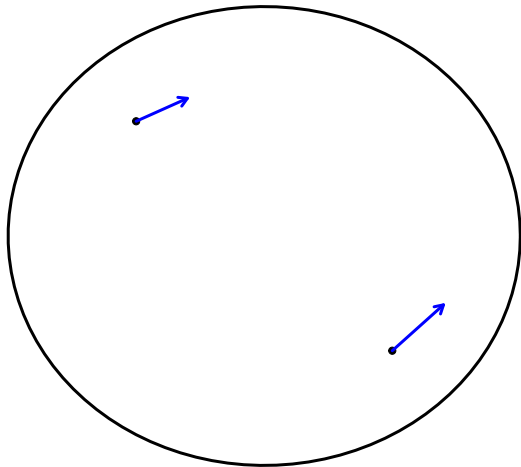


Distributional Version With Continuous Time

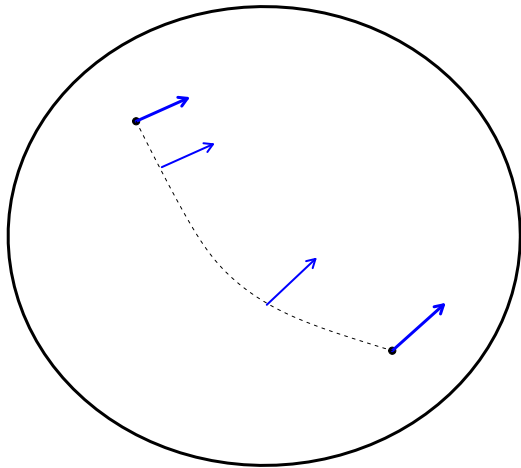


What does it mean for the curves to be parallel?

Parallel Transport



Parallel Transport



Parallel Transport

Parallel Transport

- ▶ The set of distributions endowed with Wasserstein distance W is a manifold.

Parallel Transport

- ▶ The set of distributions endowed with Wasserstein distance W is a manifold.
- ▶ Parallel transport for Wasserstein space was developed by Ambrosio and Gigli (2008), Gigli (2012) and Clancy (2021)

Parallel Transport

- ▶ The set of distributions endowed with Wasserstein distance W is a manifold.
- ▶ Parallel transport for Wasserstein space was developed by Ambrosio and Gigli (2008), Gigli (2012) and Clancy (2021)
- ▶ Given a curve μ_t defined by vector field v_t and another vector u_t , then u_t is parallel if

$$\lim_{h \rightarrow 0} \Pi_{\mu_t} \left(\frac{u_{t+h}(T_{t,t+h}(y)) - u_t(y))}{h} \right) = 0$$

where $T_{t,t+h}$ is the O map from μ_t to μ_{t+h} and Π_{μ_t} is the projection onto the tangent space at μ_t .

Parallel Transport

- ▶ The set of distributions endowed with Wasserstein distance W is a manifold.
- ▶ Parallel transport for Wasserstein space was developed by Ambrosio and Gigli (2008), Gigli (2012) and Clancy (2021)
- ▶ Given a curve μ_t defined by vector field v_t and another vector u_t , then u_t is parallel if

$$\lim_{h \rightarrow 0} \Pi_{\mu_t} \left(\frac{u_{t+h}(T_{t,t+h}(y)) - u_t(y))}{h} \right) = 0$$

where $T_{t,t+h}$ is the O map from μ_t to μ_{t+h} and Π_{μ_t} is the projection onto the tangent space at μ_t .

- ▶ Computing is challenging but doable.

Distributional Version With Continuous Time

Distributional Version With Continuous Time

- ▶ Treated process μ_t with velocity v_t .

Distributional Version With Continuous Time

- ▶ Treated process μ_t with velocity v_t .
- ▶ Counterfactual (untreated) μ_t^* with velocity v_t^* .

Distributional Version With Continuous Time

- ▶ Treated process μ_t with velocity v_t .
- ▶ Counterfactual (untreated) μ_t^* with velocity v_t^* .
- ▶ Untreated (control) process ν_t with velocity u_t .

Distributional Version With Continuous Time

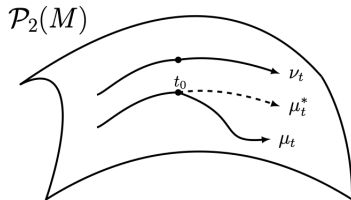
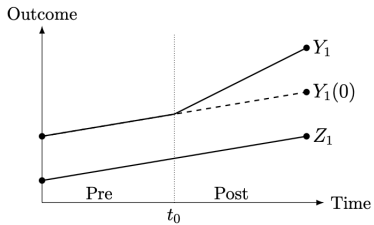
- ▶ Treated process μ_t with velocity v_t .
- ▶ Counterfactual (untreated) μ_t^* with velocity v_t^* .
- ▶ Untreated (control) process ν_t with velocity u_t .
- ▶ Want the counterfactual μ_t^* (if the treated had not been treated).

Distributional Version With Continuous Time

- ▶ Treated process μ_t with velocity v_t .
- ▶ Counterfactual (untreated) μ_t^* with velocity v_t^* .
- ▶ Untreated (control) process ν_t with velocity u_t .
- ▶ Want the counterfactual μ_t^* (if the treated had not been treated).
- ▶ Assumption:

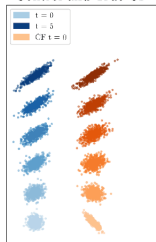
$$v_t^* = \text{PT}_{\nu_t \rightarrow \mu_t^*}(u_t).$$

DiD

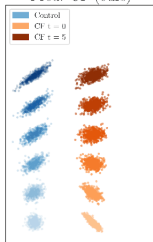


Example 1

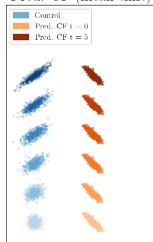
Control and True CF



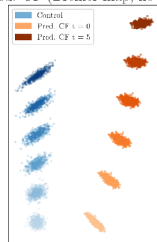
Pred. CF (ours)



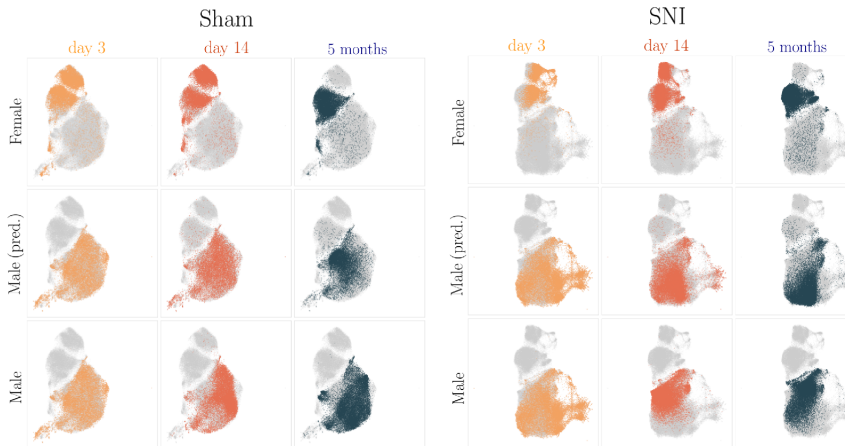
Pred. CF (mean shift)



Pred. CF (Brenier map, no PT)



Example 2



Extension to Time Varying Treatments

Extension to Time Varying Treatments

- ▶ Time varying confounders, treatments and outcome:

$$(X_1, A_1, Y_1), \dots, (X_T, A_T, Y_T)$$

Notation:

$$\overline{A}_t = (A_1, \dots, A_t)$$

etc.

Extension to Time Varying Treatments

- ▶ Time varying confounders, treatments and outcome:

$$(X_1, A_1, Y_1), \dots, (X_T, A_T, Y_T)$$

Notation:

$$\bar{A}_t = (A_1, \dots, A_t)$$

etc.

- ▶ Time varying parallel trends (Renson et al 2022, Shahn et al 2025):

$$E[Y_t(\bar{a}^*) - Y_{t-1}(\bar{a}^*) | \bar{X}_k, \bar{A}_{k-1} = \bar{a}_{k-1}^*] = E[Y_t(\bar{a}^*) - Y_{t-1}(\bar{a}^*) | \bar{X}_k, \bar{A}_k = \bar{a}_k^*]$$

for $k \leq t$. This identifies $E[Y_t(\bar{a}^*)]$ even if there is residual unobserved confounding.

Extension to Time Varying Treatments

- ▶ Time varying confounders, treatments and outcome:

$$(X_1, A_1, Y_1), \dots, (X_T, A_T, Y_T)$$

Notation:

$$\bar{A}_t = (A_1, \dots, A_t)$$

etc.

- ▶ Time varying parallel trends (Renson et al 2022, Shahn et al 2025):

$$E[Y_t(\bar{a}^*) - Y_{t-1}(\bar{a}^*) | \bar{X}_k, \bar{A}_{k-1} = \bar{a}_{k-1}^*] = E[Y_t(\bar{a}^*) - Y_{t-1}(\bar{a}^*) | \bar{X}_k, \bar{A}_k = \bar{a}_k^*]$$

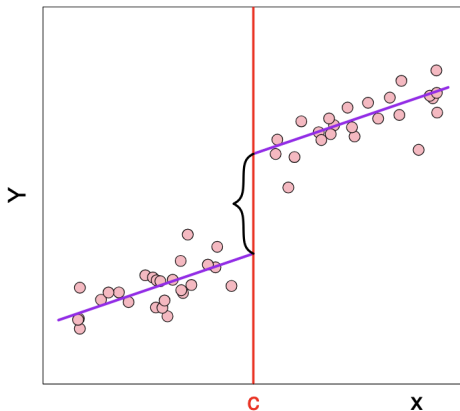
for $k \leq t$. This identifies $E[Y_t(\bar{a}^*)]$ even if there is residual unobserved confounding.

- ▶ Replacing the above assumption with the distributional version:

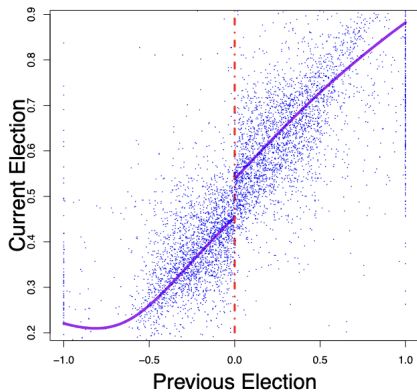
$$\nu_t^*(y | \bar{X}_k, \bar{X}_k, \bar{A}_{k-1} = \bar{a}_{k-1}^*) = \text{PT}_{\nu_t \rightarrow \mu_t^*} u_t(y | \bar{X}_k, \bar{X}_k, \bar{A}_{k-1} = \bar{a}_{k-1}^*)$$

identifies the whole distribution.

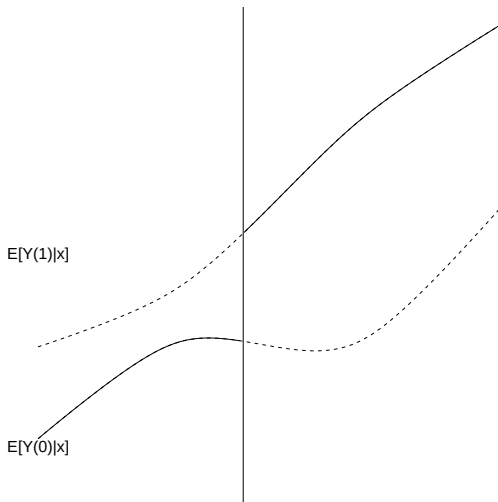
Distributional Discontinuity Design



Discontinuity Design



Discontinuity Design



Discontinuity Design

Discontinuity Design

- ▶ $Z = (A, X, Y)$ and $A = I(X \geq x_0)$

Discontinuity Design

- ▶ $Z = (A, X, Y)$ and $A = I(X \geq x_0)$
- ▶ Let

$$\psi = W(P_{Y(1)|X=x_0}, P_{Y(0)|X=x_0}).$$

Discontinuity Design

► $Z = (A, X, Y)$ and $A = I(X \geq x_0)$

► Let

$$\Psi = W(P_{Y(1)|X=x_0}, P_{Y(0)|X=x_0}).$$

► Let $Q_{Y|X} = F_{Y|X}^{-1}$ and

$$Q_0(u) = \lim_{x \uparrow x_0} Q_{Y|X=x}(u)$$

$$Q_1(u) = \lim_{x \downarrow x_0} Q_{Y|X=x}(u).$$

Discontinuity Design

► $Z = (A, X, Y)$ and $A = I(X \geq x_0)$

► Let

$$\psi = W(P_{Y(1)|X=x_0}, P_{Y(0)|X=x_0}).$$

► Let $Q_{Y|X} = F_{Y|X}^{-1}$ and

$$Q_0(u) = \lim_{x \uparrow x_0} Q_{Y|X=x}(u)$$

$$Q_1(u) = \lim_{x \downarrow x_0} Q_{Y|X=x}(u).$$

► Under mild regularity conditions

$$\psi = \int_0^1 (Q_1(u) - Q_0(u))^2 du.$$

Discontinuity Design

Discontinuity Design

- ▶ We estimate $F_{Y|X}$ using local linear regression with bandwidth h and then

$$\sqrt{nh}(\hat{Q}_1(u) - \hat{Q}_2(u)) \rightsquigarrow \mathbb{G}$$

where $\mathbb{G}$ is a Gaussian process.

Discontinuity Design

- ▶ We estimate $F_{Y|X}$ using local linear regression with bandwidth h and then

$$\sqrt{nh}(\hat{Q}_1(u) - \hat{Q}_2(u)) \rightsquigarrow \mathbb{G}$$

where $\mathbb{G}$ is a Gaussian process.

- ▶ From this we construct tests and confidence intervals.

Example: Election Data

Example: Election Data

- ▶ $\hat{\Psi} = 7.6$ with 95 percent confidence interval $[5.0, 9.4]$.

Example: Election Data

- ▶ $\hat{\Psi} = 7.6$ with 95 percent confidence interval $[5.0, 9.4]$.
- ▶ L -moment decomposition:

Moment	Percent Effect
$k = 1$	0.56
$k = 2$	0.04
$k = 3$	0.11
$k \geq 4$	0.29

Example: Election Data

- ▶ $\hat{\Psi} = 7.6$ with 95 percent confidence interval $[5.0, 9.4]$.
- ▶ L -moment decomposition:

Moment	Percent Effect
$k = 1$	0.56
$k = 2$	0.04
$k = 3$	0.11
$k \geq 4$	0.29

- ▶ Mean only accounts for 56 percent of the difference

Example: Election Data

- ▶ $\hat{\Psi} = 7.6$ with 95 percent confidence interval $[5.0, 9.4]$.
- ▶ L -moment decomposition:

Moment	Percent Effect
$k = 1$	0.56
$k = 2$	0.04
$k = 3$	0.11
$k \geq 4$	0.29

- ▶ Mean only accounts for 56 percent of the difference
- ▶ Extension to kinked design.

Bounds

Bounds

- ▶ Suppose $A \in \{0, 1\}$. The counterfactuals are $(Y(0), Y(1))$.

Bounds

- ▶ Suppose $A \in \{0, 1\}$. The counterfactuals are $(Y(0), Y(1))$.
- ▶ The law of $Y(0)$ and the law of $Y(1)$ are identified but the joint law $(Y(0), Y(1))$ is not.

Bounds

- ▶ Suppose $A \in \{0, 1\}$. The counterfactuals are $(Y(0), Y(1))$.
- ▶ The law of $Y(0)$ and the law of $Y(1)$ are identified but the joint law $(Y(0), Y(1))$ is not.
- ▶ For example:

$$\mathbb{E}[Y(0)|X, Y, A = 1]$$

is not identified.

Bounds

- ▶ Suppose $A \in \{0, 1\}$. The counterfactuals are $(Y(0), Y(1))$.
- ▶ The law of $Y(0)$ and the law of $Y(1)$ are identified but the joint law $(Y(0), Y(1))$ is not.
- ▶ For example:

$$\mathbb{E}[Y(0)|X, Y, A = 1]$$

is not identified.

- ▶ And

$$\mathbb{E}[(Y(1) - Y(0))^2]$$

is not identified.

- ▶ Common strategy: put bounds on such quantities.

Bounds

Bounds

- Frechet-Hoeffding bounds on the joint cdf $F(y(0), y(1))$:

$$L(y(0), y(1)) \leq F(y(0), y(1)) \leq U(y(0), y(1))$$

where

$$L(y(0), y(1)) = \max\{F_0(y(0)) + F_1(y(1)) - 1, 0\}$$

$$U(y(0), y(1)) = \min\{F_0(y(0)), F_1(y(1))\}.$$

Bounds

- Frechet-Hoeffding bounds on the joint cdf $F(y(0), y(1))$:

$$L(y(0), y(1)) \leq F(y(0), y(1)) \leq U(y(0), y(1))$$

where

$$L(y(0), y(1)) = \max\left\{F_0(y(0)) + F_1(y(1)) - 1, 0\right\}$$

$$U(y(0), y(1)) = \min\left\{F_0(y(0)), F_1(y(1))\right\}.$$

- Let $\psi = \mathbb{E}[(Y(1) - Y(0))^2]$. The lower bound is obtained from the optimal transport map T taking $P_0 \rightarrow P_1$:

$$\psi \geq \mathbb{E}[(T(Y(0)) - Y(0))^2].$$

Bounds

- ▶ Frechet-Hoeffding bounds on the joint cdf $F(y(0), y(1))$:

$$L(y(0), y(1)) \leq F(y(0), y(1)) \leq U(y(0), y(1))$$

where

$$L(y(0), y(1)) = \max\{F_0(y(0)) + F_1(y(1)) - 1, 0\}$$

$$U(y(0), y(1)) = \min\{F_0(y(0)), F_1(y(1))\}.$$

- ▶ Let $\psi = \mathbb{E}[(Y(1) - Y(0))^2]$. The lower bound is obtained from the optimal transport map T taking $P_0 \rightarrow P_1$:

$$\psi \geq \mathbb{E}[(T(Y(0)) - Y(0))^2].$$

- ▶ This corresponds to the joint distribution J defined by

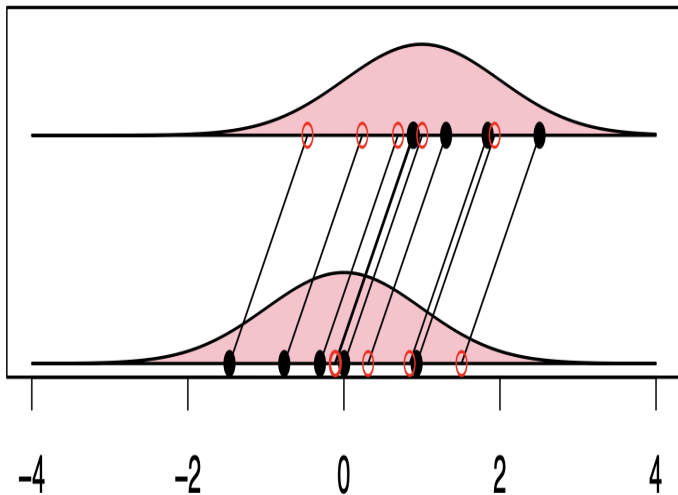
$$(Y(0), Y(1)) = (Y(0), T(Y(0))).$$



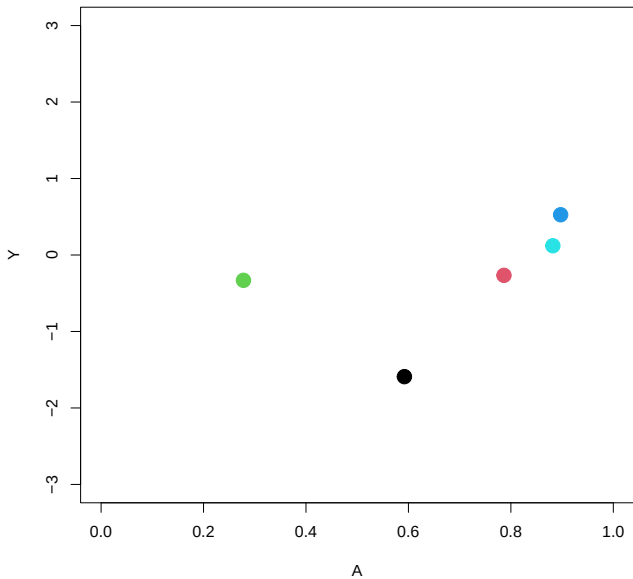
$$E_J[Y(1)|X, A = 0, Y] = T(Y(0))$$

$$E_J[Y(0)|X, A = 1, Y] = T^{-1}(Y(1)).$$

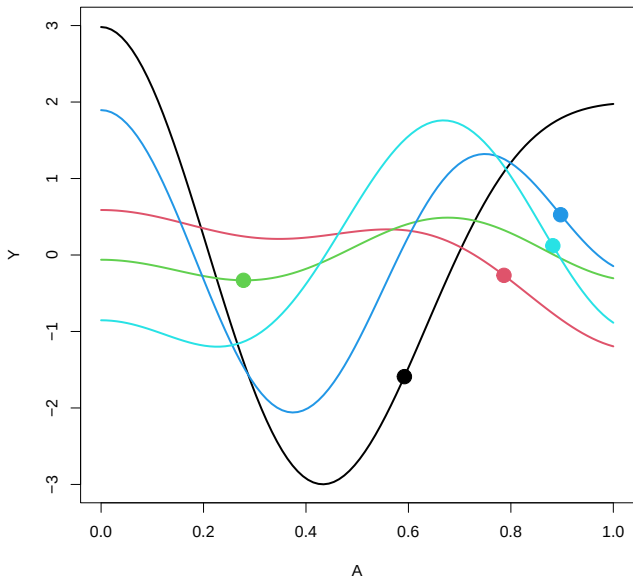
Minimal/Conservative Effect



Extension to Continuous A



Extension to Continuous A



Extension to Continuous A

Extension to Continuous A

- ▶ The marginal P_a is identified for each a :

$$P_a(\cdot) = \int P(\cdot | X = x, A = a) dP_X(x).$$

Extension to Continuous A

- ▶ The marginal P_a is identified for each a :

$$P_a(\cdot) = \int P(\cdot | X = x, A = a) dP_X(x).$$

- ▶ But the joint distribution of $(Y(a) : 0 \leq a \leq 1)$ is not identified.

Extension to Continuous A

- ▶ The marginal P_a is identified for each a :

$$P_a(\cdot) = \int P(\cdot | X = x, A = a) dP_X(x).$$

- ▶ But the joint distribution of $(Y(a) : 0 \leq a \leq 1)$ is not identified.
- ▶ Any function of $(Y(a) : 0 \leq a \leq 1)$ is not identified. For example

$$\mathbb{E}[Y(a)|X, A, Y].$$

Extension to Continuous A

- ▶ The marginal P_a is identified for each a :

$$P_a(\cdot) = \int P(\cdot | X = x, A = a) dP_X(x).$$

- ▶ But the joint distribution of $(Y(a) : 0 \leq a \leq 1)$ is not identified.
- ▶ Any function of $(Y(a) : 0 \leq a \leq 1)$ is not identified. For example

$$\mathbb{E}[Y(a) | X, A, Y].$$

- ▶ We want to find the minimal causal effect. Which joint distribution makes $Y(a)$ as “flat” as possible?

Extension to Continuous A

Extension to Continuous A

► Define

$$\psi = \mathbb{E} \left[\int \int (Y(a_1) - Y(a_2))^2 dP_A(a_1) dP_A(a_2) \right].$$

Extension to Continuous A

- Define

$$\psi = \mathbb{E} \left[\int \int (Y(a_1) - Y(a_2))^2 dP_A(a_1) dP_A(a_2) \right].$$

- Let $\mathcal{J}$ denote all distributions on paths $(Y(a) : 0 \leq 1)$ such that the marginal distribution J_a for $Y(a)$ satisfies

$$J_a = P_a \quad \text{for all } a.$$

(Recall $P_a = \int P(\cdot | X = x, A = a) dP_X(x)$.)

Extension to Continuous A

- Define

$$\psi = \mathbb{E} \left[\int \int (Y(a_1) - Y(a_2))^2 dP_A(a_1) dP_A(a_2) \right].$$

- Let $\mathcal{J}$ denote all distributions on paths $(Y(a) : 0 \leq 1)$ such that the marginal distribution J_a for $Y(a)$ satisfies

$$J_a = P_a \quad \text{for all } a.$$

(Recall $P_a = \int P(\cdot | X = x, A = a) dP_X(x)$.)

- The solution was found in Pass (2013).

Continuous A

Continuous A

- Let B be the **barycenter** which minimizes

$$\int W^2(P_a, B) dP_A(a).$$

Continuous A

- ▶ Let B be the **barycenter** which minimizes

$$\int W^2(P_a, B) dP_A(a).$$

- ▶ The minimizer J_* is given by

$$Y(a) = T_a(Z)$$

where $Z \sim B$ and T_a is the OT map from B to P_a .

Continuous A

- ▶ Let B be the **barycenter** which minimizes

$$\int W^2(P_a, B) dP_A(a).$$

- ▶ The minimizer J_* is given by

$$Y(a) = T_a(Z)$$

where $Z \sim B$ and T_a is the OT map from B to P_a .

- ▶ Under J_* , the law of $Y(a)$ given (X, A, Y) is deterministic and is given by

$$Y_*(a) = F_a^{-1}(F_A(Y)).$$

Implementation

- Perform (causal) quantile estimation:

$$\hat{Q}(a, \tau) = \hat{F}_{Y(a)}^{-1}(\tau)$$

where

$$F_{Y(a)}(y) = \int F(y|x, a) dP_X(x).$$

Implementation

- Perform (causal) quantile estimation:

$$\hat{Q}(a, \tau) = \hat{F}_{Y(a)}^{-1}(\tau)$$

where

$$F_{Y(a)}(y) = \int F(y|x, a) dP_X(x).$$

- Let $\tau_i = \hat{F}_{Y(A_i)}(Y_i)$.

Implementation

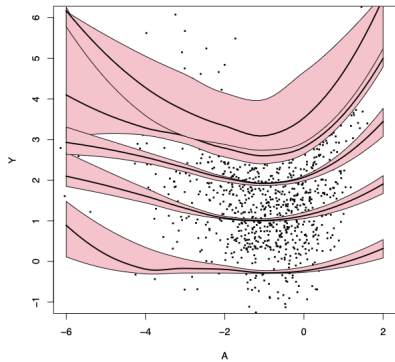
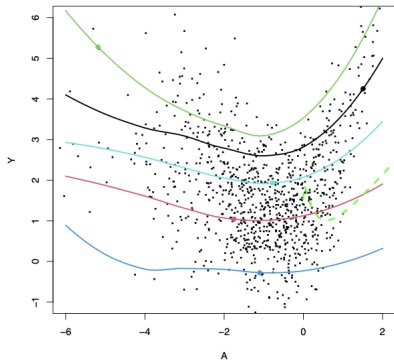
- Perform (causal) quantile estimation:

$$\hat{Q}(a, \tau) = \hat{F}_{Y(a)}^{-1}(\tau)$$

where

$$F_{Y(a)}(y) = \int F(y|x, a) dP_X(x).$$

- Let $\tau_i = \hat{F}_{Y(A_i)}(Y_i)$.
- Find $\hat{Q}(a, \tau_i)$.



Causal Geodesy

Causal Geodesy

- ▶ Usual interventions in causal inference: set $A = a$.

Causal Geodesy

- ▶ Usual interventions in causal inference: set $A = a$.
- ▶ Stochastic interventions: draw $A \sim Q(a|x)$.

Causal Geodesy

- ▶ Usual interventions in causal inference: **set** $A = a$.
- ▶ Stochastic interventions: draw $A \sim Q(a|x)$.
- ▶ Incremental interventions (Kennedy 2019):

$$q_{\delta}(A = 1|X = x) = \frac{\delta P(A = 1|X = x)}{\delta P(A = 1|X = x) + P(A = 0|X = x)}.$$

Causal Geodesy

- ▶ Usual interventions in causal inference: **set** $A = a$.
- ▶ Stochastic interventions: draw $A \sim Q(a|x)$.
- ▶ Incremental interventions (Kennedy 2019):

$$q_{\delta}(A = 1|X = x) = \frac{\delta P(A = 1|X = x)}{\delta P(A = 1|X = x) + P(A = 0|X = x)}.$$

- ▶ δ is the change in odds of treatment:

$$\delta = \frac{\text{odds}_q(A = 1|X = x)}{\text{odds}_{\pi}(A = 1|X = x)}.$$

Causal Geodesy

- ▶ Usual interventions in causal inference: set $A = a$.
- ▶ Stochastic interventions: draw $A \sim Q(a|x)$.
- ▶ Incremental interventions (Kennedy 2019):

$$q_{\delta}(A = 1|X = x) = \frac{\delta P(A = 1|X = x)}{\delta P(A = 1|X = x) + P(A = 0|X = x)}.$$

- ▶ δ is the change in odds of treatment:

$$\delta = \frac{\text{odds}_q(A = 1|X = x)}{\text{odds}_{\pi}(A = 1|X = x)}.$$

- ▶ $\delta \rightarrow \infty$ gives $A = 1$ and $\delta \rightarrow 0$ gives $A = 0$.

Causal Geodesy

- ▶ Usual interventions in causal inference: set $A = a$.
- ▶ Stochastic interventions: draw $A \sim Q(a|x)$.
- ▶ Incremental interventions (Kennedy 2019):

$$q_{\delta}(A = 1|X = x) = \frac{\delta P(A = 1|X = x)}{\delta P(A = 1|X = x) + P(A = 0|X = x)}.$$

- ▶ δ is the change in odds of treatment:

$$\delta = \frac{\text{odds}_q(A = 1|X = x)}{\text{odds}_{\pi}(A = 1|X = x)}.$$

- ▶ $\delta \rightarrow \infty$ gives $A = 1$ and $\delta \rightarrow 0$ gives $A = 0$.
- ▶ Interpolates between “no intervention” to “sharp intervention”

Incremental

Incremental

- ▶ Schindl, Shen and Kennedy (2024) considered the exponential tilt

$$q_{\delta}(a|x) \propto e^{\delta a} \pi(a|x)$$

where $\pi(a|x)$ is the density of A given X (propensity score).

Incremental

- Schindl, Shen and Kennedy (2024) considered the exponential tilt

$$q_{\delta}(a|x) \propto e^{\delta a} \pi(a|x)$$

where $\pi(a|x)$ is the density of A given X (propensity score).

- We consider a family of stochastic interventions that vary from

$$\underbrace{\pi(a|x)}_{\text{no intervention}} \longrightarrow \underbrace{\delta_{a_0}}_{\text{sharp intervention}}$$

Incremental

- Schindl, Shen and Kennedy (2024) considered the exponential tilt

$$q_{\delta}(a|x) \propto e^{\delta a} \pi(a|x)$$

where $\pi(a|x)$ is the density of A given X (propensity score).

- We consider a family of stochastic interventions that vary from

$$\underbrace{\pi(a|x)}_{\text{no intervention}} \longrightarrow \underbrace{\delta_{a_0}}_{\text{sharp intervention}}$$

Correlation $\longrightarrow$ Causation

Incremental

- ▶ Schindl, Shen and Kennedy (2024) considered the exponential tilt

$$q_{\delta}(a|x) \propto e^{\delta a} \pi(a|x)$$

where $\pi(a|x)$ is the density of A given X (propensity score).

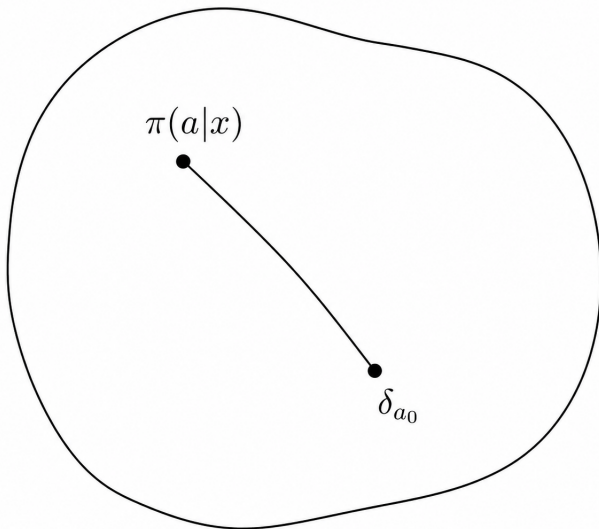
- ▶ We consider a family of stochastic interventions that vary from

$$\underbrace{\pi(a|x)}_{\text{no intervention}} \longrightarrow \underbrace{\delta_{a_0}}_{\text{sharp intervention}}$$

$$\text{Correlation} \longrightarrow \text{Causation}$$

- ▶ There are many such paths. We consider the shortest path: the Wasserstein geodesic.

Geodesic



Geodesics

Geodesics

- ▶ Given P and Q there is a geodesic (shortest path) $(\nu_t : 0 \leq t \leq 1)$ where

$$\nu_0 = P, \quad \nu_1 = Q.$$

Geodesics

- ▶ Given P and Q there is a geodesic (shortest path) $(\nu_t : 0 \leq t \leq 1)$ where

$$\nu_0 = P, \quad \nu_1 = Q.$$

- ▶ ν_t is the distribution of X_t where

$$X_t = (1 - t)X + tT(X)$$

where $X \sim P$ and T is the OT map from P to Q .

In other words, $\nu_t = ((1 - t)id + tT)_\# P$.

Geodesy

Geodesy

- Now let ν_t be the geodesic connecting $\pi(a|x)$ to δ_{a_*} .

Geodesy

- ▶ Now let ν_t be the geodesic connecting $\pi(a|x)$ to δ_{a_*} .
- ▶ Then

$$\nu_t(a|x) = \frac{1}{1-t} \pi \left(\frac{a - ta_*}{1-t} \middle| x \right)$$

and this corresponds to the following stochastic intervention:

$$(1-t)A + ta_*, \quad A \sim \pi(a|x).$$

Geodesy

- ▶ Now let ν_t be the geodesic connecting $\pi(a|x)$ to δ_{a_*} .
- ▶ Then

$$\nu_t(a|x) = \frac{1}{1-t} \pi \left(\frac{a - ta_*}{1-t} \middle| x \right)$$

and this corresponds to the following stochastic intervention:

$$(1-t)A + ta_*, \quad A \sim \pi(a|x).$$

- ▶ $t = 0$, parametric rate, no confounding, no positivity restrictions
→
 $t = 1$, nonparametric rate, adjust for confounding, positivity restrictions

Geodesy

- ▶ Now let ν_t be the geodesic connecting $\pi(a|x)$ to δ_{a_*} .
- ▶ Then

$$\nu_t(a|x) = \frac{1}{1-t} \pi \left(\frac{a - ta_*}{1-t} \middle| x \right)$$

and this corresponds to the following stochastic intervention:

$$(1-t)A + ta_*, \quad A \sim \pi(a|x).$$

- ▶ $t = 0$, parametric rate, no confounding, no positivity restrictions
→
 $t = 1$, nonparametric rate, adjust for confounding, positivity restrictions
- ▶ Estimation gets harder as $t \rightarrow 1$.

Geodesy

Geodesy

- ▶ Estimating gets harder with increasing t .

Geodesy

- ▶ Estimating gets harder with increasing t .
- ▶ In fact:

$$\inf_{\hat{\psi}} \sup_P \mathbb{E} \int_0^t (\hat{\psi}_s - \psi_s)^2 \geq \frac{C}{n} \int_0^t (1 + \chi^2(\rho_t, \pi)) ds$$

Geodesy

- ▶ Estimating gets harder with increasing t .
- ▶ In fact:

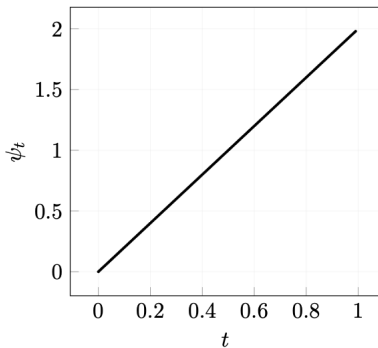
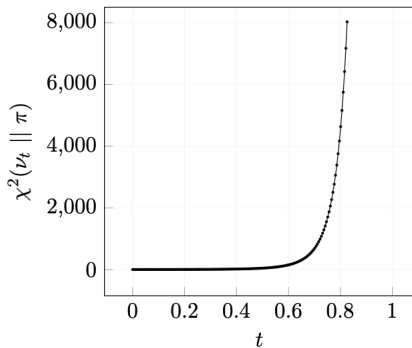
$$\inf_{\hat{\psi}} \sup_P \mathbb{E} \int_0^t (\hat{\psi}_s - \psi_s)^2 \geq \frac{C}{n} \int_0^t (1 + \chi^2(\rho_t, \pi)) ds$$

- ▶ For each $0 \leq t < 1$, using a one-step estimator,

$$\frac{\sqrt{n}}{\sigma(t)} (\hat{\psi}(t) - \psi(t)) \rightsquigarrow N(0, 1)$$

but this is not uniform in t as $t \rightarrow 1$.

Geodesy: Marginal Structural Model



Geodesy: Marginal Structural Model

Geodesy: Marginal Structural Model

- ▶ Model $m(a_*, t, \beta)$ for $\psi(a_*, t)$.

Geodesy: Marginal Structural Model

- ▶ Model $m(a_*, t, \beta)$ for $\psi(a_*, t)$.
- ▶ For example:

$$m(a_*, t, \beta) = \mu + t \sum_{j=1}^k \beta_j \phi_j(a).$$

Geodesy: Marginal Structural Model

- ▶ Model $m(a_*, t, \beta)$ for $\psi(a_*, t)$.
- ▶ For example:

$$m(a_*, t, \beta) = \mu + t \sum_{j=1}^k \beta_j \phi_j(a).$$

- ▶ Estimator: choose $0 < t < 1$

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \int_0^1 \int_0^t (\hat{\psi}(a, s) - m(a, s, \beta))^2 ds \, h(a) da$$

where $\hat{\psi}(a, s)$ is the nonparametric estimator.

Geodesy: Marginal Structural Model

- ▶ Model $m(a_*, t, \beta)$ for $\psi(a_*, t)$.
- ▶ For example:

$$m(a_*, t, \beta) = \mu + t \sum_{j=1}^k \beta_j \phi_j(a).$$

- ▶ Estimator: choose $0 < t < 1$

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \int_0^1 \int_0^t (\hat{\psi}(a, s) - m(a, s, \beta))^2 ds \, h(a) da$$

where $\hat{\psi}(a, s)$ is the nonparametric estimator.

- ▶ Then $\operatorname{Var}(\hat{\beta})$ is minimized for some $0 < t_0 < 1$.

Others

Others

- ▶ Mediation: Machado, Voitsitska, Charpentier and Gallic (2026).

Others

- ▶ Mediation: Machado, Voitsitska, Charpentier and Gallic (2026).
- ▶ Sinkhorn causal effects: Agarwal and Luedtke (2026). (Regularized Wasserstein distance.)

Others

- ▶ Mediation: Machado, Voitsitska, Charpentier and Gallic (2026).
- ▶ Sinkhorn causal effects: Agarwal and Luedtke (2026). (Regularized Wasserstein distance.)
- ▶ Sensitivity analysis for stochastic interventions: Levis, Kennedy, McClean, Balakrishnan and Wasserman (2024) and Liang (2026).

Others

- ▶ Mediation: Machado, Voitsitska, Charpentier and Gallic (2026).
- ▶ Sinkhorn causal effects: Agarwal and Luedtke (2026). (Regularized Wasserstein distance.)
- ▶ Sensitivity analysis for stochastic interventions: Levis, Kennedy, McClean, Balakrishnan and Wasserman (2024) and Liang (2026).
- ▶ Causal graphs Tu et al (2022) and Cheridito and Echstein (2025).

Others

- ▶ Mediation: Machado, Voitsitska, Charpentier and Gallic (2026).
- ▶ Sinkhorn causal effects: Agarwal and Luedtke (2026). (Regularized Wasserstein distance.)
- ▶ Sensitivity analysis for stochastic interventions: Levis, Kennedy, McClean, Balakrishnan and Wasserman (2024) and Liang (2026).
- ▶ Causal graphs Tu et al (2022) and Cheridito and Echstein (2025).
- ▶ and more ...

Part IV: Limitations

Limitations

- ▶ Why optimal transport?

Limitations

- ▶ Why optimal transport?
- ▶ The multivariate case.

$$Y_0 \sim \mu_0, \quad Y_1 \sim \mu_1, \quad Y_0, Y_1 \in \mathbb{R}^d.$$

Computation hard is hard. Inference is hard.

Limitations

- ▶ Why optimal transport?
- ▶ The multivariate case.

$$Y_0 \sim \mu_0, \quad Y_1 \sim \mu_1, \quad Y_0, Y_1 \in \mathbb{R}^d.$$

Computation hard is hard. Inference is hard.

- ▶ Often replace OT with entropic transport:

$$\inf_{\pi} \int \int \|y_1 - y_0\|^2 d\pi(y_0, y_1) + \epsilon KL(\pi | \mu_0 \times \mu_1)$$

and take

$$T_{\epsilon}(y_0) = \mathbb{E}_{\pi_{\epsilon}}[Y_1 | Y_0 = y_0].$$

Limitations

- ▶ Why optimal transport?
- ▶ The multivariate case.

$$Y_0 \sim \mu_0, \quad Y_1 \sim \mu_1, \quad Y_0, Y_1 \in \mathbb{R}^d.$$

Computation hard is hard. Inference is hard.

- ▶ Often replace OT with entropic transport:

$$\inf_{\pi} \int \int \|y_1 - y_0\|^2 d\pi(y_0, y_1) + \epsilon KL(\pi | \mu_0 \times \mu_1)$$

and take

$$T_{\epsilon}(y_0) = \mathbb{E}_{\pi_{\epsilon}}[Y_1 | Y_0 = y_0].$$

- ▶ Computation and inference are easier. How do we choose ϵ ?

Limitations

- ▶ Why optimal transport?
- ▶ The multivariate case.

$$Y_0 \sim \mu_0, \quad Y_1 \sim \mu_1, \quad Y_0, Y_1 \in \mathbb{R}^d.$$

Computation hard is hard. Inference is hard.

- ▶ Often replace OT with entropic transport:

$$\inf_{\pi} \int \int \|y_1 - y_0\|^2 d\pi(y_0, y_1) + \epsilon KL(\pi | \mu_0 \times \mu_1)$$

and take

$$T_{\epsilon}(y_0) = \mathbb{E}_{\pi_{\epsilon}}[Y_1 | Y_0 = y_0].$$

- ▶ Computation and inference are easier. How do we choose ϵ ?
- ▶ Usually need to transport $p(y_0|x)$ to $p(y_1|x)$ so we need to do this for each x .

Conclusion

Conclusion

- ▶ Optimal transport/Wasserstein distance has potential for causal inference.

Conclusion

- ▶ Optimal transport/Wasserstein distance has potential for causal inference.
- ▶ Useful for distributional effects.

Conclusion

- ▶ Optimal transport/Wasserstein distance has potential for causal inference.
- ▶ Useful for distributional effects.
- ▶ Has been used before, implicitly.

Conclusion

- ▶ Optimal transport/Wasserstein distance has potential for causal inference.
- ▶ Useful for distributional effects.
- ▶ Has been used before, implicitly.
- ▶ Interest is growing rapidly.

Conclusion

- ▶ Optimal transport/Wasserstein distance has potential for causal inference.
- ▶ Useful for distributional effects.
- ▶ Has been used before, implicitly.
- ▶ Interest is growing rapidly.
- ▶ Computation still a challenge.

Conclusion

- ▶ Optimal transport/Wasserstein distance has potential for causal inference.
- ▶ Useful for distributional effects.
- ▶ Has been used before, implicitly.
- ▶ Interest is growing rapidly.
- ▶ Computation still a challenge.

THE END