Applications of Optimal Transport in Causal Inference

Florian Gunsilius

Kantorovich Initiative Seminar

UNIVERSITY OF MICHIGAN

Goal of this talk

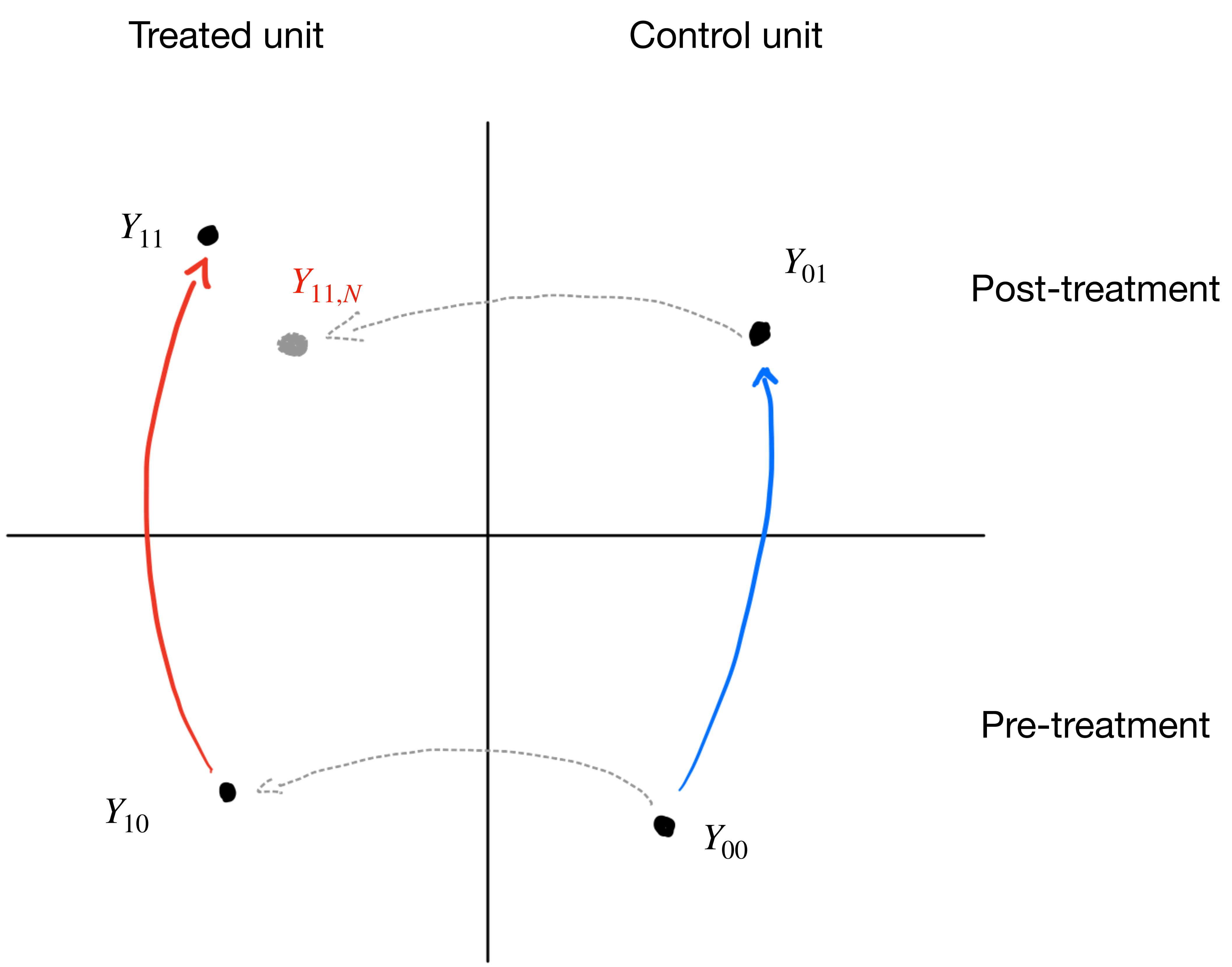
Very narrow overview of approaches in causal inference and econometrics where optimal transport can be useful:

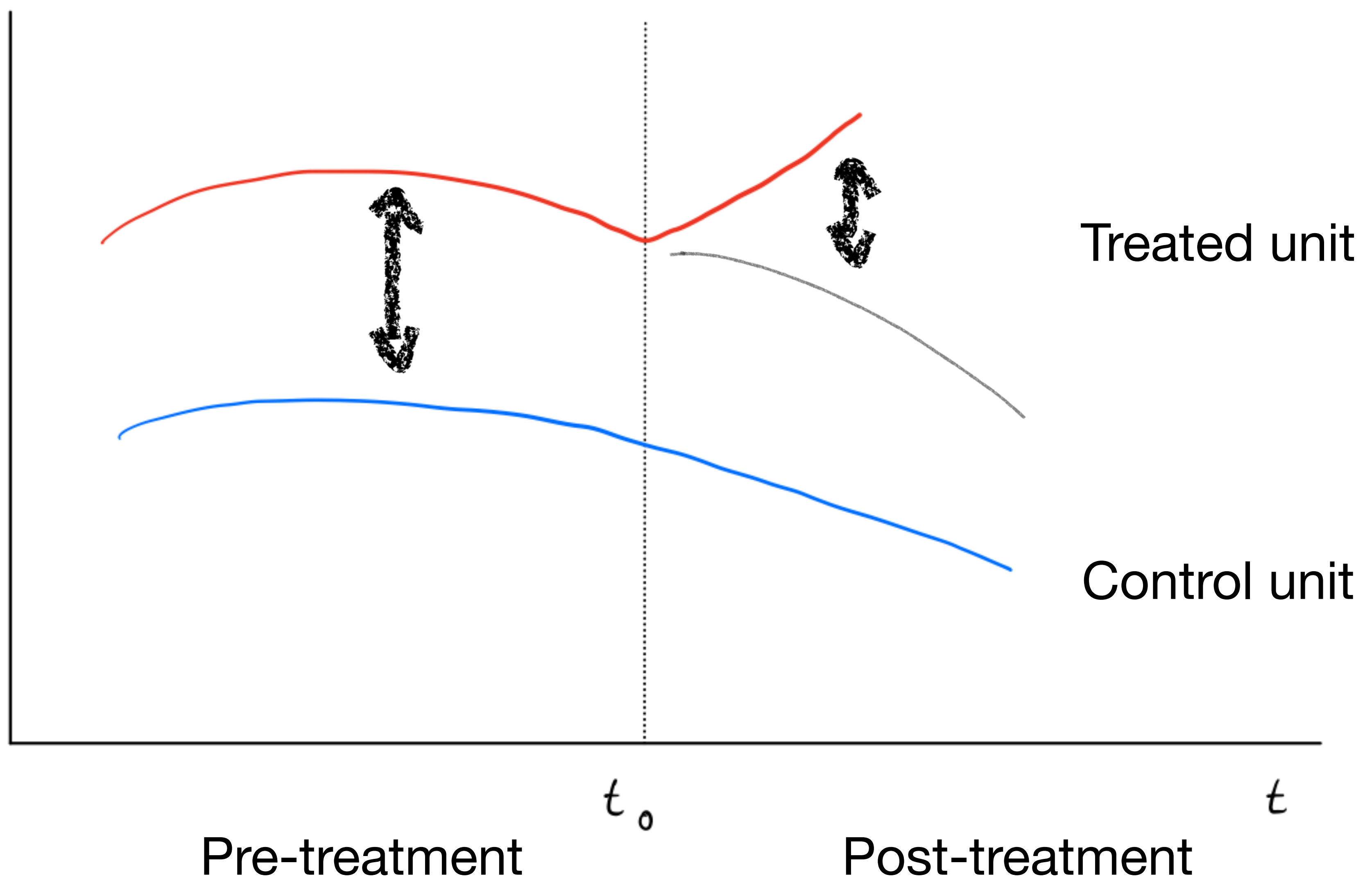
1. Difference-in-differences
2. Synthetic Controls
3. Matching

Many more: instrumental variables, domain adaptation, fairness, ...

➡ OT can be very useful when dealing with **treatment heterogeneity**

1. Optimal transport and difference-in-differences





joint work with
Philippe Rigollet and William Torous

Treatment effect via difference-in-differences

Key: account for the underlying trend of the outcome of the treated group

→ Subtract the trend of the control unit, then any difference between treated- and control unit will be due to the causal effect of the treatment (Abadie 2005, Rev. Econ. Stud., Heckman et.al. 1997, Rev. Econ. Stud.)

$$\boxed{E\left[Y_{11} - Y_{10} \,|\, T = 1\right] = \Big(E\left[Y(1) \,|\, D = 1\right] - E\left[Y(1) \,|\, D = 0\right]\Big) - \Big(E\left[Y(0) \,|\, D = 1\right] - E\left[Y(0) \,|\, D = 0\right]\Big)}$$

ATT

Change in observed outcomes of treated unit

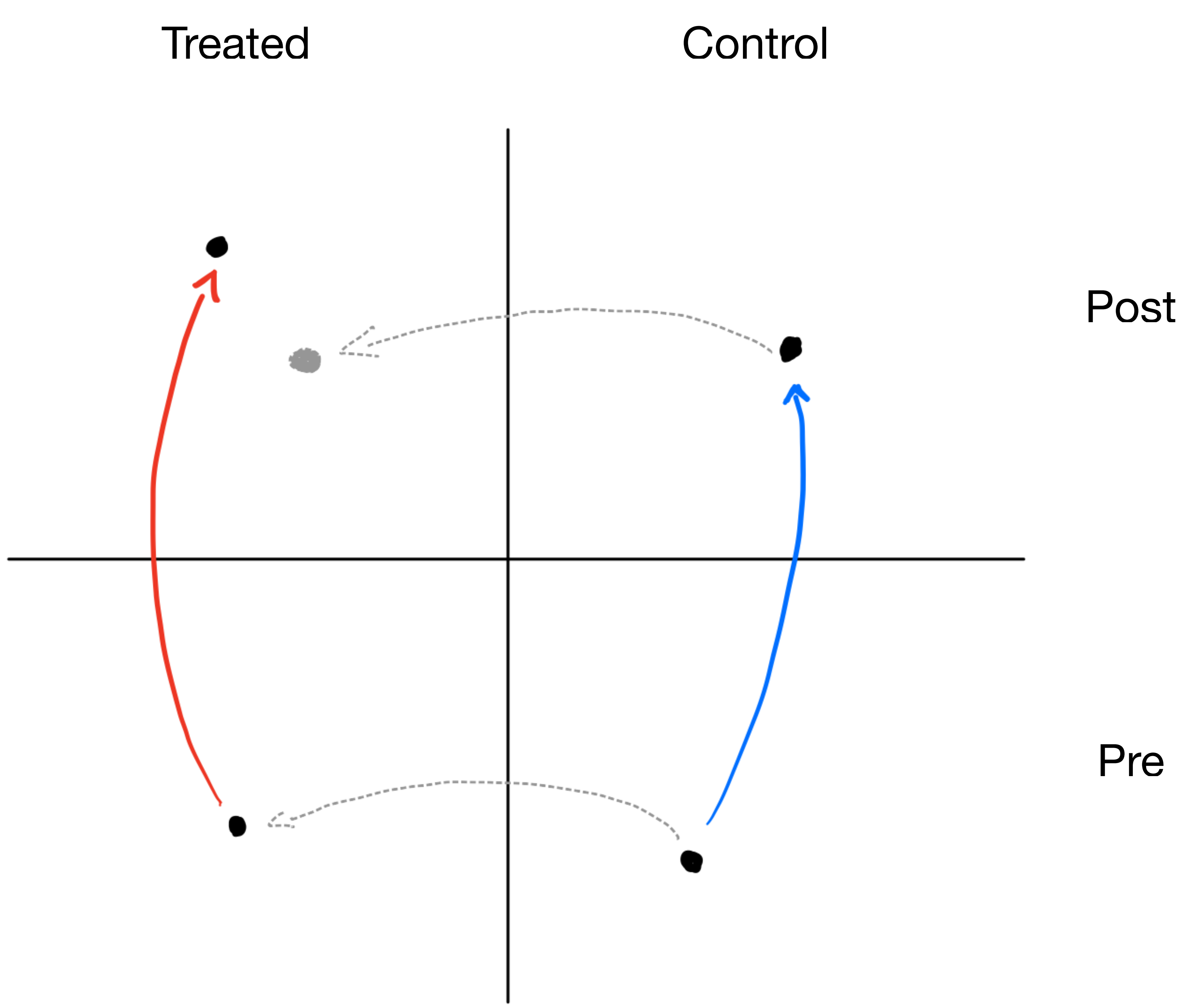
Change in observed outcomes of control unit

Main Assumption: Parallel trends, i.e. treated group would have followed the same trend as control group without treatment

The changes-in-changes estimator and OT

Classical difference in differences is for aggregate outcomes
In many setting one cares about **individual heterogeneity,** captures by **probability distributions**

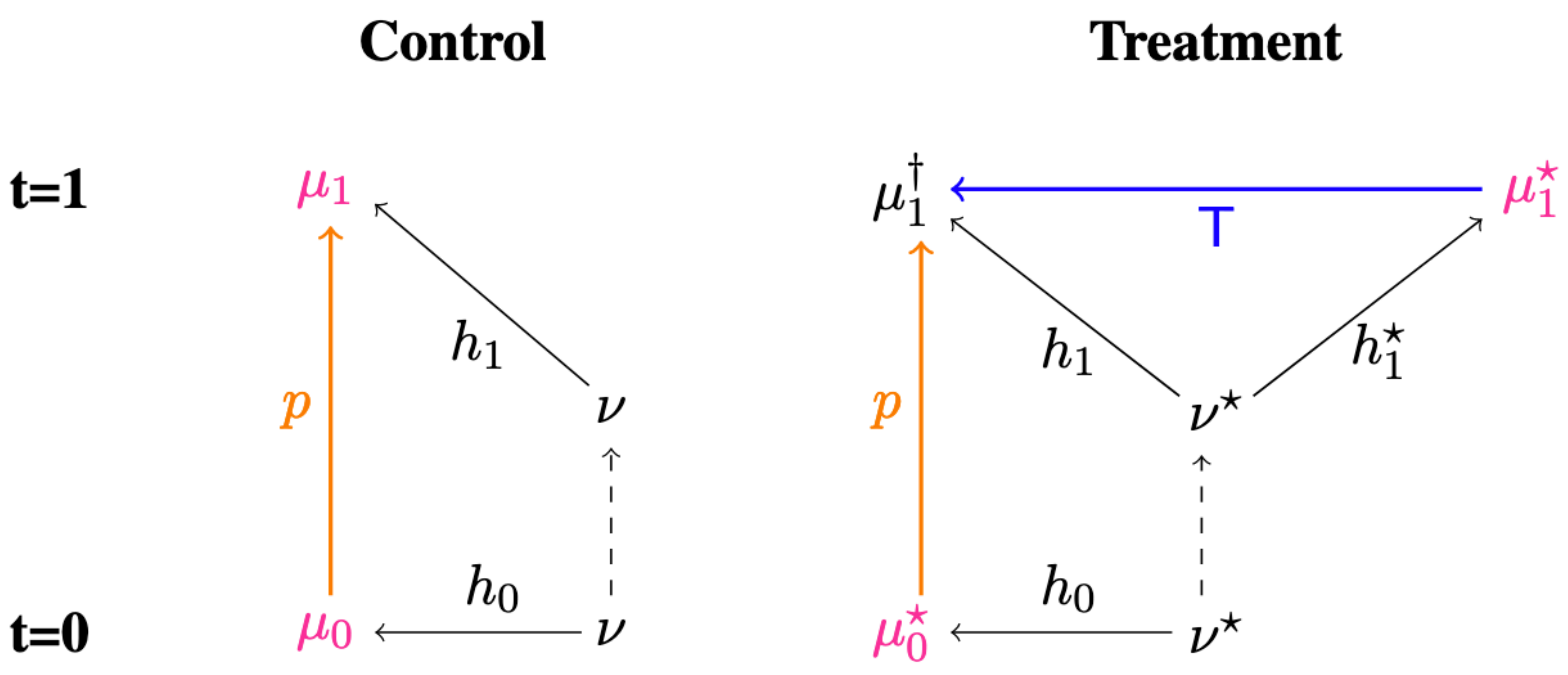
Athey & Imbens (2006) introduce the **changes in changes estimator:**



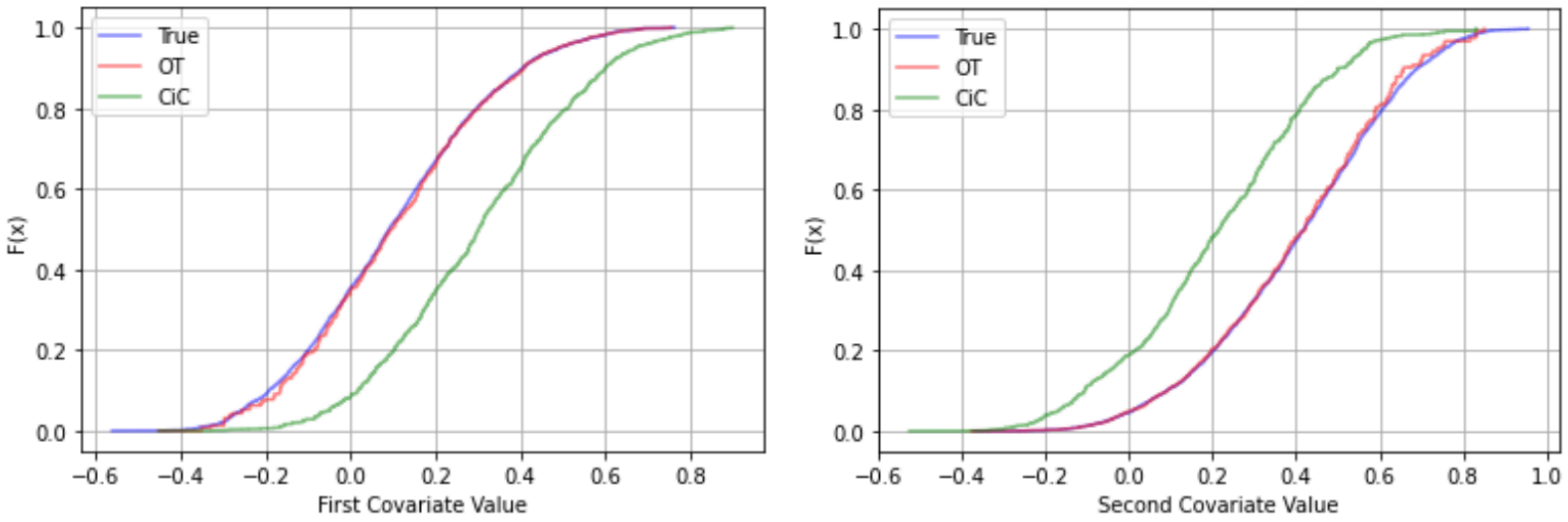
$$\boxed{F_{Y_{11}^N}(y) = F_{Y_{10}}\left(F^{-1}_{Y_{00}}\left(F_{Y_{01}}(y)\right)\right)}$$

Monotone rearrangement

Multivariate extension



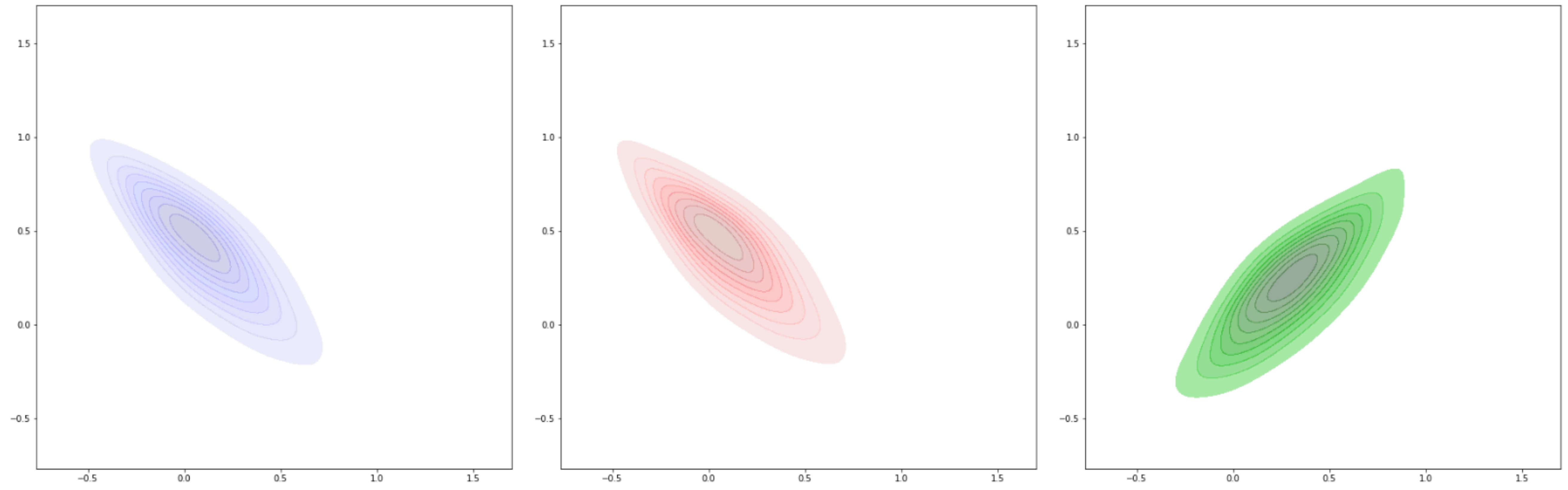
The main identifying assumption now is not monotonicity, but **cyclic monotonicity**



(a) First Marginal

(b) Second Marginal

Figure 2: Recovery of counterfactual marginals by OT and CiC.



(a) True CF

(b) OT Estimate

(c) CiC Estimate

2. Optimal transport for synthetic controls

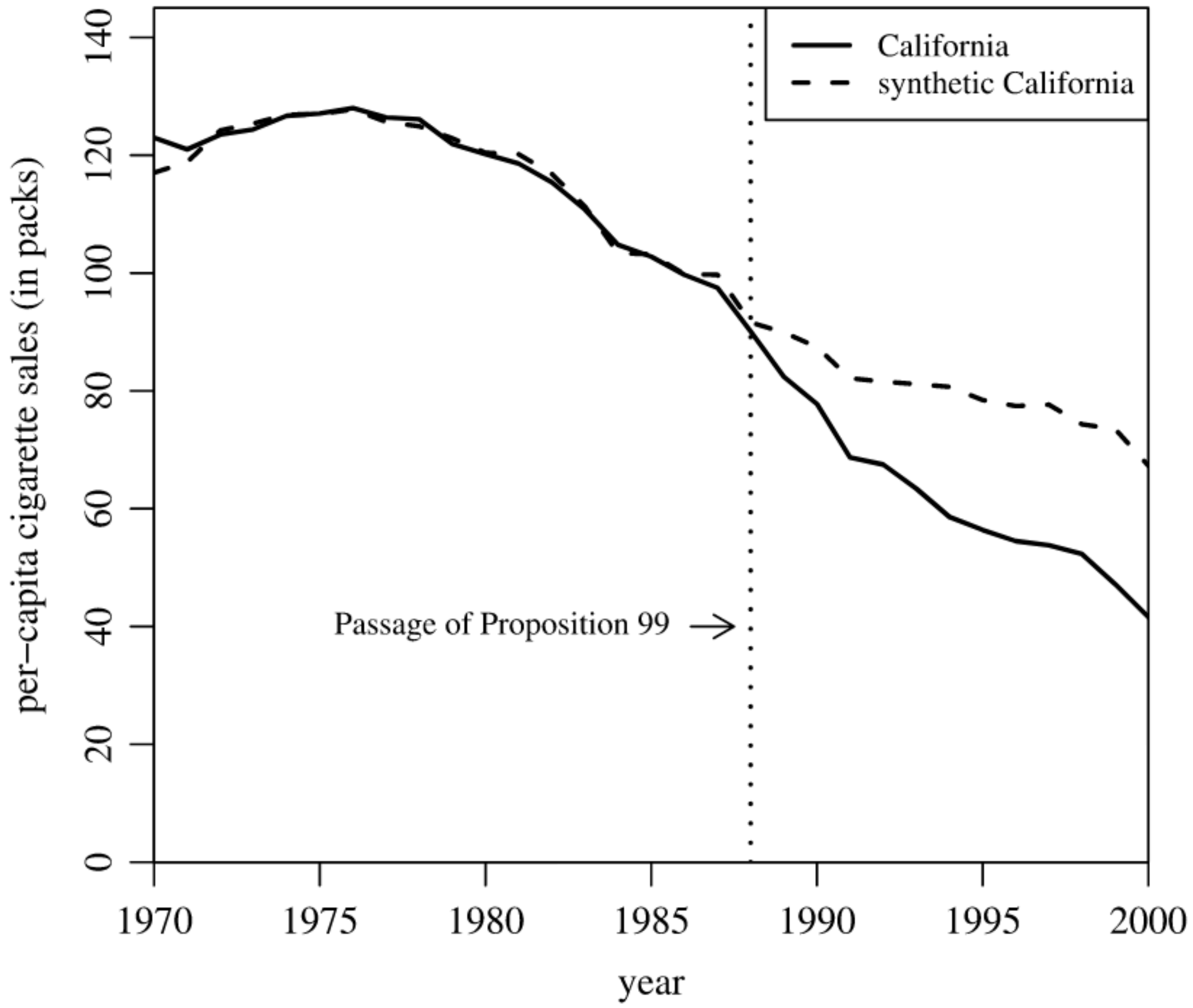
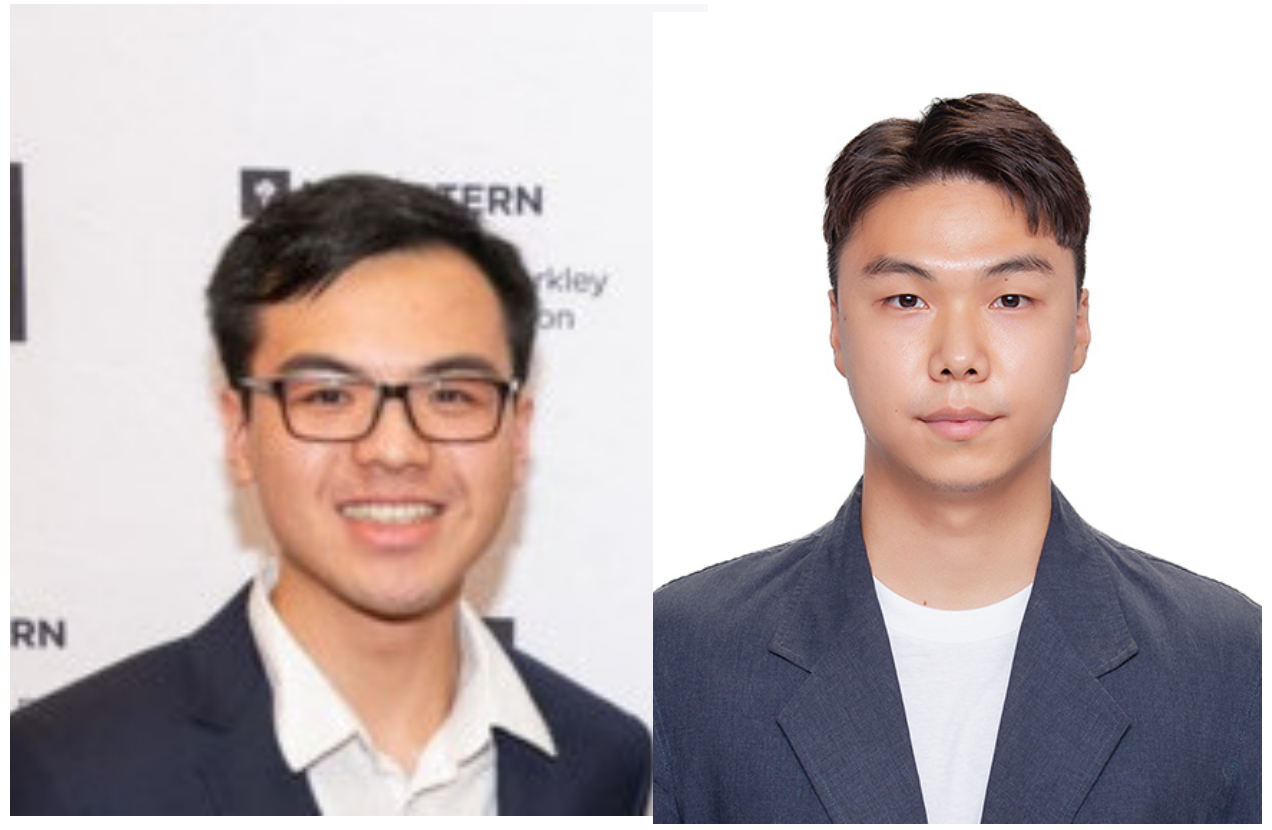


Figure 2 from
Abadie, Diamond & Hainmueller (2010)



In parts joint with

Rex Hsieh and MJ Lee

Classical approach

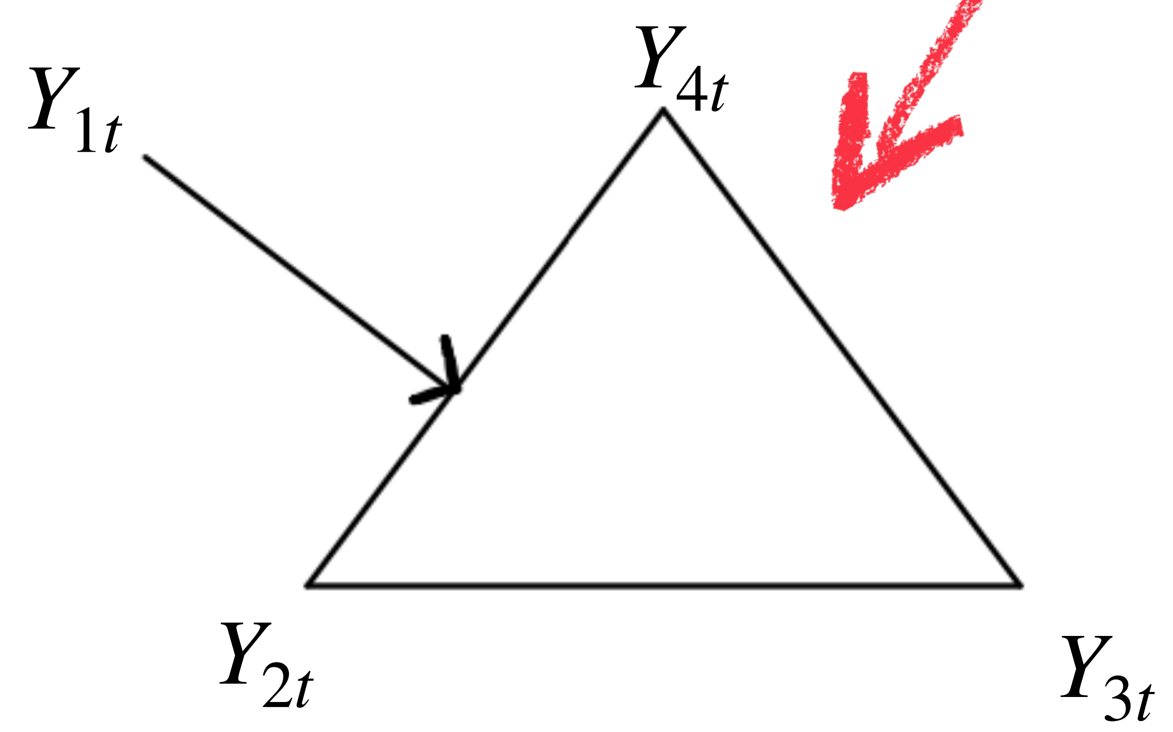
$0 \leq t \leq T$ time periods

$0 < T_0 < T$ pre-intervention periods

$j = 2, \ldots, J+1$ control units

$j = 1$ treatment/target unit

$\left\{ Y_{jt} \right\}_{j=1,\ldots,J+1}$ observable outcomes



Two steps:

1. For all $0 \leq t \leq T_0$:

$$\lambda_t^* = \arg \min_{\lambda \in \Delta^{J-1}} \left| Y_{1t} - \sum_{j=2}^{J+1} \lambda_j Y_{jt} \right|^2$$

(obtain optimal weights over all periods as $\lambda^* = \sum_{t \leq T_0} w_t \lambda_t^*$)

2. For all $t > T_0$:

$$Y_{1t,N} = \sum_{j=2}^{J+1} \lambda_{jt}^* Y_{jt}$$

Distributional approach

Classical version only deals with **aggregate outcomes:**

> **e.g.** aggregate household income, population size, average income, etc.

OT allows to design a **distributional approach** that can deal with entire distributions, which allows to take into account **general heterogeneity of treatment**

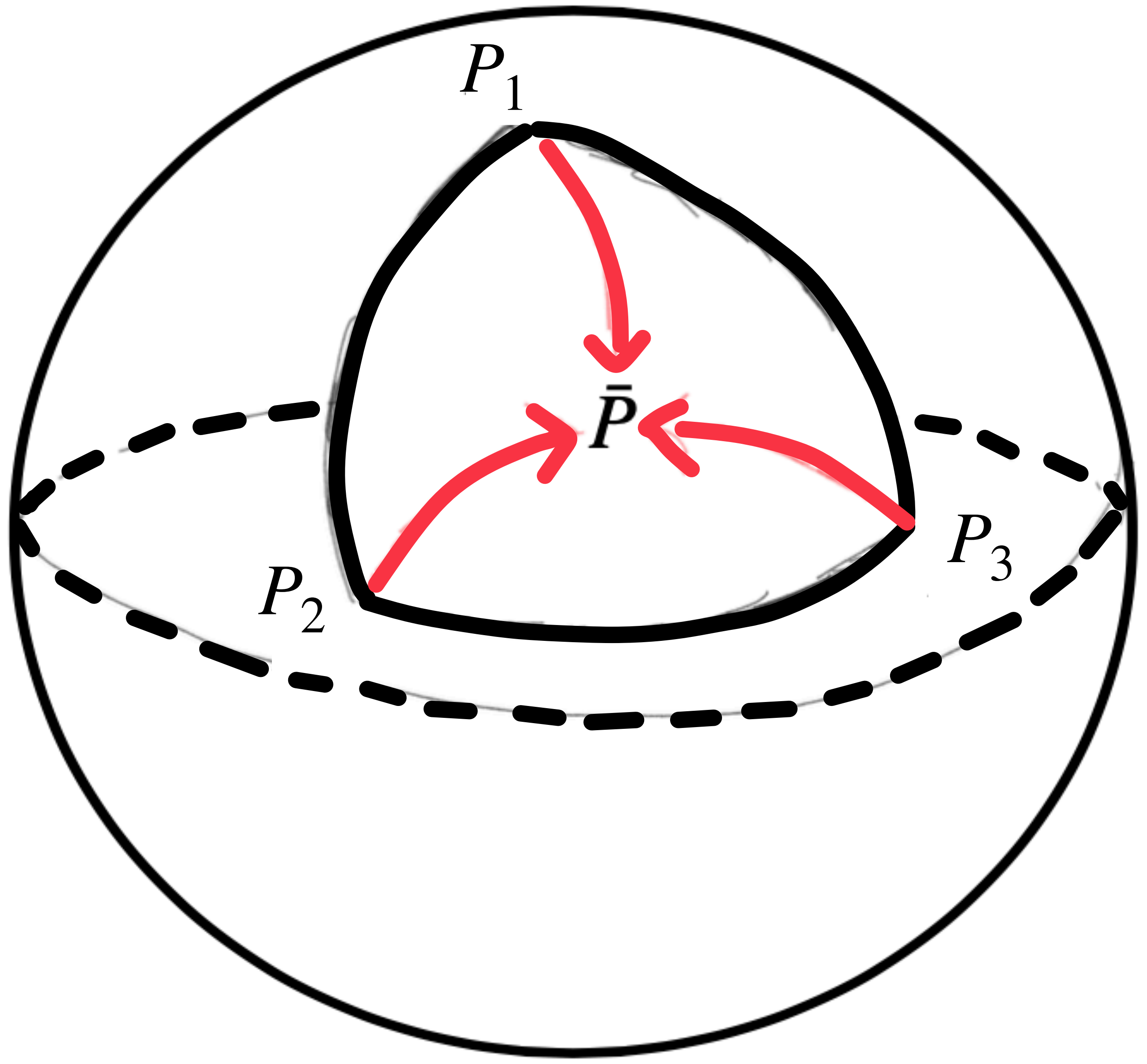
> **e.g.** distribution of individual household income, population movement patterns, individual income, etc.

Wasserstein Projections

Weighted Barycenter

$$\bar{P}(\lambda) = \arg \inf_{P \in \mathscr{P}(\mathbb{R}^d)} \sum_{j=1}^{J} \frac{\lambda_j}{2} W_2^2 \left(P_j, P\right)$$

$$\lambda \equiv (\lambda_1, \ldots, \lambda_d) \in \Delta^J$$

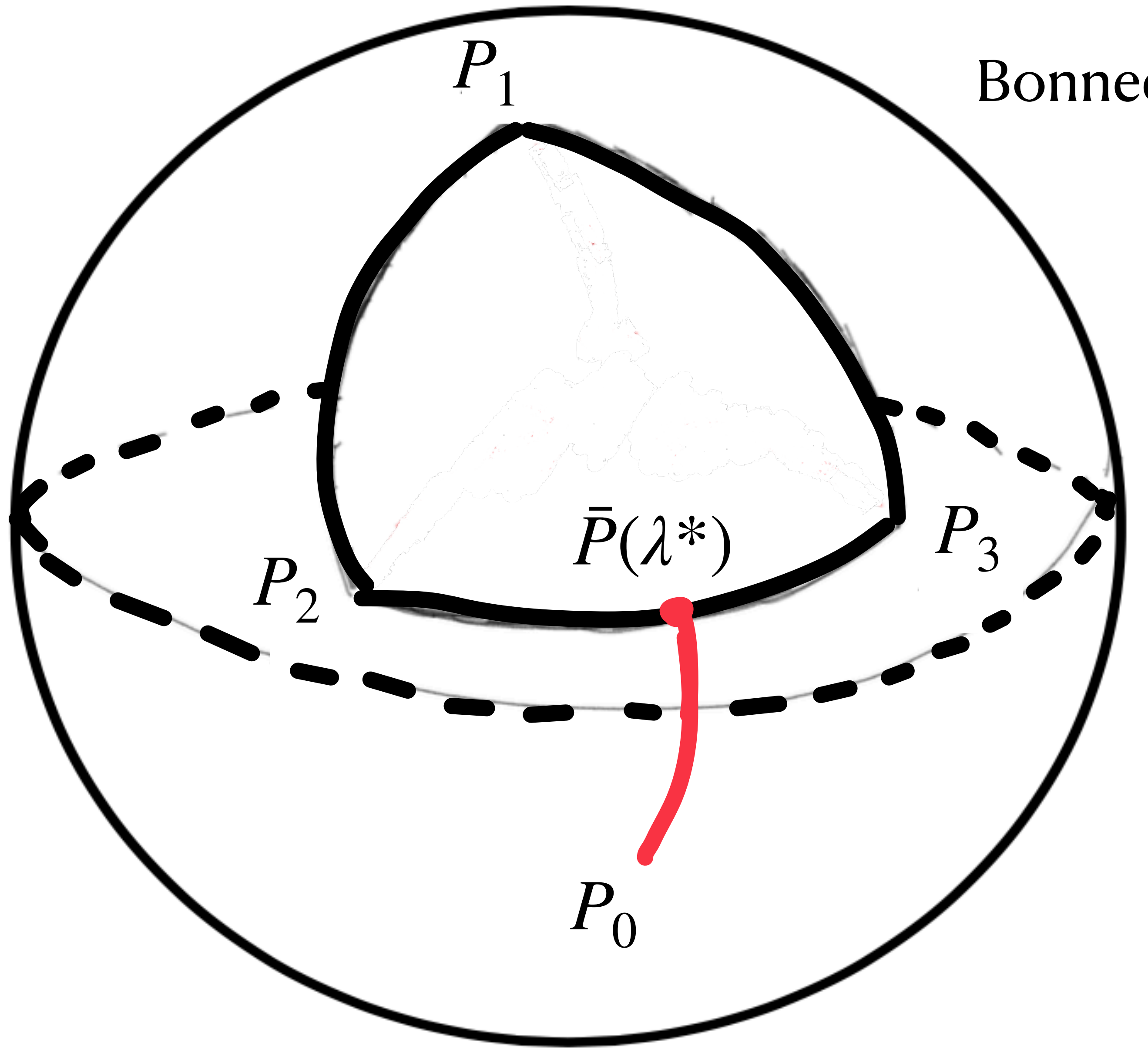


Projection

Bilevel program

$$\lambda^* = \arg\min_{\lambda \in \Delta^J} W_2^2 \left(P_0, \bar{P}(\lambda)\right)$$

$$\text{s.t.} \quad \bar{P}(\lambda) = \arg \inf_{P \in \mathscr{P}(\mathbb{R}^d)} \sum_{j=1}^{J} \frac{\lambda_j}{2} W_2^2 \left(P_j, P\right)$$



Bonneel et. al. (2016)

Using the tangential structure of $\mathscr{W}_2\left(\mathbb{R}^d\right)$

Tangent cone structure for general target measures (AGS 2005):

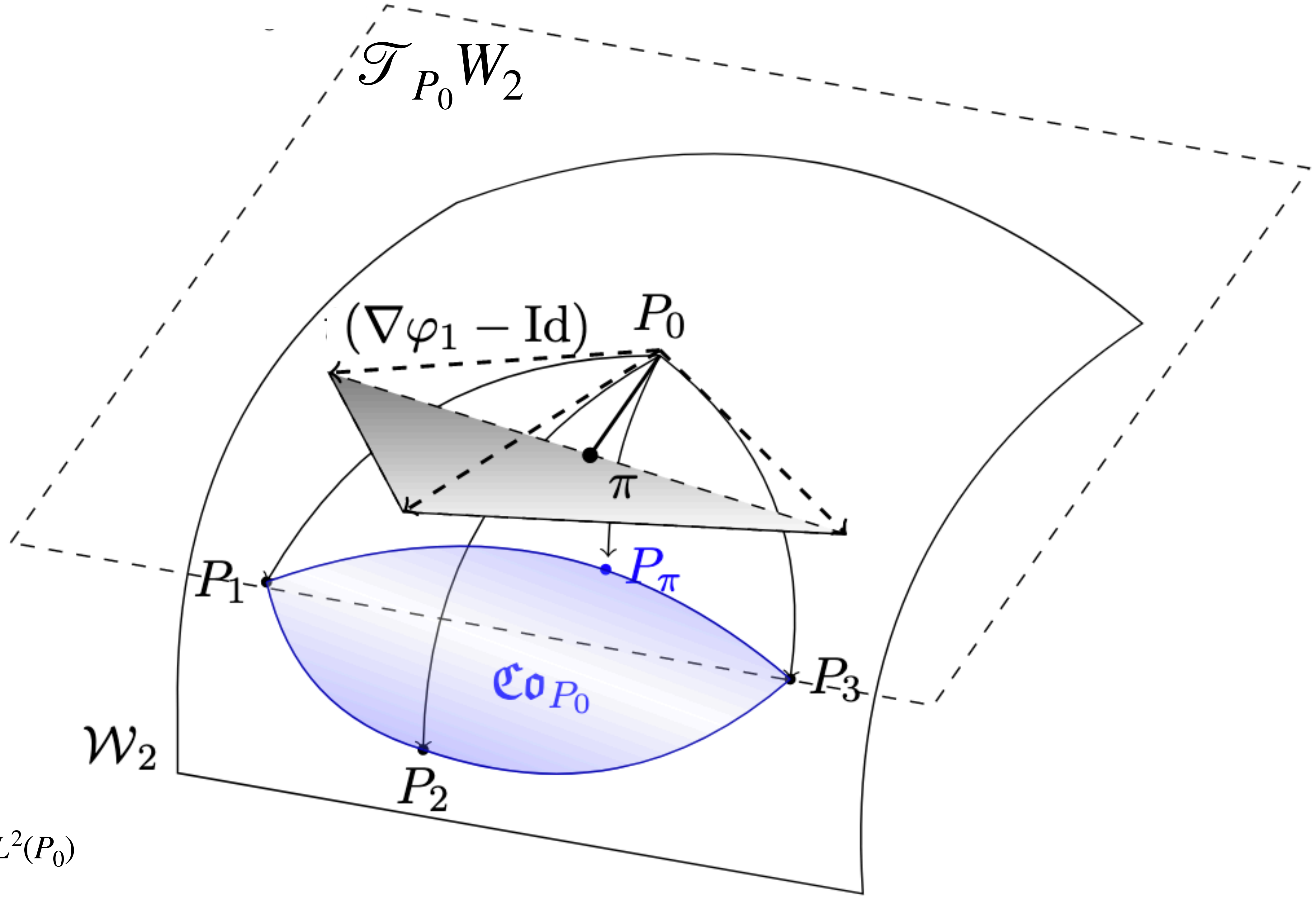
$\mathscr{G}(P_0) \equiv \left\{\gamma \in \mathscr{P}_2(\mathbb{R}^d \times \mathbb{R}^d) : (\pi_1)_\# \gamma = P_0,\ (\pi_1, \pi_1 + \varepsilon\pi_2)_\# \gamma \text{ is optimal for some } \varepsilon > 0\right\}$

closed under local Wasserstein distance

$$W_P^2(\gamma_{12}, \gamma_{13}) \equiv \min\left\{\int_{(\mathbb{R}^d)^3} \left|x_2 - x_3\right|^2 d\gamma_{123} : \gamma_{123} \in \Gamma_1(\gamma_{12}, \gamma_{13})\right\}$$

with the corresponding exponential map

$$\exp_P(\gamma) = (\pi_1 + \pi_2)_\# \gamma\,.$$



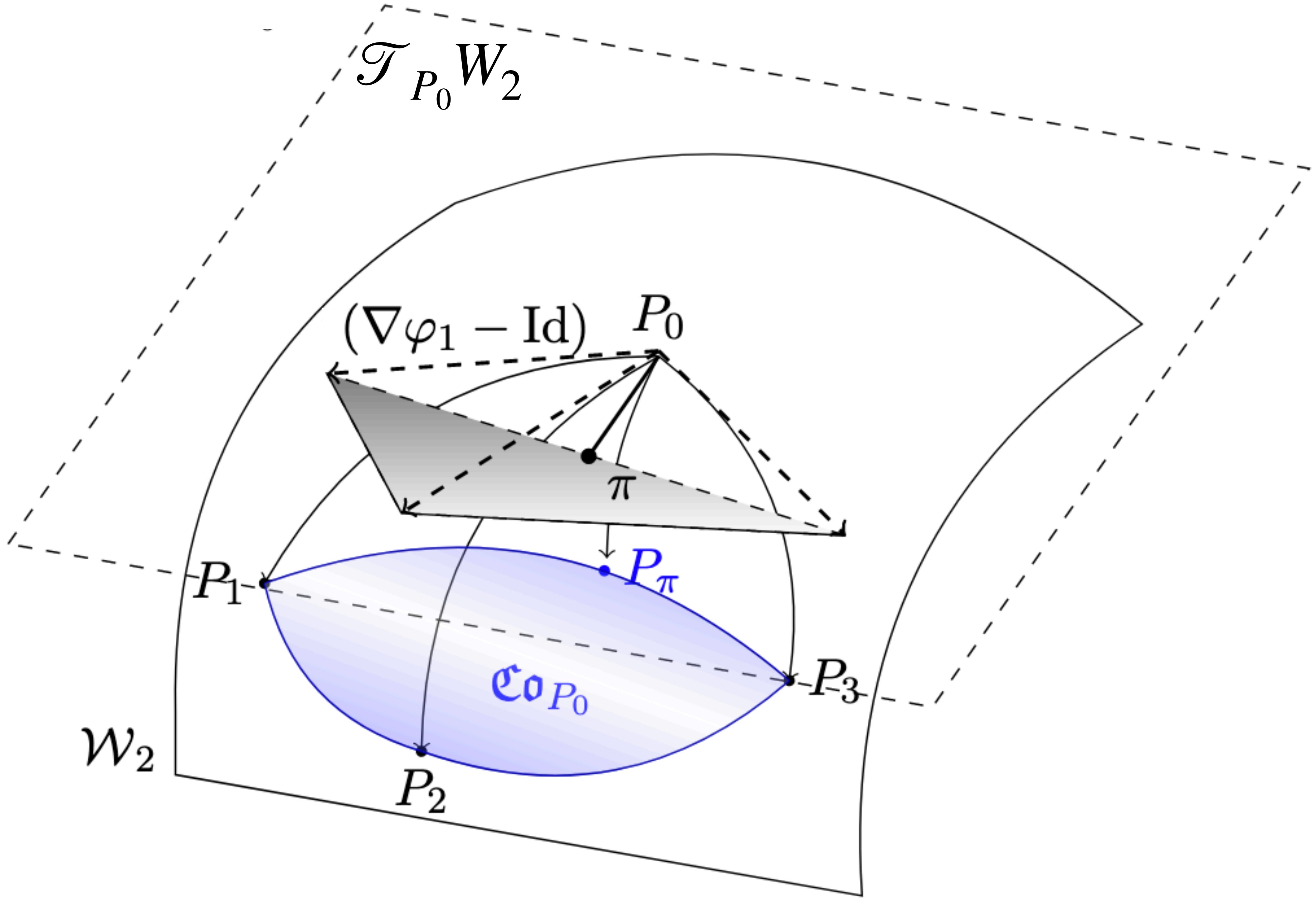
Tangent space for regular targets:

$$\mathscr{T}_{P_0} W_2(\mathbb{R}^d) \equiv \overline{\left\{t(\nabla\varphi_j - \mathrm{Id}) : \left(\mathrm{Id} \times \nabla\varphi_j\right)_\# P_0 \text{ is optimal in } \Gamma(P_0, (\nabla\varphi_j)_\# P_0),\ t > 0\right\}}^{L^2(P_0)}$$

Implementation

Tangential Wasserstein projections (regular target measure):

$$\lambda^* \equiv \arg\min_{\lambda \in \Delta^J} \left\| \sum_{j=1}^{J} \lambda_j \left(\nabla \varphi_j - \mathrm{Id} \right) \right\|^2_{L^2(P_0)}$$

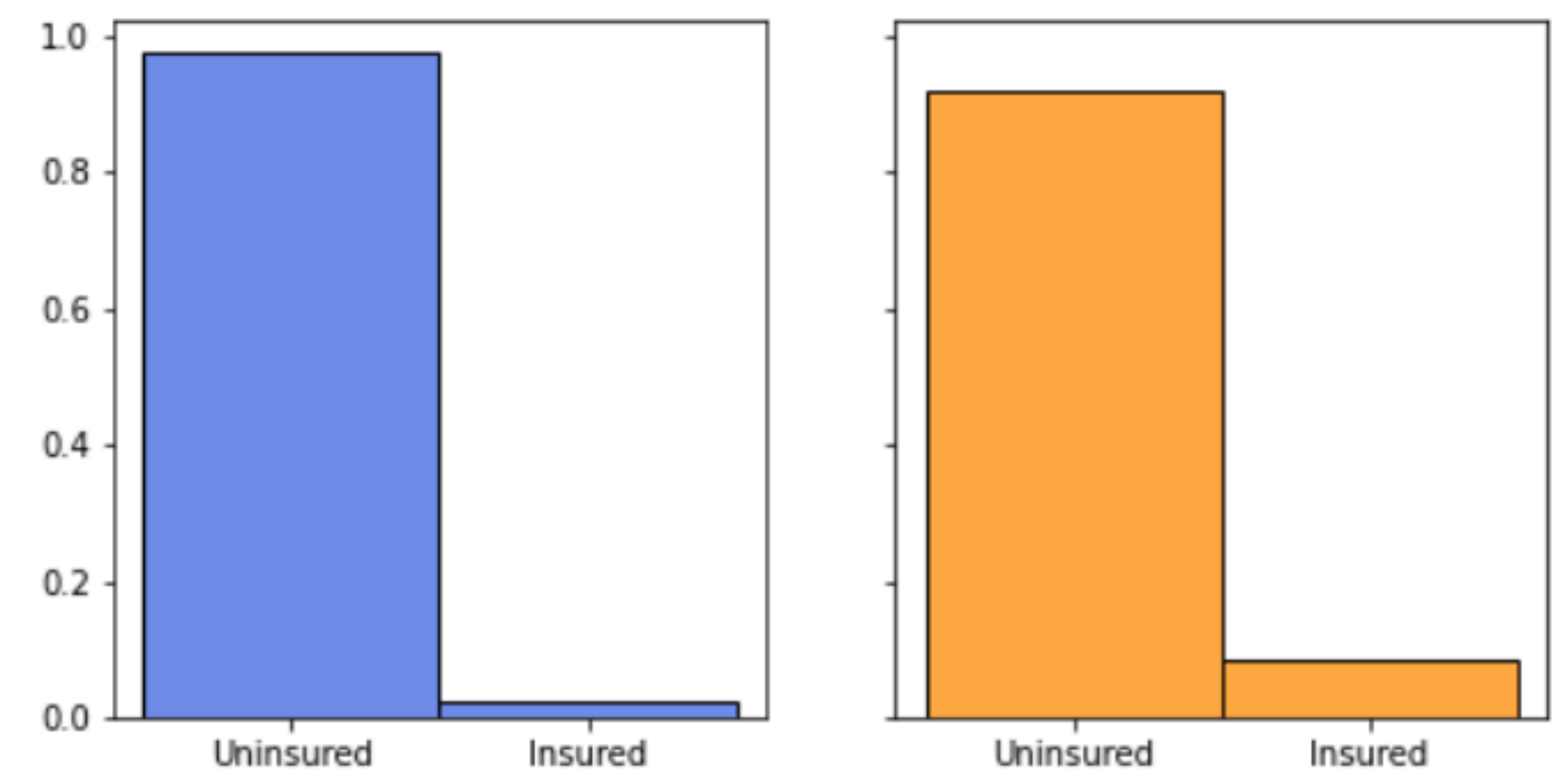


General target measure:

$$\lambda^* \equiv \arg\min_{\lambda \in \Delta^J} \left\| \sum_{j=1}^{J} \lambda_j \left(b_{\gamma_{0j}} - \mathrm{Id} \right) \right\|^2_{L^2(P_0)}$$

$b_{\gamma_{0j}}(x_1) \equiv \int_{\mathbb{R}^d} x_2 \, d\gamma_{0j,x_1}(x_2)$ is the barycentric projection of the optimal transport plans γ_{0j}

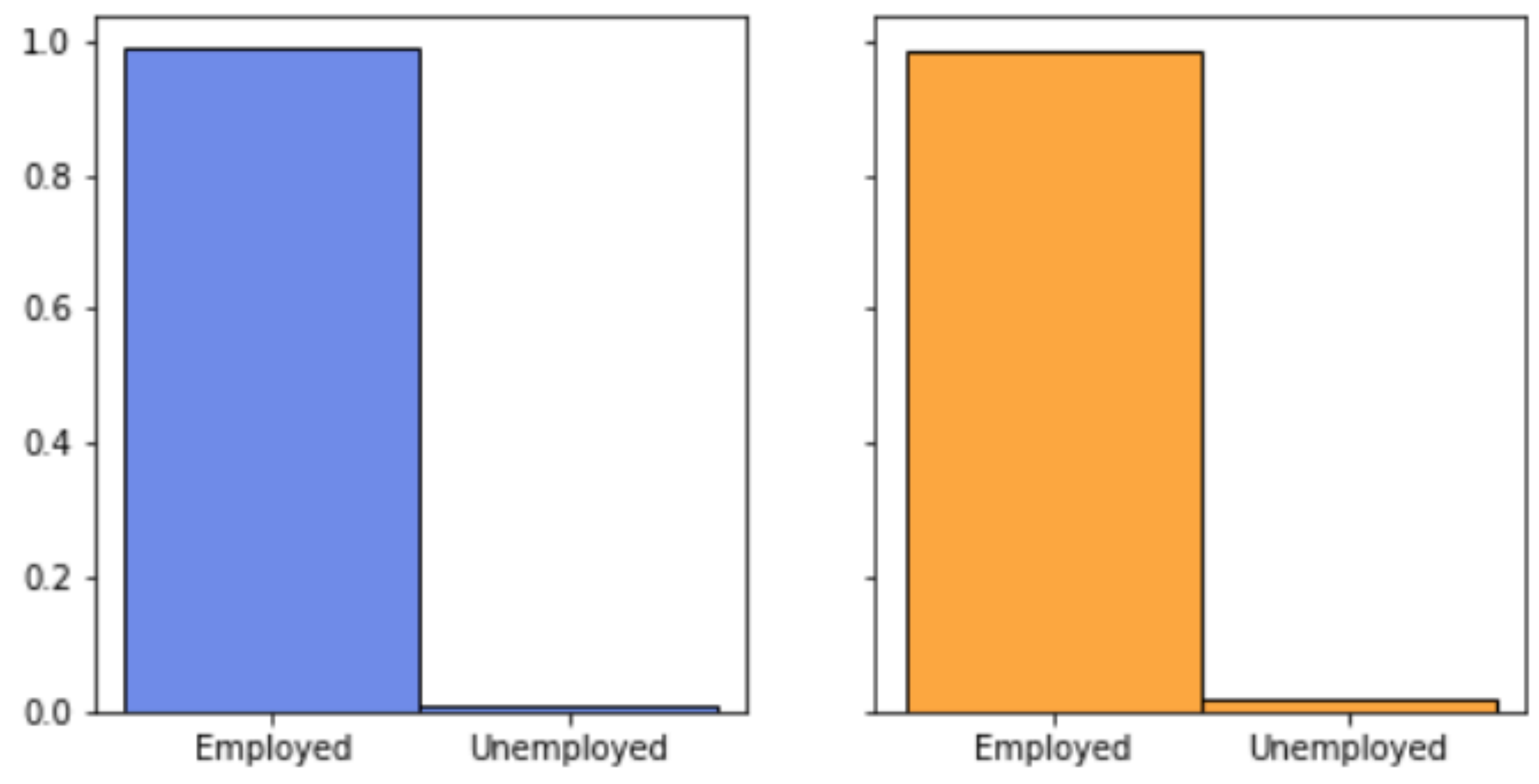
Application: Medicaid in Montana



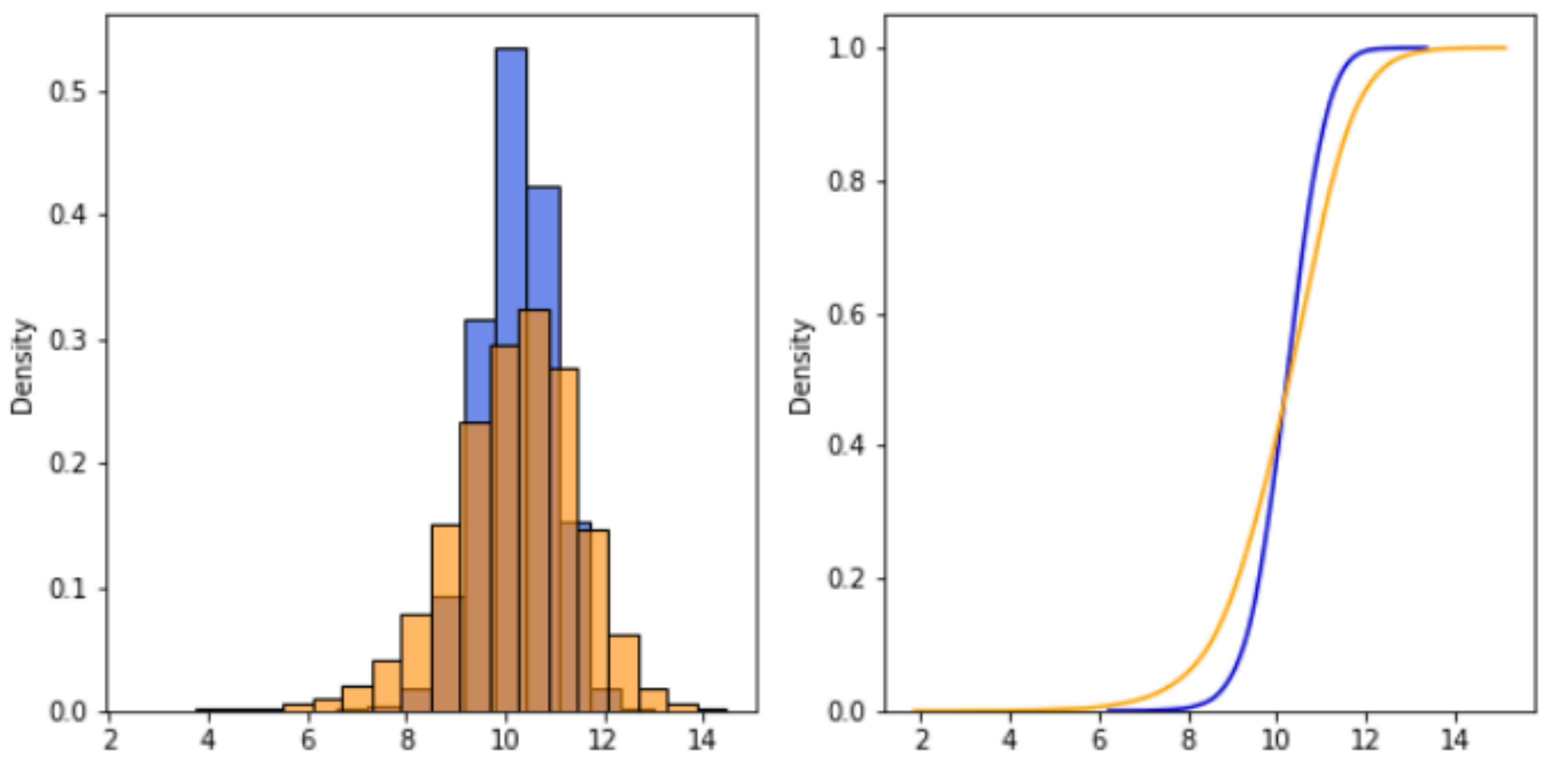
Medicaid coverage

Counterfactual

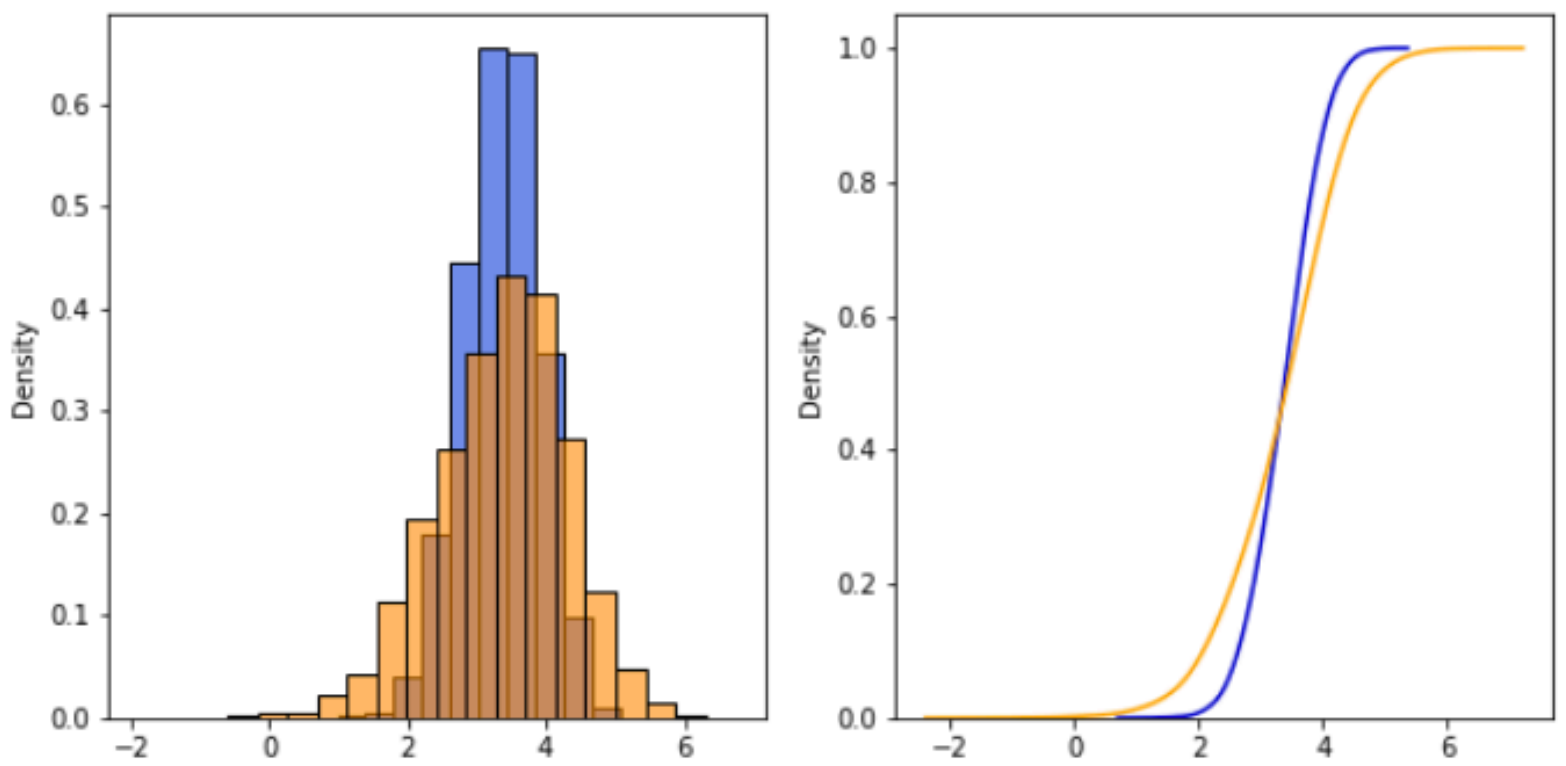
Actual



Employment status



Log wage

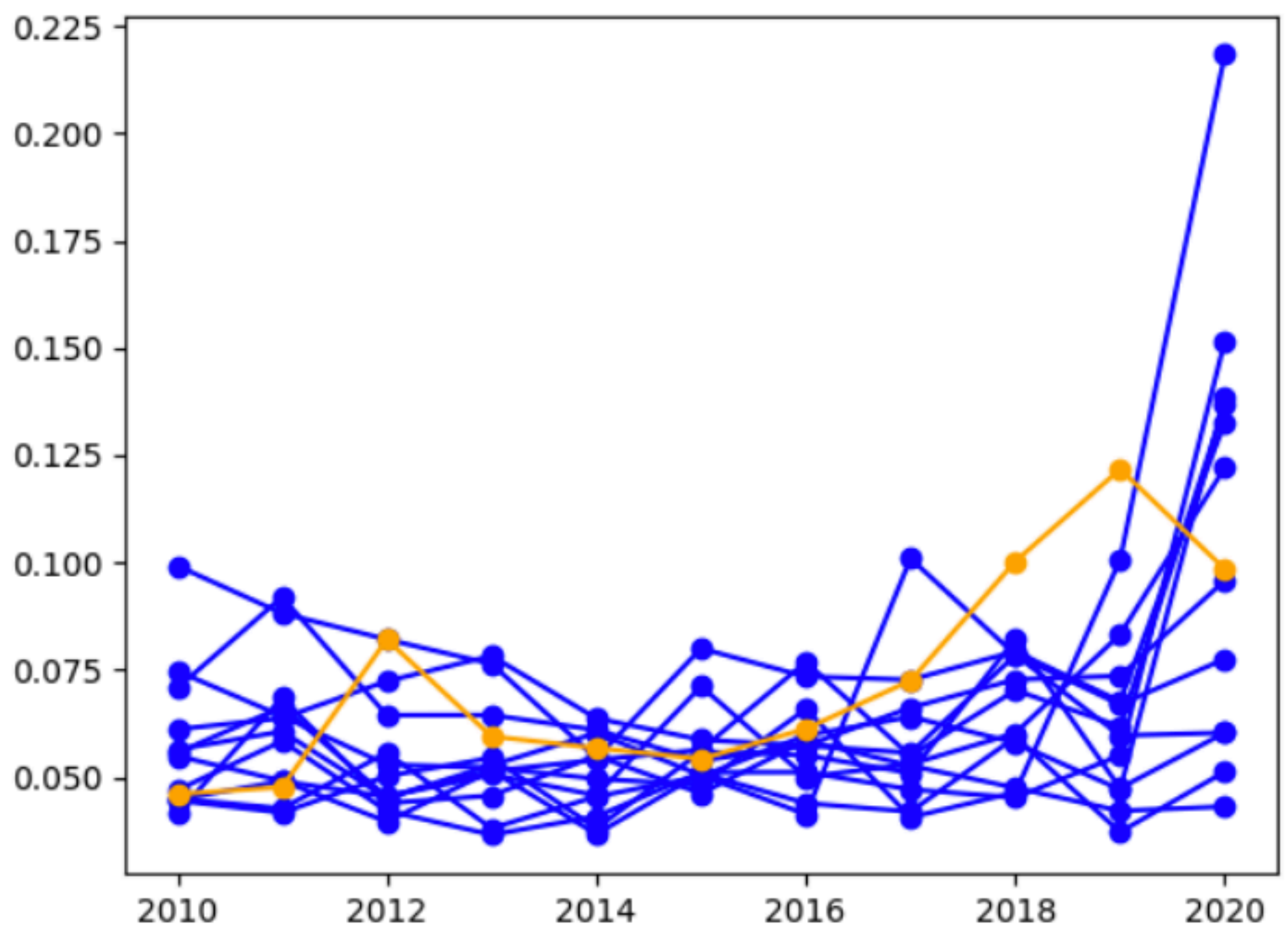


Log labor hours supplied

Weights of control states

Permutation test over time

State	AL	FL	GA	KS	MS	NC	SC	SD	TN	TX	WI	WY
Weight	0.184	0	0	0	0.174	0	0.010	0.513	0	0	0.119	0

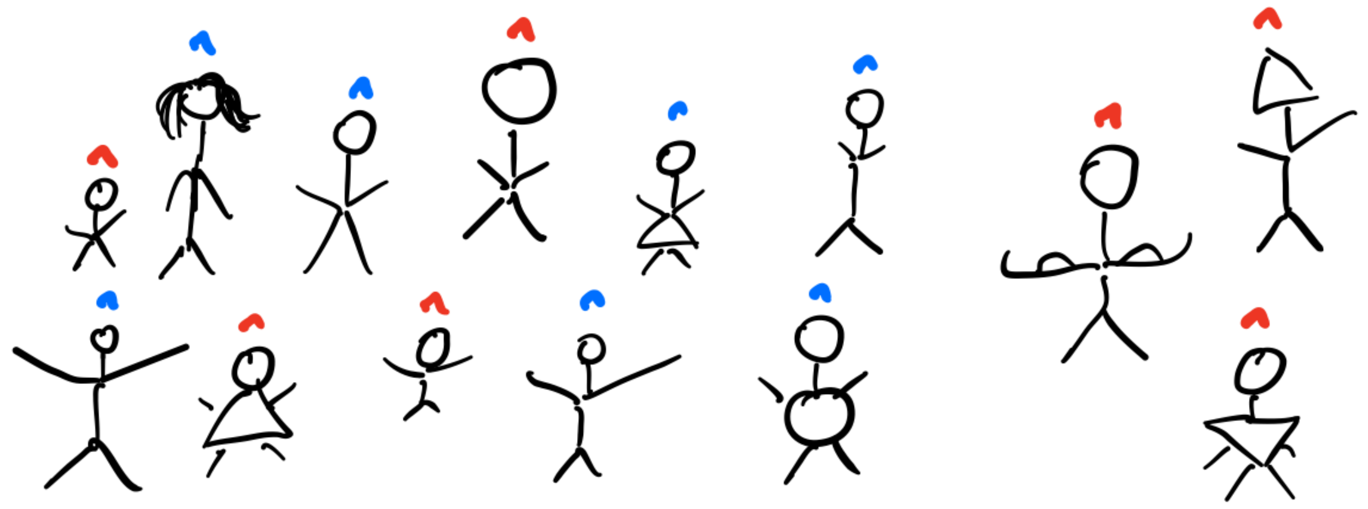


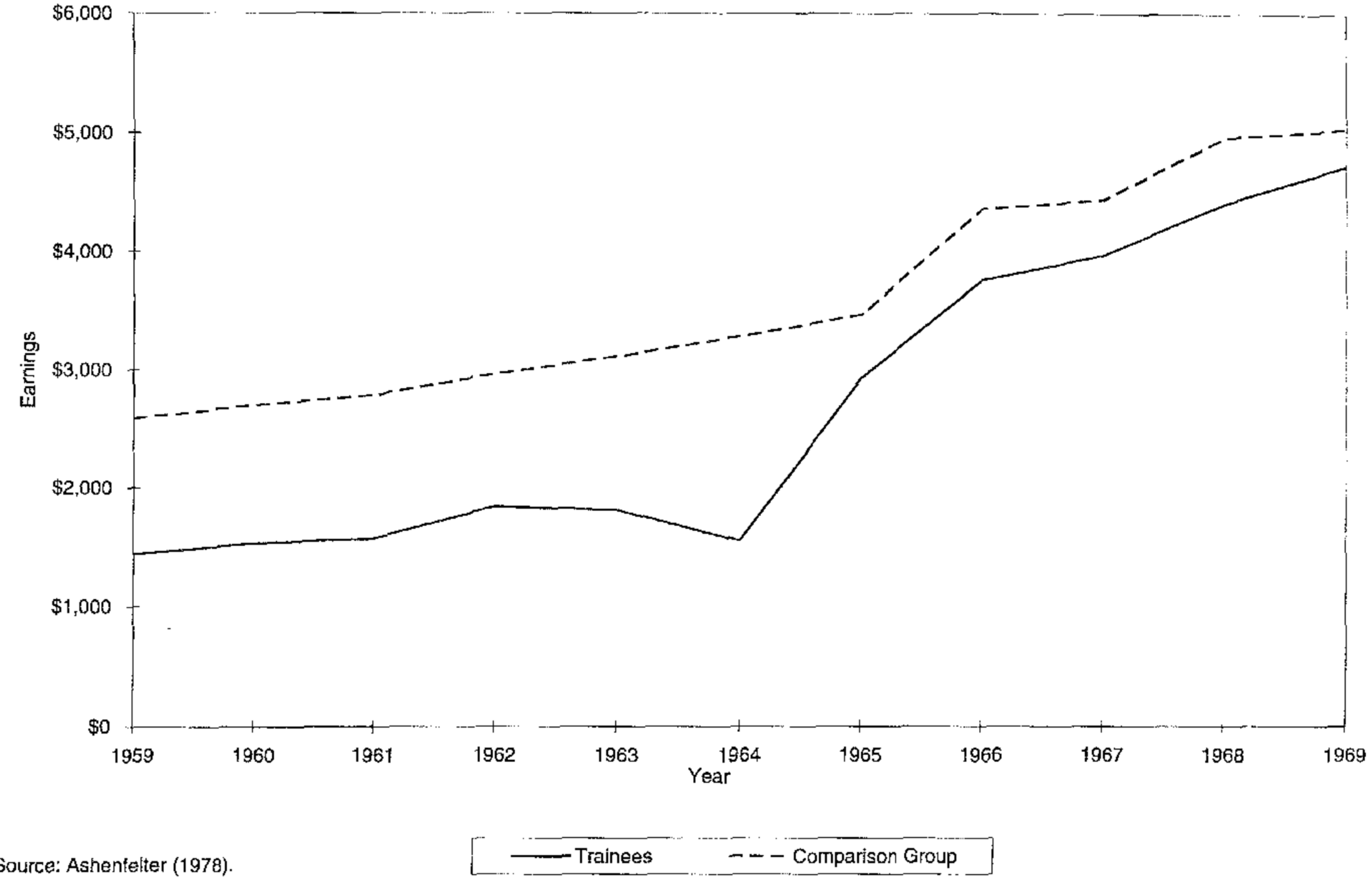
"p-values"

Year (t)	p_t (Weights Using All Years)	p_t (Averaged Weights Over All Years)
2017	0.231	0.308
2018	0.077	0.077
2019	0.077	0.077
2020	0.535	0.385

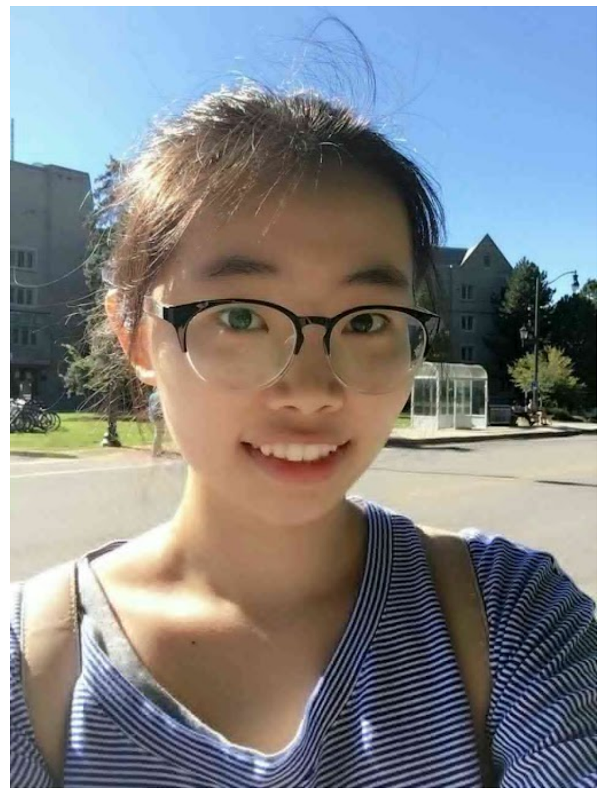
3. Optimal transport as a matching estimator

Treatment Control



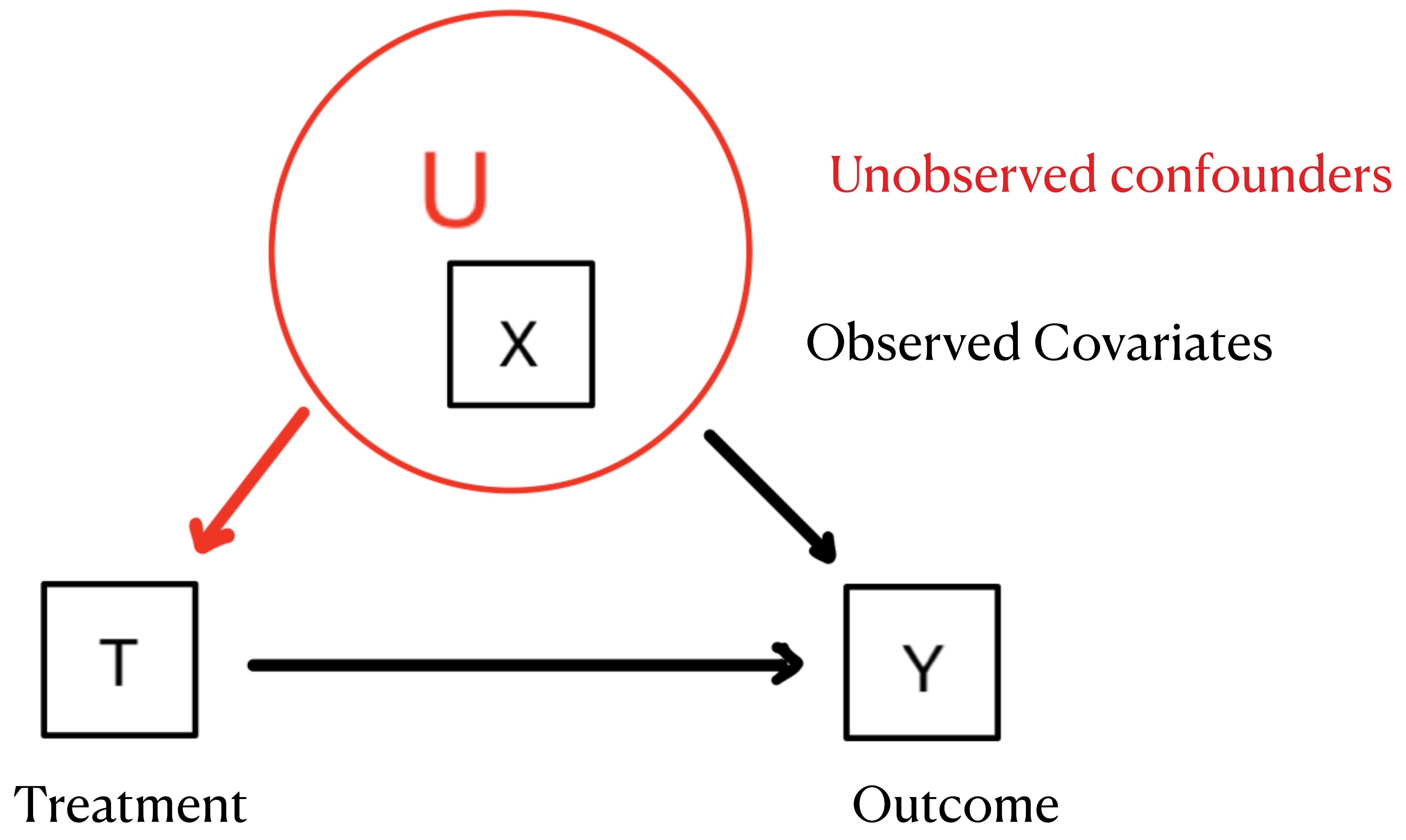


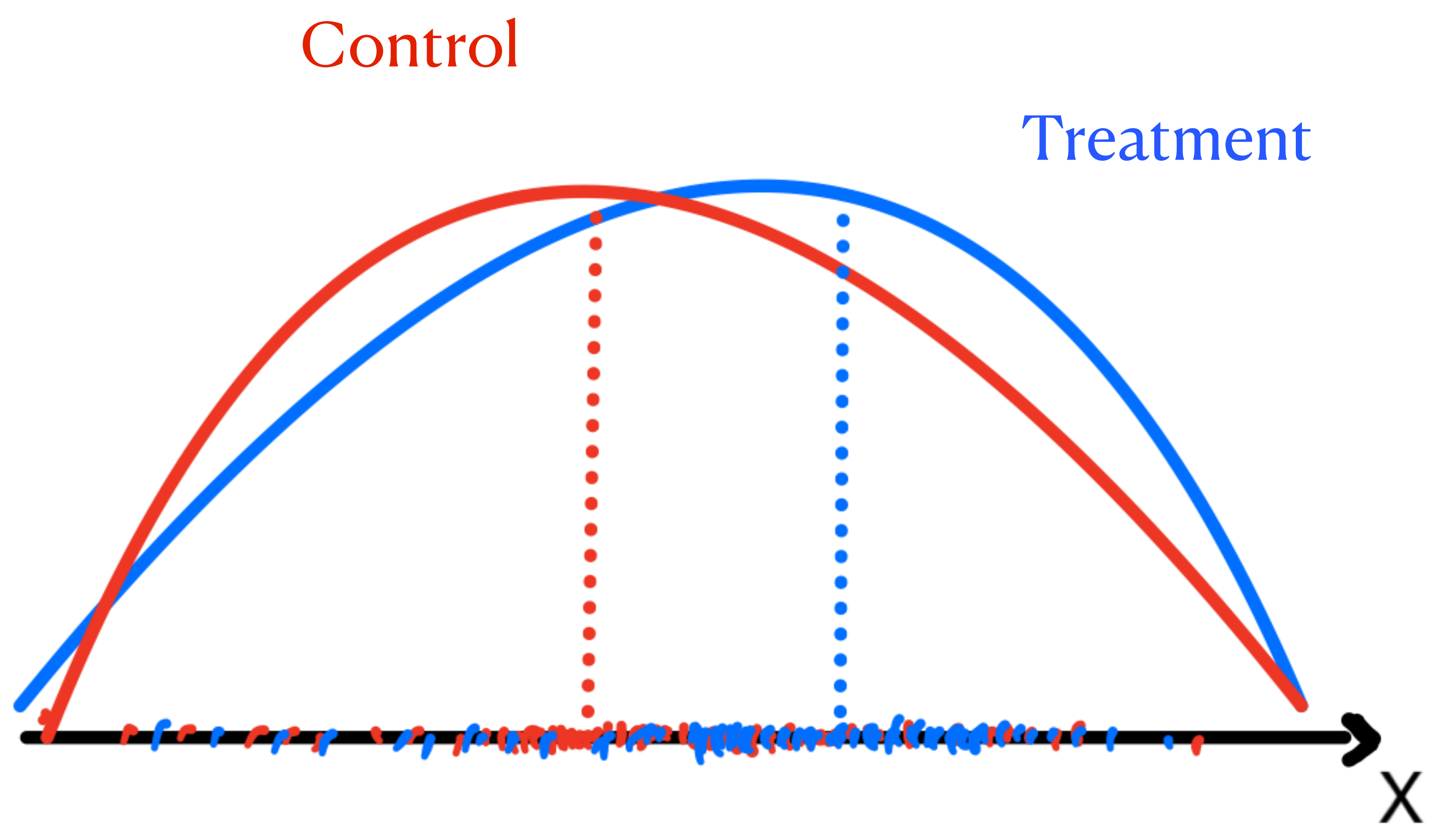
Source: Ashenfelter (1978).



joint work with Yuliang Xu

The problem: Unobserved confounders





Matching to correct for endogeneity bias

$\boxed{\text{Use information from covariates!}}$

Potential outcome notation

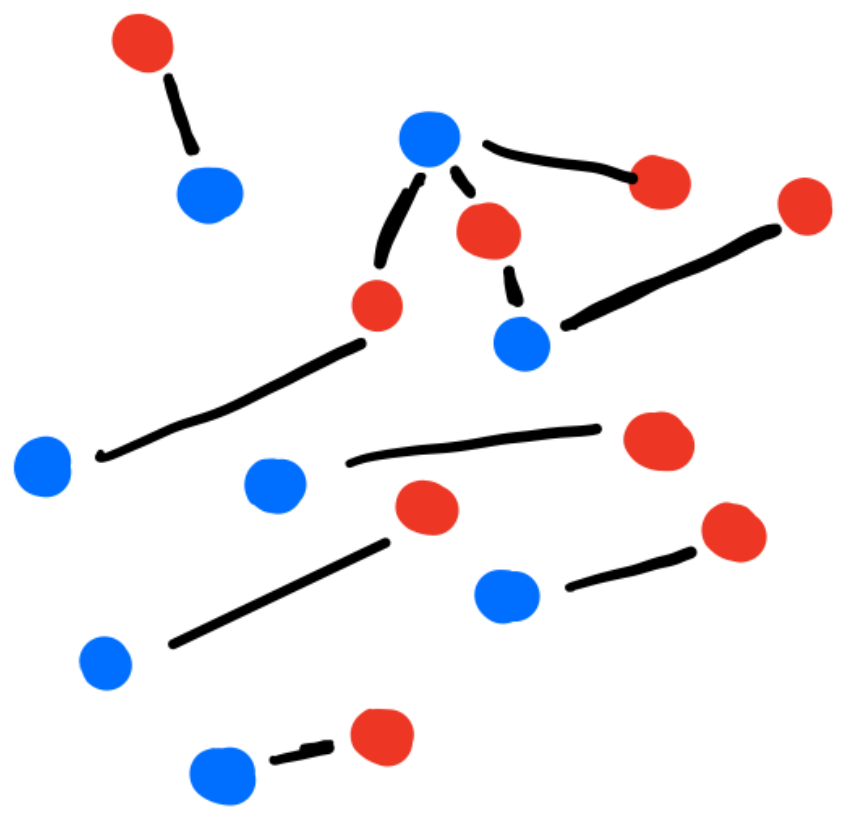
$T = 1,\ldots,J$	Treatment
$Y(j) \in \mathbb{R}, \quad j = 1,\ldots,J$	Potential outcomes
$X \in \mathbb{R}^d$	Observed covariates
$D(j) \in \left\{ \tau \in \{0,1\}^J : \sum_{j=1}^{J} \tau_j = 1 \right\}$	Treatment indicator

$T = j \quad \Longleftrightarrow \quad D(j) = 1$

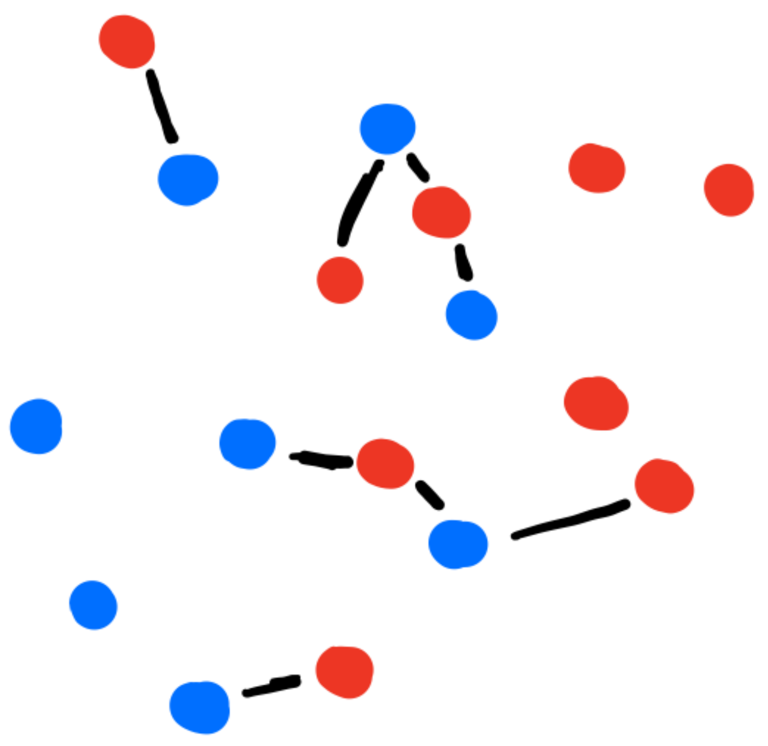
Assumptions for causal inference

$\boxed{Y(j) \perp D(j) \mid X}$		Weak unconfoundedness
$0 < \rho \leq P(D(j) = 1 \mid X) \leq 1 - \rho < 1$		Overlap
$X \mid D(j) \sim \mu_j$		Covariate distribution
$y(X, j) = \mathbb{E}(Y \mid T = j, X)$	bd, cont	Regularity

Key to reduce bias: use unbalanced optimal transport



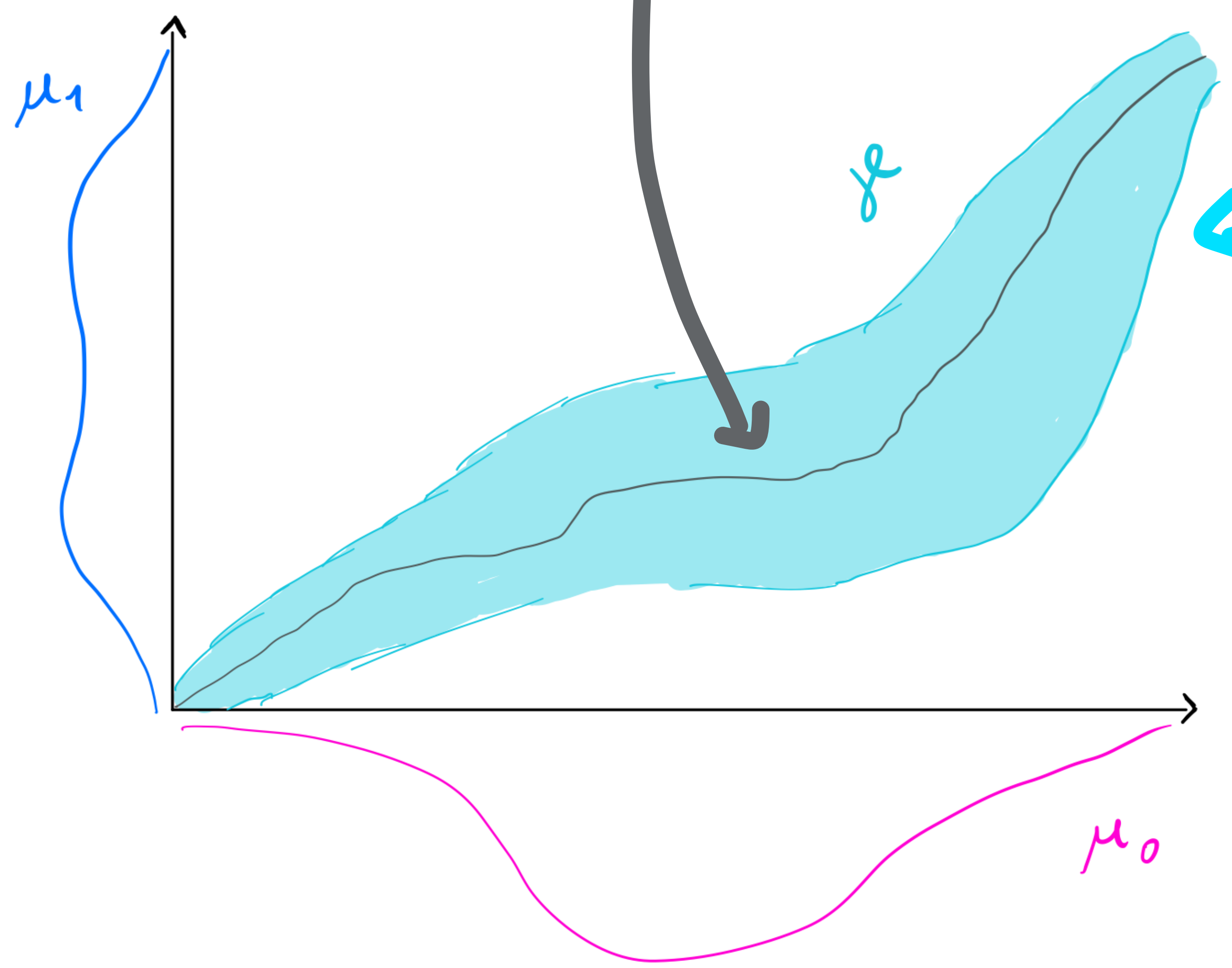
Classical (balanced) OT:
All elements are matched



Unbalanced OT:
Only keep good matches

Unbalanced optimal transport

$$\inf_{\gamma\in\mathcal{M}^+(\mathcal{X})}\int_{\mathcal{X}} c(x)d\gamma(x) \quad + \quad \varepsilon KL\Big(\gamma\,||\bigotimes_{j=1}^{J}\mu_j\Big) \quad + \quad \sum_{j=1}^{J} D_\phi\left(\pi_j\gamma\,||\,\mu_j\right)$$



Relaxing the constraint that measures have to have the same overall mass

$$KL\Big(\gamma\,||\bigotimes_j\mu_j\Big) \equiv \int_{\mathcal{X}} \ln\frac{d\gamma}{d\bigotimes_j\mu_j}(x)d\gamma(x) + \left(\int_{\mathcal{X}} d\bigotimes_j\mu_j(x) - \int_{\mathcal{X}} d\gamma(x)\right)$$

$$D_\phi\left(\mu\,||\,\nu\right) \equiv \int_{\mathcal{X}}\phi\left(\frac{d\mu}{d\nu}\right)d\nu + \phi'_\infty\int_{\mathcal{X}} d\mu^\perp$$

Galichon & Salanié (2010), Cuturi (2013)

Causal effects via unbalanced OT matching

DEFINITION 1. For the j-th treatment and $t \neq j$ denote $\gamma_{j|t}$ as the conditional measure of covariates in group j given the covariates in group t under the joint distribution γ. Then under Assumption 1, the expected potential outcome can be expressed in the sample version as

$$\hat{\mathbb{E}}_N\left[\hat{Y}(j)\right] = \frac{1}{N}\sum_{i=1}^{N} Y_i I(D_i(j)=1) + \frac{1}{N}\sum_{i=1}^{N}\sum_{t\neq j}\sum_{k\neq i} Y_k \hat{\gamma}_{N,j|t}(X_k|X_i) I(D_i(t)=1), \tag{10}$$

where $N = \sum_{j=1}^{J} N_j$ is the overall number of sample points over all treatment arms and $\hat{\gamma}_N$ is the empirical counterpart to the optimal matching estimated via the generalized Sinkhorn algorithm (8) by replacing μ_j with the empirical measures $\hat{\mu}_{N_j}$ defined below in (11).

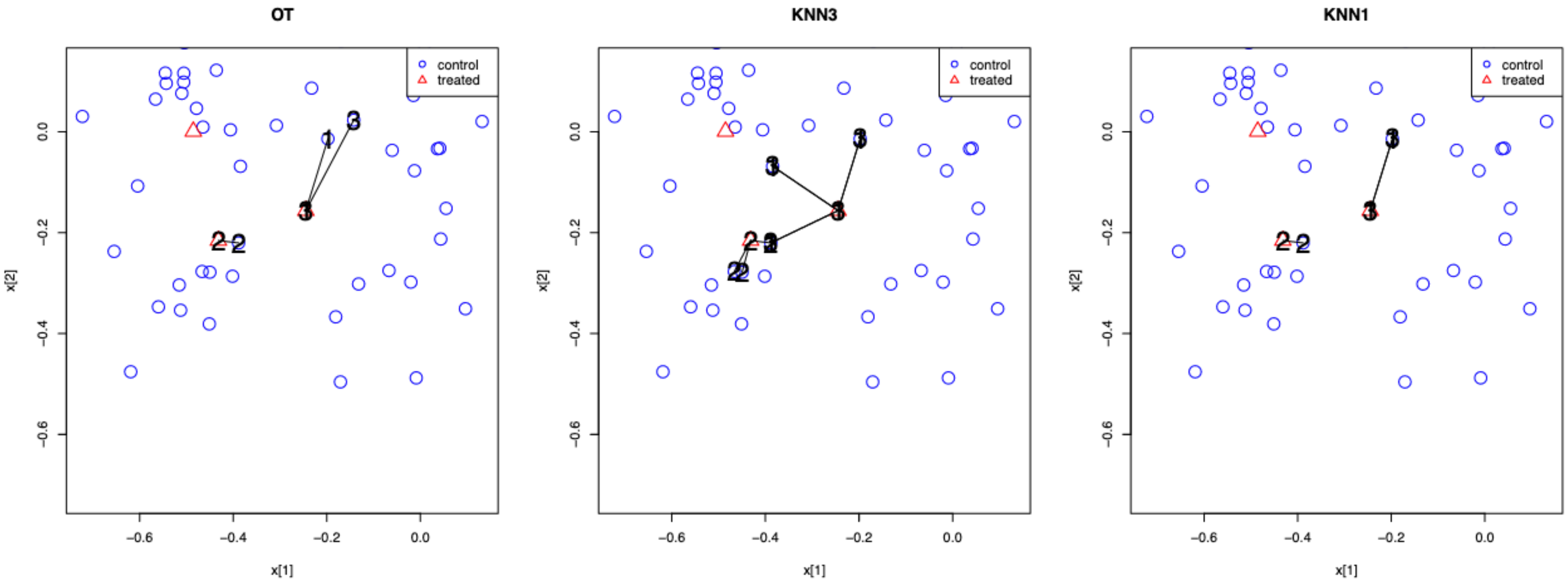


Fig 2: Top 3 pairs with the largest unbalanced OT weights in the simulation case 2, and the corresponding KNN matches for the selected treated individuals.

Conclusion

- Overview of **some** settings in causal inference where optimal transport is interesting and can be useful.
- OT is often an interesting choice when dealing with **heterogeneity** in the treatment effect
- Many other applications and settings: instrumental variables, domain adaptation, etc.
- Problems in causal inference can inform new optimal transport estimators