Counterfactual Cocycles: A Framework for Robust and Coherent Counterfactual Transports

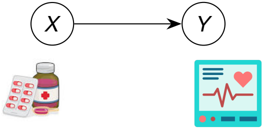
Hugh Dance[1], Benjamin Bloem-Reddy[2]

October 22, 2025

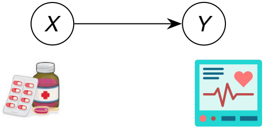
[1] PhD Student, Gatsby Unit (P. Orbanz lab), UCL
[2] Assistant Professor, Department of Statistics, University of British Columbia

RCT as the “gold-standard” for marginal counterfactual inference

Goal: Assess effectiveness of medication $X \in \{0, 1\}$ on symptoms $Y := (Y_1, \dots, Y_p)$.



Goal: Assess effectiveness of medication $X \in \{0, 1\}$ on symptoms $Y := (Y_1, \ldots, Y_p)$.



Randomized Control Trial (RCT)

- Observe treatment units $\{Y^{(i)}(1)\}_{i=1}^{n_1}$ and control units $\{Y^{(i)}(0)\}_{i=n_1+1}^{n_1+n_0}$.
- Potential outcomes satisfy $Y^{(i)}(x) \sim \mathbb{P}_{Y|X=x}$

RCT as the “gold-standard” for marginal counterfactual inference

Goal: Assess effectiveness of medication $X \in \{0,1\}$ on symptoms $Y := (Y_1, \dots, Y_p)$.

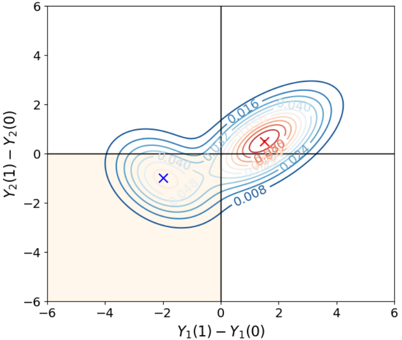
$$X \longrightarrow Y$$

Randomized Control Trial (RCT)

- Observe treatment units $\{Y^{(i)}(1)\}_{i=1}^{n_1}$ and control units $\{Y^{(i)}(0)\}_{i=n_1+1}^{n_1+n_0}$.
- Potential outcomes satisfy $Y^{(i)}(x) \sim \mathbb{P}_{Y|X=x}$

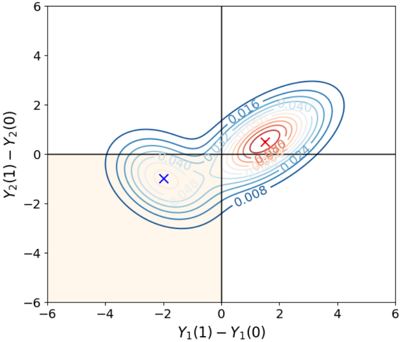
$$\text{ATE} = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] \implies \widehat{ATE} = \frac{1}{n_1}\sum_{i=1}^{n_1} Y^{(i)}(1) - \frac{1}{n_0}\sum_{i=n_1+1}^{n_0} Y^{(i)}(0)$$

Goal: Quantify possible treatment harms



3 Shen et al. (2013). Treatment Benefit and Treatment Harm Rate to Characterize Heterogeneity in Treatment Effect, Biometrics.

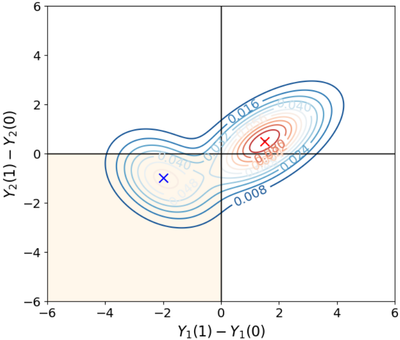
Goal: Quantify possible treatment harms



Treatment Harm Rate (THR)[3] $:= \mathbb{P}(Y(1) \prec Y(0)) = \mathbb{P}\big(\{Y_1(1) \prec Y_1(0)\} \cap \{Y_2(1) - Y_2(0)\}\big)$

[3] Shen et al. (2013). Treatment Benefit and Treatment Harm Rate to Characterize Heterogeneity in Treatment Effect, Biometrics.

Goal: Quantify possible treatment harms



Treatment Harm Rate (THR)[3] $:= \mathbb{P}(Y(1) \prec Y(0)) = \mathbb{P}\big(\{Y_1(1) \prec Y_1(0)\} \cap \{Y_2(1) - Y_2(0)\}\big)$

Issue: Only observe $Y^{(i)}(0)$ or $Y^{(i)}(1)$ for each unit - cannot estimate THR from data!

[3] Shen et al. (2013). Treatment Benefit and Treatment Harm Rate to Characterize Heterogeneity in Treatment Effect, Biometrics.

The Need for Counterfactual Couplings

Underlying issue: need coupling $\mathbb{P}_{Y(1),Y(0)}$ to estimate $\mathbb{P}_{Y(1)-Y(0)}$

$$\text{THR} := \iint \mathbf{1}\{y_1 - y_0 \leq 0\}\, d\mathbb{P}_{Y(1),Y(0)}(y_1, y_0)$$

Other Examples

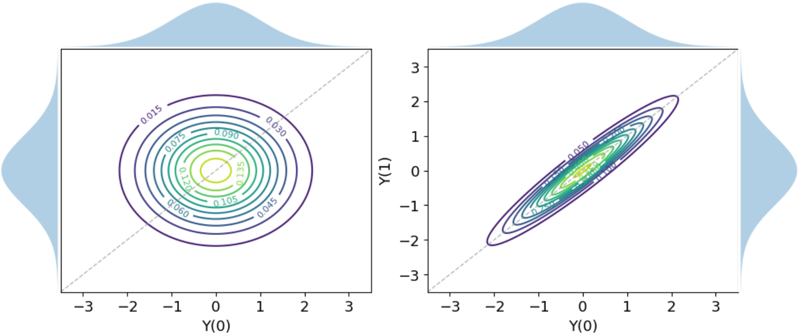
- Median Treatment Effect (MTE)[5]: $\text{Med}(Y(1) - Y(0))$
- Conditional Value at Risk (CVar)[9]: $\mathbb{E}[Y(1) - Y(0) | Y(1) - Y(0) \leq q_\alpha]$
- Treatment Benefit Rate (TBR)[13]: $\mathbb{P}(Y(1) \succ Y(0))$

[7] Lee, M.J. (2000). Median Treatment Effect in Randomized Trials, JRSSB-B.

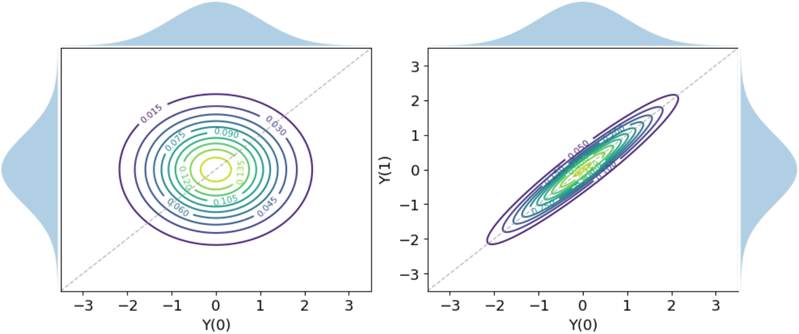
[11] Kallus, N. (2023). Treatment effect risk: Bounds and inference, Management Science.

[15] Shen et al. (2013). Treatment Benefit and Treatment Harm Rate to Characterize Heterogeneity in Treatment Effect, Biometrics.

- Infinitely many admissible couplings $\pi \in \Pi(\mathbb{P}_{Y(1)}, \mathbb{P}_{Y(0)})$
- Can never observe joint samples $(Y(1), Y(0))$



- Infinitely many admissible couplings $\pi \in \Pi(\mathbb{P}_{Y(1)}, \mathbb{P}_{Y(0)})$
- Can never observe joint samples $(Y(1), Y(0))$

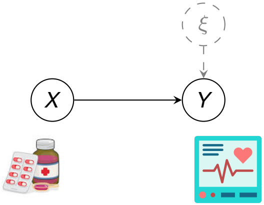


Existing Approaches

- Structural Causal Models (SCMs)
- Optimal Transport Methods (OT)

Existing Methods

Structural Causal Models[16]

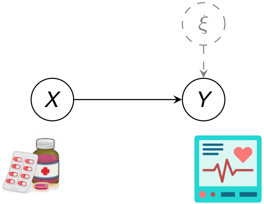


Structural Causal Model (SCM): $Y = f(X, \xi), \quad \xi \sim \mathbb{P}_\xi, \quad \xi \perp\!\!\!\perp X$

Counterfactuals: $Y(x) = f_x(\xi) := f(x, \xi) \implies$ induced coupling: $(Y(1), Y(0)) = (f_1(\xi), f_0(\xi))$

[16] Spirtes et al. (2000), Pearl et al. (2009), Bongers et al. (2021).

Structural Causal Models[16]

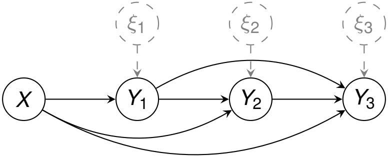


Structural Causal Model (SCM): $Y = f(X, \xi), \quad \xi \sim \mathbb{P}_\xi, \quad \xi \perp\!\!\!\perp X$

Counterfactuals: $Y(x) = f_x(\xi) := f(x, \xi) \implies$ induced coupling: $(Y(1), Y(0)) = (f_1(\xi), f_0(\xi))$

. . . how to identify $(f, \mathbb{P}_\xi)$ from data?

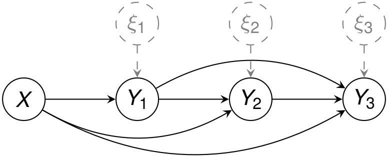
[16] Spirtes et al. (2000), Pearl et al. (2009), Bongers et al. (2021).



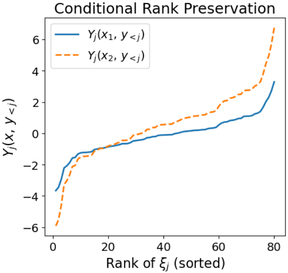
Causal Ordering: $Y_1 \prec Y_2 \prec \cdots \prec Y_p \quad \Longrightarrow \quad$ **Acyclic-SCM:** $Y_j = f_j(X, Y_{<j}, \xi_j) \quad \forall j$

[17] Bogachev, V.I. et al. (2005) "Triangular transformations of measures", Sbornik: Mathematics.

Identifiability via Bijective Causal Models (BCMs)



Causal Ordering: $Y_1 \prec Y_2 \prec \cdots \prec Y_p \quad \Longrightarrow \quad$ **Acyclic-SCM:** $Y_j = f_j(X, Y_{<j}, \xi_j) \quad \forall j$



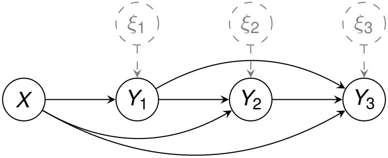
Conditions for Identifiability[17]

- Each f_j is monotone increasing on ξ_j
- $\mathbb{P}_{Y|X=x}$ and $\mathbb{P}_\xi$ are abs. continuous on $\mathbb{R}^d$

$\Rightarrow f_x : \mathcal{E} \to \mathbb{Y}$ is *triangular, monotone, increasing*

[17] Bogachev, V.I. et al. (2005) "Triangular transformations of measures", Sbornik: Mathematics.

Identifiability via Bijective Causal Models (BCMs)

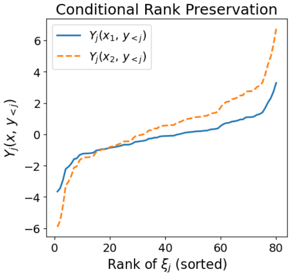


Causal Ordering: $Y_1 \prec Y_2 \prec \cdots \prec Y_p \implies$ **Acyclic-SCM:** $Y_j = f_j(X, Y_{<j}, \xi_j) \quad \forall j$

Conditions for Identifiability[17]

- Each f_j is monotone increasing on ξ_j
- $\mathbb{P}_{Y|X=x}$ and $\mathbb{P}_\xi$ are abs. continuous on $\mathbb{R}^d$

$\Rightarrow f_x : \mathcal{E} \to \mathbb{Y}$ is *triangular, monotone, increasing*

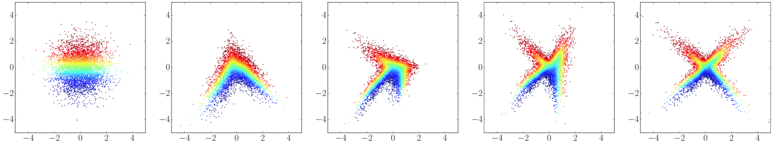


Can augment model with measured causes Z (age, gender, gene expression...)

[17] Bogachev, V.I. et al. (2005) "Triangular transformations of measures", Sbornik: Mathematics.

Popular Approach[18]: Fix 'base' distribution $\mathbb{P}_\xi$ and learn diffeomorphisms $(f_x)_{x \in \mathbb{X}}$

$$y(x) = f_x^{(L)} \circ f_x^{(L-1)} \circ \cdots \circ f_x^{(1)}(\xi) \implies \log p(y|x) = \log p_\xi(f_x^{-1}(y)) + \log |\nabla_y f_x^{-1}(y)|$$

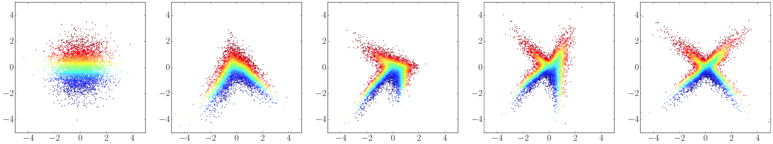


[18] Khemekhem et al. AISTATS'21, Nasr-Esfahany et al. ICML'22, Javaloy et al. NeurIPS'24, Zhou et al. AAAI'25

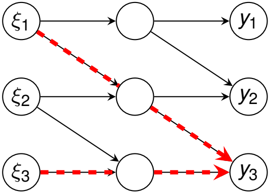
How to model BCMs?

Popular Approach[18]: Fix 'base' distribution $\mathbb{P}_\xi$ and learn diffeomorphisms $(f_x)_{x \in \mathbb{X}}$

$$y(x) = f_x^{(L)} \circ f_x^{(L-1)} \circ \cdots \circ f_x^{(1)}(\xi) \implies \log p(y|x) = \log p_\xi(f_x^{-1}(y)) + \log |\nabla_y f_x^{-1}(y)|$$



Autoregressive Normalizing Flows: Respect causal ordering when stacked!

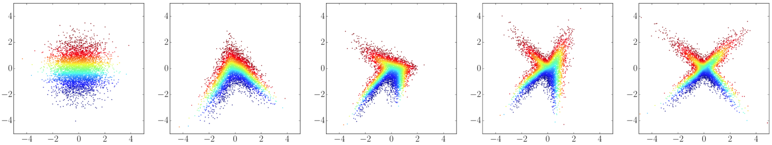


[18] Khemekhem et al. AISTATS'21, Nasr-Esfahany et al. ICML'22, Javaloy et al. NeurIPS'24, Zhou et al. AAAI'25

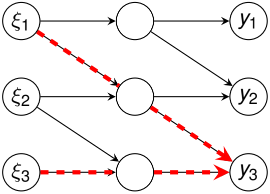
How to model BCMs?

Popular Approach[18]: Fix 'base' distribution $\mathbb{P}_\xi$ and learn diffeomorphisms $(f_x)_{x\in\mathbb{X}}$

$$y(x) = f_x^{(L)} \circ f_x^{(L-1)} \circ \cdots \circ f_x^{(1)}(\xi) \implies \log p(y|x) = \log p_\xi(f_x^{-1}(y)) + \log |\nabla_y f_x^{-1}(y)|$$



Autoregressive Normalizing Flows: Respect causal ordering when stacked!



Model	**Autoregressive transform** $f_j(v_{<j}, \xi_j)$
NICE (Additve)	$\xi_j + \mu_j(v_{<j})$
MAF (Affine)	$\xi_j \mapsto \exp\big(\lambda_j(v_{<j})\big)\,\xi_j + \mu_j(v_{<j})$
NAF (INN)	$\xi_j \mapsto \sigma^{-1}(w(v_{<j}) \cdot \sigma(\sigma_j(v_{<j})\xi_j + \mu_j(v_{<j})))$
NSF (Spline)	$\xi_j \mapsto v_j 1_{v_j \notin [-B,B]} + M_j(\xi_j; v_{<j}) 1_{v_j \in [-B,B]}$

[18] Khemekhem et al. AISTATS'21, Nasr-Esfahany et al. ICML'22, Javaloy et al. NeurIPS'24, Zhou et al. AAAI'25

Tail Mis-Specification Problems

- If tail decay of $\mathbb{P}_\xi$ doesn't match $\mathbb{P}_{Y|X=x}$, no bi-lipschitz f_x exists![19]
- Heavy tailed $\mathbb{P}_{Y|X=x}$, Gaussian $\hat{\mathbb{P}}_\xi$ = undefined likelihood:

$$|\mathbb{E} \log \hat{p}(Y(x))| \succeq \mathbb{E}\|f_x^{-1}(Y(x))\|^2 = \infty$$

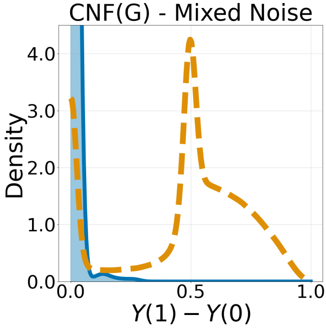
[19] Jaini et al. ICML'20, Liang et al. ICML'22

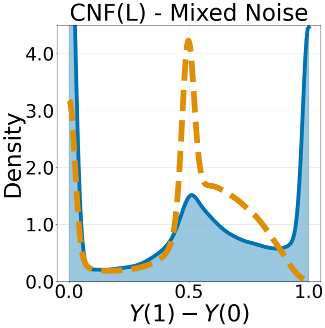
Tail Mis-Specification Problems

- If tail decay of $\mathbb{P}_\xi$ doesn't match $\mathbb{P}_{Y|X=x}$, no bi-lipschitz f_x exists![19]
- Heavy tailed $\mathbb{P}_{Y|X=x}$, Gaussian $\hat{\mathbb{P}}_\xi$ = undefined likelihood:

$$|\mathbb{E} \log \hat{p}(Y(x))| \succeq \mathbb{E}\|f_x^{-1}(Y(x))\|^2 = \infty$$

Example : $\quad Y = (X+1)\xi\ , \quad X \sim Bern(1/2), \quad \xi \sim \frac{1}{2}|\mathcal{N}(0,1)| - \frac{1}{2}|NBP(0.1, 0.1)$





[19] Jaini et al. ICML'20, Liang et al. ICML'22

Support Mis-Specification Problems

- Need diffeomorphic $\mathrm{supp}(\hat{\mathbb{P}}_\xi)$ and $\mathrm{supp}(\mathbb{P}_{Y|X=x})$ for existence of $f_x \in \mathcal{F}$
- Likelihood non-identifiability if $\mathrm{supp}\big((f_x^{-1})_\#(\mathbb{P}_{Y|X=x})\big) \not\supset \mathrm{supp}(\hat{\mathbb{P}}_\xi)$:

Support Mis-Specification Problems

- Need diffeomorphic supp$(\hat{\mathbb{P}}_\xi)$ and supp$(\mathbb{P}_{Y|X=x})$ for existence of $f_x \in \mathcal{F}$
- Likelihood non-identifiability if supp$\big((f_x^{-1})_\#(\mathbb{P}_{Y|X=x})\big) \not\supset$ supp$(\hat{\mathbb{P}}_\xi)$:

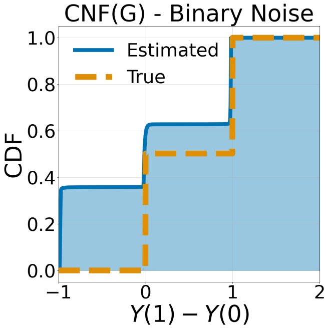
$$\mathbb{E}\log \hat{p}_f(Y(x)) = \int \big(\log \hat{p}_\xi(f_x^{-1}(y)) + \log|\nabla_y f_x^{-1}(y)|\big)\mathbb{P}_{Y|X=x}(dy)$$

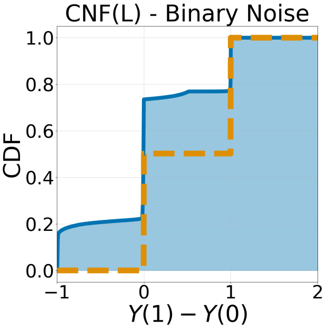
Support Mis-Specification Problems

- Need diffeomorphic supp$(\hat{\mathbb{P}}_\xi)$ and supp$(\mathbb{P}_{Y|X=x})$ for existence of $f_x \in \mathcal{F}$
- Likelihood non-identifiability if $\mathrm{supp}\big((f_x^{-1})_\#(\mathbb{P}_{Y|X=x})\big) \not\supset \mathrm{supp}(\hat{\mathbb{P}}_\xi)$:

$$\mathbb{E}\log\hat{p}_f(Y(x)) = \int \big(\log\hat{p}_\xi(f_x^{-1}(y)) + \log|\nabla_y f_x^{-1}(y)|\big)\mathbb{P}_{Y|X=x}(dy)$$

Example : $\quad Y = (X+1)\xi\ , \quad X \sim Bern(1/2), \quad \xi \sim Rad(1/2)$



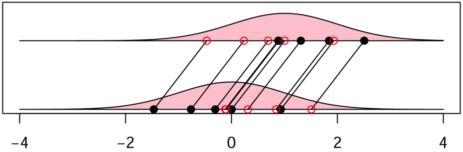


Idea: Choose coupling between $Y(1)$, $Y(0)$ using *optimal-transport*:

$$\pi^* = \arg \min_{\pi \in \Pi(\mathbb{P}_{Y(1)}, \mathbb{P}_{Y(0)})} \iint c(y(1), y(0)) d\pi(y(1), y(0))$$

Motivations:

- Conservativism[1]
- Counterfactual Similarity[2]
- Optimal Matching[3]



[1] Balakrishnan et al. (2025) "Conservative inference for counterfactuals." Journal of Causal Inference.

[2] De Lara et al. (2024) "Transport-based Counterfactual Models" JMLR.

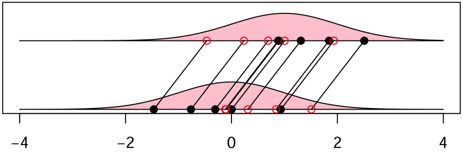
[3] Charpentier et al (2023) "Optimal transport for counterfactual estimation: A method for causal inference.", Optimal transport statistics for economics and related topics.

Idea: Choose coupling between $Y(1)$, $Y(0)$ using *optimal-transport*:

$$\pi^* = \arg\min_{\pi \in \Pi(\mathbb{P}_{Y(1)}, \mathbb{P}_{Y(0)})} \iint c(y(1), y(0)) d\pi(y(1), y(0))$$

Motivations:

- Conservativism[1]
- Counterfactual Similarity[2]
- Optimal Matching[3]



Deterministic coupling restriction: $Y(1) = T_{1,0}(Y(0))$

[1] Balakrishnan et al. (2025) "Conservative inference for counterfactuals." Journal of Causal Inference.

[2] De Lara et al. (2024) "Transport-based Counterfactual Models" JMLR.

[3] Charpentier et al (2023) "Optimal transport for counterfactual estimation: A method for causal inference.", Optimal transport statistics for economics and related topics.

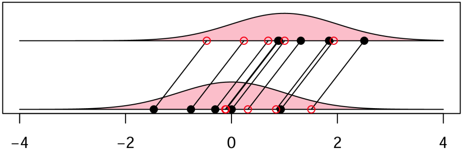
Optimal Transport Methods: An Alternative

Idea: Choose coupling between $Y(1)$, $Y(0)$ using *optimal-transport*:

$$\pi^* = \arg \min_{\pi \in \Pi(\mathbb{P}_{Y(1)}, \mathbb{P}_{Y(0)})} \iint c(y(1), y(0)) d\pi(y(1), y(0))$$

Motivations:

- Conservativism[1]
- Counterfactual Similarity[2]
- Optimal Matching[3]



Deterministic coupling restriction: $Y(1) = T_{1,0}(Y(0))$

Brenier Maps $T^*_{1,0} = \arg\min_{T_\# \mathbb{P}_{Y(0)} = \mathbb{P}_{Y(1)}} \int \|y(0) - T_{1,0}(y(0))\|^2 d\mathbb{P}_{Y(0)}$

[1] Balakrishnan et al. (2025) "Conservative inference for counterfactuals." Journal of Causal Inference.

[2] De Lara et al. (2024) "Transport-based Counterfactual Models" JMLR.

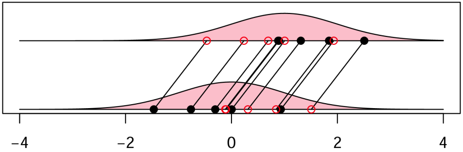
[3] Charpentier et al (2023) "Optimal transport for counterfactual estimation: A method for causal inference.", Optimal transport statistics for economics and related topics.

Idea: Choose coupling between $Y(1)$, $Y(0)$ using *optimal-transport*:

$$\pi^* = \arg \min_{\pi \in \Pi(\mathbb{P}_{Y(1)}, \mathbb{P}_{Y(0)})} \iint c(y(1), y(0)) d\pi(y(1), y(0))$$

Motivations:

- Conservativism[1]
- Counterfactual Similarity[2]
- Optimal Matching[3]



Deterministic coupling restriction: $Y(1) = T_{1,0}(Y(0))$

Brenier Maps $T^*_{1,0} = \arg\min_{T_{\#}\mathbb{P}_{Y(0)} = \mathbb{P}_{Y(1)}} \int \|y(0) - T_{1,0}(y(0))\|^2 d\mathbb{P}_{Y(0)}$

Guarantees identifiability of $T_{1,0}$ between abs. continuous distributions!

[1] Balakrishnan et al. (2025) "Conservative inference for counterfactuals." Journal of Causal Inference.

[2] De Lara et al. (2024) "Transport-based Counterfactual Models" JMLR.

[3] Charpentier et al (2023) "Optimal transport for counterfactual estimation: A method for causal inference.", Optimal transport statistics for economics and related topics.

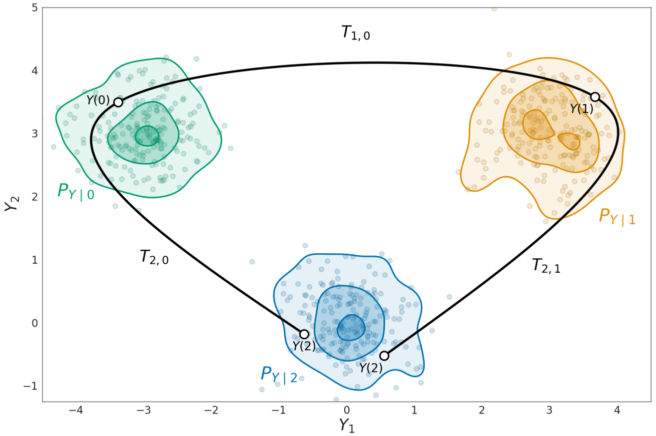
Incoherence of Transports Under Multiple treatments

Issue: If $x \in \{0, 1, 2\}$ and Y multivariate, OT maps not closed under composition!

$$T_{2,1} \circ T_{1,0} \neq T_{2,0}$$

Logical Impossibility:

$$Y(2) = T_{2,1} \circ T_{1,0}(Y(0)) \neq T_{2,0}(Y(0)) = T_{2,0} \circ T_{0,2}(Y(2)) = Y(2)$$



Example: Gaussian Transport

Three Multivariate Gaussians:

$$\mathbb{P}_0 = \mathcal{N}(\mu_0, \Sigma_0), \quad \mathbb{P}_1 = \mathcal{N}(\mu_1, \Sigma_1), \quad \mathbb{P}_2 = \mathcal{N}(\mu_2, \Sigma_2)$$

Brenier (OT) Map:

$$T_{x,x'}(y) = \mu_{x'} + \Sigma_x^{-1/2}\left(\Sigma_x^{1/2}\Sigma_{x'}\Sigma_x^{1/2}\right)^{1/2}\Sigma_x^{-1/2}(y - \mu_x)$$

Three Multivariate Gaussians:

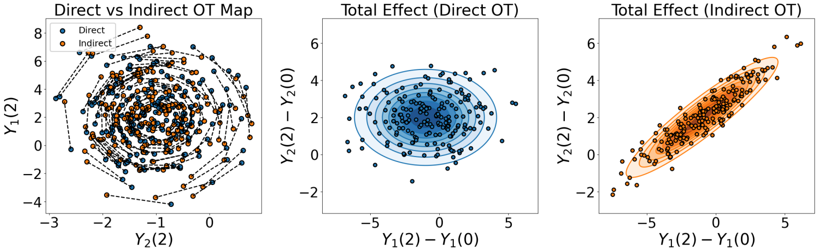
$$\mathbb{P}_0 = \mathcal{N}(\mu_0, \Sigma_0), \quad \mathbb{P}_1 = \mathcal{N}(\mu_1, \Sigma_1), \quad \mathbb{P}_2 = \mathcal{N}(\mu_2, \Sigma_2)$$

Brenier (OT) Map:

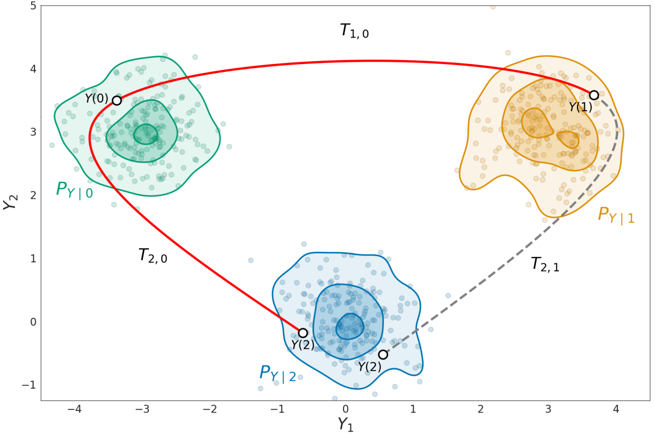
$$T_{x,x'}(y) = \mu_{x'} + \Sigma_x^{-1/2}\left(\Sigma_x^{1/2}\Sigma_{x'}\Sigma_x^{1/2}\right)^{1/2}\Sigma_x^{-1/2}(y - \mu_x)$$

Illustration: Draw $Y(0) \sim \mathbb{P}_0$ and impute $Y(1), Y(2)$

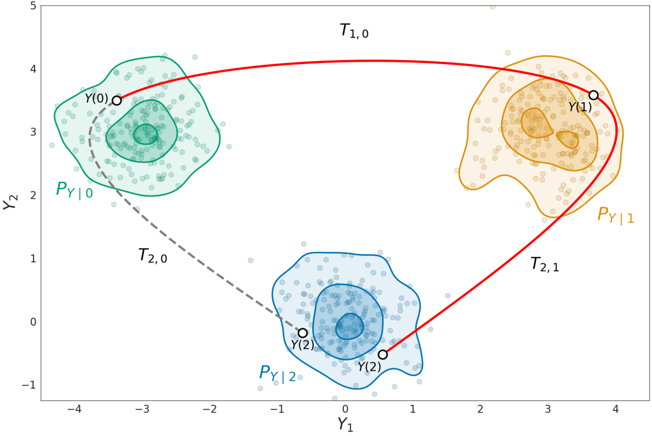
- **Direct:** $Y(1) = T_{1,0}(Y(0))$, $\quad Y(2) = T_{2,0}(Y(0))$
- **Indirect:** $Y(1) = T_{1,0}(Y(0))$, $\quad Y(2) = T_{2,1} \circ T_{1,0}(Y(0))$



Selecting Transport Subsets Induces Coupling Non-Identifiability



Selecting Transport Subsets Induces Coupling Non-Identifiability



Transport-based Models with Cocycles

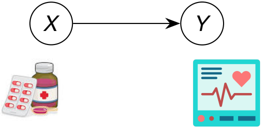
Goals

Modeling and Estimation Framework for counterfactual couplings that:

- Guarantees Coherence
- Avoids Latent Noise Assumptions

Formal Set-up

Treatment $X \in \mathbb{R}$, Outcomes $Y := (Y_1, \dots, Y_p) \in \mathbb{R}^p$



Counterfactuals: There exist 'potential outcomes' $\{Y(x)\}_{x \in \mathbb{X}}$ satisfying

- Consistency: $X = x \implies Y(x) = Y$
- Exchangeability: $Y(x) \perp\!\!\!\perp X$

Formal Set-up

Treatment $X \in \mathbb{R}$, Outcomes $Y := (Y_1, \ldots, Y_p) \in \mathbb{R}^p$

$$\begin{tikzcd} X \arrow[r] & Y \end{tikzcd}$$

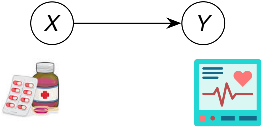
Counterfactuals: There exist 'potential outcomes' $\{Y(x)\}_{x \in \mathbb{X}}$ satisfying

- Consistency: $X = x \implies Y(x) = Y$
- Exchangeability: $Y(x) \perp\!\!\!\perp X$

$$\implies Y(x) \sim \mathbb{P}_{Y|X=x}$$

Formal Set-up

Treatment $X \in \mathbb{R}$, Outcomes $Y := (Y_1, \ldots, Y_p) \in \mathbb{R}^p$



Counterfactuals: There exist 'potential outcomes' $\{Y(x)\}_{x \in \mathbb{X}}$ satisfying

- Consistency: $X = x \implies Y(x) = Y$
- Exchangeability: $Y(x) \perp\!\!\!\perp X$

$$\implies Y(x) \sim \mathbb{P}_{Y|X=x}$$

Transport-based model:

$$Y(x) = T_{x,x'}(Y(x')), \quad \forall x, x' \in \mathbb{X} \tag{CC}$$

Necessary Algebraic Properties

$$Y(x) = T_{x,x'}(Y(x')), \quad \forall x, x' \in \mathbb{X} \tag{CC}$$

$$Y(x) = T_{x,x'}(Y(x')), \quad \forall x, x' \in \mathbb{X} \tag{CC}$$

1. **Distribution Adaptedness**: For each $x, x' \in \mathbb{X}$,

$$(T_{x',x})_{\#}\mathbb{P}_{Y|X=x} = \mathbb{P}_{Y|X=x'} \tag{DA}$$

2. **Identity**: For each $x \in \mathbb{X}$:

$$T_{x,x}(y) = y, \quad \forall y \in \mathbb{Y}_x \tag{ID}$$

$$Y(x) = T_{x,x'}(Y(x')), \quad \forall x, x' \in \mathbb{X} \tag{CC}$$

1. **Distribution Adaptedness**: For each $x, x' \in \mathbb{X}$,

$$(T_{x',x})_{\#}\mathbb{P}_{Y|X=x} = \mathbb{P}_{Y|X=x'} \tag{DA}$$

2. **Identity**: For each $x \in \mathbb{X}$:

$$T_{x,x}(y) = y, \quad \forall y \in \mathbb{Y}_x \tag{ID}$$

3. **Path Independence**: For each $x, x', x'' \in \mathbb{X}$:

$$T_{x'',x'} \circ T_{x',x}(y) = T_{x'',x}(y), \quad \forall y \in \mathbb{Y}_x \tag{PI}$$

$$Y(x) = T_{x,x'}(Y(x')), \quad \forall x, x' \in \mathbb{X} \tag{CC}$$

1. **Distribution Adaptedness**: For each $x, x' \in \mathbb{X}$,

$$(T_{x',x})_{\#}\mathbb{P}_{Y|X=x} = \mathbb{P}_{Y|X=x'} \tag{DA}$$

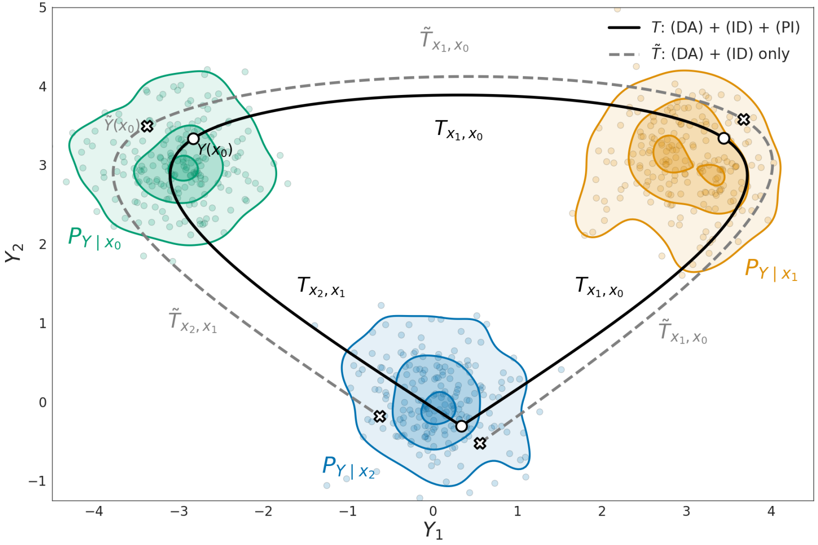
2. **Identity**: For each $x \in \mathbb{X}$:

$$T_{x,x}(y) = y, \quad \forall y \in \mathbb{Y}_x \tag{ID}$$

3. **Path Independence**: For each $x, x', x'' \in \mathbb{X}$:

$$T_{x'',x'} \circ T_{x',x}(y) = T_{x'',x}(y), \quad \forall y \in \mathbb{Y}_x \tag{PI}$$

ID + PI = properties of a *cocycle!*



T: (DA) + (ID) + (PI)
$\tilde{T}$: (DA) + (ID) only
$\tilde{T}_{x_1,x_0}$
T_{x_1,x_0}
$\tilde{Y}(x_0)$
$Y(x_0)$
$P_{Y\,|\,x_0}$
$P_{Y\,|\,x_1}$
$P_{Y\,|\,x_2}$
T_{x_2,x_1}
$\tilde{T}_{x_2,x_1}$
T_{x_1,x_0}
$\tilde{T}_{x_1,x_0}$
Y_2
Y_1
5
4
3
2
1
0
−1
−4
−3
−2
−1
0
1
2
3
4

Sufficiency of Cocycle Properties for Admissible Transports

(DA) + (ID) + (PI) guarantee existence of *some* counterfactuals $\{\tilde{Y}(x)\}_{x \in \mathbb{X}}$:

$$\widetilde{Y}(x) =_{\text{a.s.}} T_{x,x'}\big(\widetilde{Y}(x')\big) \text{ for every } x, x' \in \mathbb{X}$$

(DA) + (ID) + (PI) guarantee existence of *some* counterfactuals $\{\tilde{Y}(x)\}_{x\in\mathbb{X}}$:

$$\widetilde{Y}(x) =_{\text{a.s.}} T_{x,x'}\left(\widetilde{Y}(x')\right) \text{ for every } x, x' \in \mathbb{X}$$

...how to enforce (ID) + (PI) + (DA) in practice?

Theorem 1 (Cocycle Factorization)

Every cocycle T satisfying (ID), (PI)*, and* (DA) *w.r.t.* $\mathbb{P}_{Y|X}$ *can be represented as:*

$$T_{x,x'} = f_x \circ f_{x'}^{+}, \quad \mathbb{P}_{Y|X=x'}\text{-a.s.}$$

Where $f_x : \mathbb{Y}_0 \to \mathbb{Y}$ *is injective with left-inverse* f_x^{+}

Theorem 1 (Cocycle Factorization)

Every cocycle T satisfying (ID), (PI)*, and* (DA) *w.r.t.* $\mathbb{P}_{Y|X}$ *can be represented as:*

$$T_{x,x'} = f_x \circ f_{x'}^{+}, \quad \mathbb{P}_{Y|X=x'}\text{-a.s.}$$

Where $f_x : \mathbb{Y}_0 \to \mathbb{Y}$ *is injective with left-inverse* f_x^{+}

Implication: Can construct classes of cocycles via parameterized bijections!

$$\mathcal{F} \subseteq \left\{ f : \mathbb{X} \times \mathbb{Y} \to \mathbb{Y} \;\middle|\; f_x := f(x, \bullet) \text{ bijective } \forall x \in \mathbb{X} \right\}$$

Examples: Normalizing flows, invertible NNs...

Structure of Counterfactual Cocycles

Theorem 1 (Cocycle Factorization)

Every cocycle T satisfying (ID), (PI)*, and* (DA) *w.r.t.* $\mathbb{P}_{Y|X}$ *can be represented as:*

$$T_{x,x'} = f_x \circ f_{x'}^{+}, \quad \mathbb{P}_{Y|X=x'}\text{-a.s.}$$

Where $f_x : \mathbb{Y}_0 \to \mathbb{Y}$ *is injective with left-inverse* f_x^{+}

Implication: Can construct classes of cocycles via parameterized bijections!

$$\mathcal{F} \subseteq \left\{ f : \mathbb{X} \times \mathbb{Y} \to \mathbb{Y} \;\middle|\; f_x := f(x, \bullet) \text{ bijective } \forall x \in \mathbb{X} \right\}$$

Examples: Normalizing flows, invertible NNs...

Wait... isn't this just counterfactuals in a bijective SCM?

$$Y(x) = f_x(\xi) \implies Y(x) = f_x \circ f_{x'}^{-1}(Y(x'))$$

Theorem 2 (Cocycle Equivalence to Structural Model)

$\{Y(x)\}_{x\in\mathbb{X}}$ *satisfies Exchangeability, Consistency and*

$$Y(x) = T_{x,x'}(Y(x')) \quad \forall x, x' \in \mathbb{X}$$

if and only if $Y = f(X,\xi), \quad \xi \perp\!\!\!\perp X$.

Theorem 2 (Cocycle Equivalence to Structural Model)

$\{Y(x)\}_{x \in \mathbb{X}}$ *satisfies Exchangeability, Consistency and*

$$Y(x) = T_{x,x'}(Y(x')) \quad \forall x, x' \in \mathbb{X}$$

if and only if $Y = f(X, \xi), \quad \xi \perp\!\!\!\perp X$.

Proof Sketch ($\Longrightarrow$)

$$Y(x) = T_{x,x_0}(Y(x_0)) = f_x(Y(x_0)) =: f_x(\xi)$$

Theorem 2 (Cocycle Equivalence to Structural Model)

$\{Y(x)\}_{x \in \mathbb{X}}$ *satisfies Exchangeability, Consistency and*

$$Y(x) = T_{x,x'}(Y(x')) \quad \forall x, x' \in \mathbb{X}$$

if and only if $Y = f(X, \xi), \quad \xi \perp\!\!\!\perp X$.

Proof Sketch ($\Longrightarrow$)

$$Y(x) = T_{x,x_0}(Y(x_0)) = f_x(Y(x_0)) =: f_x(\xi)$$

Theorem 2 (Cocycle Equivalence to Structural Model)

$\{Y(x)\}_{x\in\mathbb{X}}$ *satisfies Exchangeability, Consistency and*

$$Y(x) = T_{x,x'}(Y(x')) \quad \forall x, x' \in \mathbb{X}$$

if and only if $Y = f(X,\xi), \quad \xi \perp\!\!\!\perp X$.

Proof Sketch ($\implies$)

$$Y(x) = T_{x,x_0}(Y(x_0)) = f_x(Y(x_0)) =: f_x(\xi)$$

Implication:

- (injective) SCMs characterize space of valid counterfactual transport models
- Under TMI restriction on f_x, we recover Acyclic BCM $Y_j = f_j(Y_{<j}, X, \xi_j)$ again!

Our Idea: Center everything around the cocycle

Aim: Do all estimation and inference without ever referencing or specifying $\mathbb{P}_\xi$

Our Idea: Center everything around the cocycle

Aim: Do all estimation and inference without ever referencing or specifying $\mathbb{P}_\xi$

1. Specify cocycle model $T^\theta_{x,x'} = f^\theta_x \circ (f^\theta_{x'})^{-1}$

2. Directly target $T_{x,x'}$ between conditionals:

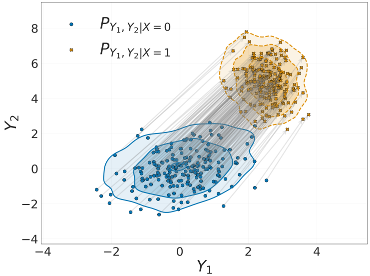
$$\ell(\theta) = \mathbb{E}_{x,x'} D(\mathbb{P}_{Y|X=x}, (T^\theta_{x,x'})_\# \mathbb{P}_{Y|X=x'})^2$$

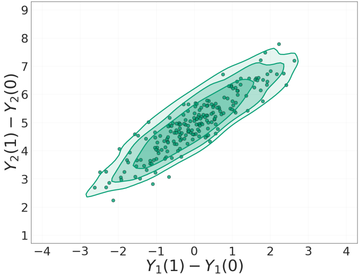
3. Use $T^{\hat{\theta}}$ to impute counterfactuals:

$$\hat{Y}^{(i)}(x) = T^{\hat{\theta}}_{x,X^{(i)}}(Y^{(i)})$$

... and empirically estimate quantities:

$$\widehat{THR} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(\hat{Y}^{(i)}(1) \prec \hat{Y}^{(i)}(0))$$





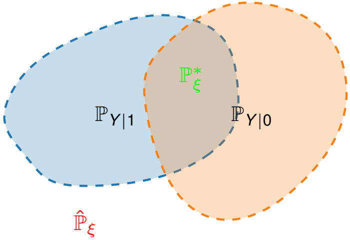
Benefits of Counterfactual Cocycle Modelling

$$(\hat{f}_x)_\# : \hat{\mathbb{P}}_\xi \mapsto \mathbb{P}_{Y|X=x}, \qquad (T_{x,x'})_\# = (f_x \circ f_{x'}^{+})_\# : \mathbb{P}_{Y|X=x'} \mapsto \mathbb{P}_{Y|X=x}$$

$$(\hat{f}_x)_\# : \hat{\mathbb{P}}_\xi \mapsto \mathbb{P}_{Y|X=x}, \qquad (T_{x,x'})_\# = (f_x \circ f_{x'}^+)_\# : \mathbb{P}_{Y|X=x'} \mapsto \mathbb{P}_{Y|X=x}$$

RCT case: $\boldsymbol{x} \in \{0,1\}$

- Consider class of functions $\mathcal{F}$ for modelling $(f_x)_{x\in\{0,1\}}$
- $\mathcal{P}_1(\mathcal{F}), \mathcal{P}_0(\mathcal{F})$ = distributions reachable by pushing $\mathbb{P}_{Y|1}, \mathbb{P}_{Y|0}$ through $f^{-1} \in \mathcal{F}^{-1}$

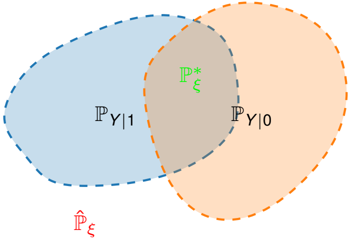


Noise Invariance and Model Mis-Specification

$$(\hat{f}_x)_\# : \hat{\mathbb{P}}_\xi \mapsto \mathbb{P}_{Y|X=x}, \qquad (T_{x,x'})_\# = (f_x \circ f^+_{x'})_\# : \mathbb{P}_{Y|X=x'} \mapsto \mathbb{P}_{Y|X=x}$$

RCT case: $x \in \{0,1\}$

- Consider class of functions $\mathcal{F}$ for modelling $(f_x)_{x\in\{0,1\}}$
- $\mathcal{P}_1(\mathcal{F}), \mathcal{P}_0(\mathcal{F})$ = distributions reachable by pushing $\mathbb{P}_{Y|1}, \mathbb{P}_{Y|0}$ through $f^{-1} \in \mathcal{F}^{-1}$



- $(\mathcal{F}, \hat{\mathbb{P}}_\xi)$ well-specified for $(f, \mathbb{P}_\xi)$ if

  $\hat{\mathbb{P}}_\xi \in$ intersection

- $\mathcal{F}$ well specified for T if

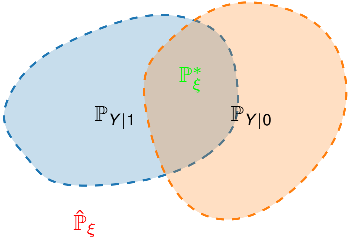
  $\exists\ \mathbb{P}^*_\xi \in$ intersection

Noise Invariance and Model Mis-Specification

$$(\hat{f}_x)_\# : \hat{\mathbb{P}}_\xi \mapsto \mathbb{P}_{Y|X=x}, \qquad (T_{x,x'})_\# = (f_x \circ f_{x'}^+)_\# : \mathbb{P}_{Y|X=x'} \mapsto \mathbb{P}_{Y|X=x}$$

RCT case: $x \in \{0,1\}$

- Consider class of functions $\mathcal{F}$ for modelling $(f_x)_{x\in\{0,1\}}$
- $\mathcal{P}_1(\mathcal{F}), \mathcal{P}_0(\mathcal{F})$ = distributions reachable by pushing $\mathbb{P}_{Y|1}, \mathbb{P}_{Y|0}$ through $f^{-1} \in \mathcal{F}^{-1}$



- $(\mathcal{F}, \hat{\mathbb{P}}_\xi)$ well-specified for $(f, \mathbb{P}_\xi)$ if

  $\hat{\mathbb{P}}_\xi \in$ intersection

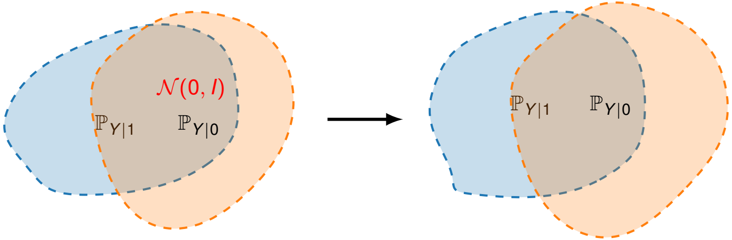
- $\mathcal{F}$ well specified for T if

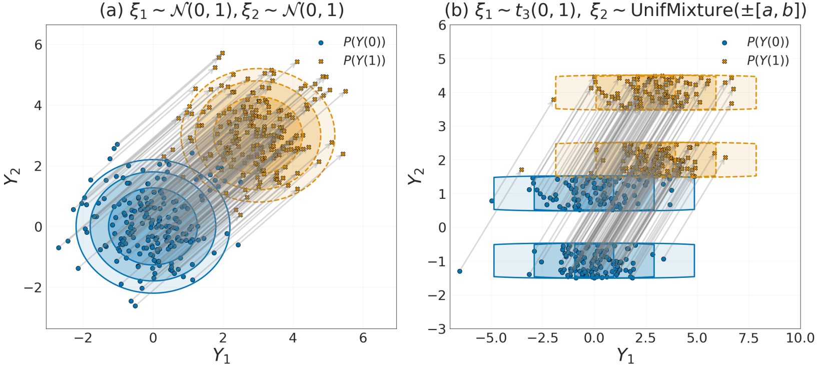
  $\exists\ \mathbb{P}_\xi^* \in$ intersection

Invariance: Changing true $\mathbb{P}_\xi$ doesn't change whether the intersection is empty!

$$\tilde{Y}(x) = f_x(\tilde{\xi}) \implies \tilde{Y}(x) = f_x \circ f_{x'}^+(\tilde{Y}(x'))$$

Invariance Example





Noise Invariance and Minimal Complexity

Can reparameterize SCM using any bijection $g \in \mathrm{Aut}(\mathcal{E})$:

$$Y = f(X, \xi) = f(X, g \circ g^{-1}(\xi)) = f^{(g)}(X, \xi^{(g)})$$

Noise Invariance and Minimal Complexity

Can reparameterize SCM using any bijection $g \in \mathrm{Aut}(\mathcal{E})$:

$$Y = f(X, \xi) = f(X, g \circ g^{-1}(\xi)) = f^{(g)}(X, \xi^{(g)})$$

Cocycle Invariant to Reparameterization:

$$T^{(g)}_{x,x'} = f^{(g)}_x \circ f^{(g)}_{x'} = f_x \circ \cancel{g \circ g^{-1}} \circ f^{+}_{x'} = T_{x,x'}$$

Noise Invariance and Minimal Complexity

Can reparameterize SCM using any bijection $g \in \mathrm{Aut}(\mathcal{E})$:

$$Y = f(X, \xi) = f(X, g \circ g^{-1}(\xi)) = f^{(g)}(X, \xi^{(g)})$$

Cocycle Invariant to Reparameterization:

$$T^{(g)}_{x,x'} = f^{(g)}_x \circ f^{(g)}_{x'} = f_x \circ \cancel{g \circ g^{-1}} \circ f^{+}_{x'} = T_{x,x'}$$

Implication: Can use *any* member of $f^{(g)}$ to construct cocycle!

$\Longrightarrow$ Just need "simplest representative" $f^{\star} \in (f^{(g)})_{g \in \mathrm{Aut}(\mathcal{E})}$ to lie in model class $\mathcal{F}$

Noise Invariance and Minimal Complexity

Can reparameterize SCM using any bijection $g \in \mathrm{Aut}(\mathcal{E})$:

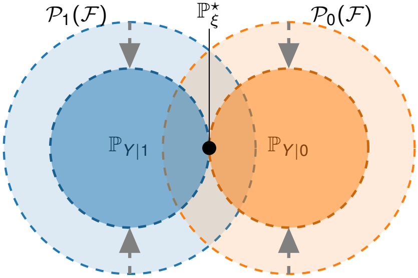
$$Y = f(X, \xi) = f(X, g \circ g^{-1}(\xi)) = f^{(g)}(X, \xi^{(g)})$$

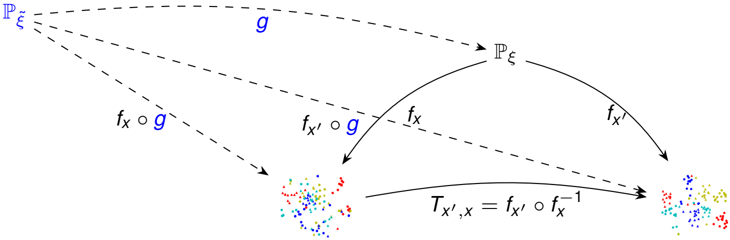
Cocycle Invariant to Reparameterization:

$$T^{(g)}_{x,x'} = f^{(g)}_x \circ f^{(g)}_{x'} = f_x \circ \cancel{g \circ g^{-1}} \circ f^{+}_{x'} = T_{x,x'}$$

Implication: Can use *any* member of $f^{(g)}$ to construct cocycle!

$\Longrightarrow$ Just need "simplest representative" $f^{\star} \in (f^{(g)})_{g \in \mathrm{Aut}(\mathcal{E})}$ to lie in model class $\mathcal{F}$





Example: let $\mathbb{P}_0 = \mathcal{N}(0,1)$, $\mathbb{P}_{Y|X} = \text{Cauchy}(\beta x, \sigma)$, then:

$$f_x \circ g(\xi) = \underbrace{x}_{f_x} + \underbrace{\sigma \tan\left[\frac{\pi}{2}\text{erf}\left(\frac{\xi}{\sqrt{2}}\right)\right]}_{g(u)} \qquad \text{...but } T_{x,x'}(y) = x - x' + y$$

So, need "complicated" $\hat{f}_x$ from $\mathcal{N}(0,1)$, but cocycle can be modelled via linear class

$$\{f_x^{\beta}(y) = \beta x + y \mid \beta \in \mathbb{R}\}$$

What is the optimal base distribution?

Define smallest group of transformations containing (f_0, f_1): $\mathbb{G}_f := \langle f_0, f_1 \rangle$

What is the optimal base distribution?

Define smallest group of transformations containing (f_0, f_1): $\mathbb{G}_f := \langle f_0, f_1 \rangle$

Size of $\mathbb{G}_f$: can depend on *parameterization* $(f_1^{(g)}, f_0^{(g)}) = (f^{(1)} \circ g, f^{(0)} \circ g)$

What is the optimal base distribution?

Define smallest group of transformations containing (f_0, f_1): $\mathbb{G}_f := \langle f_0, f_1 \rangle$

Size of $\mathbb{G}_f$: can depend on *parameterization* $(f_1^{(g)}, f_0^{(g)}) = (f^{(1)} \circ g, f^{(0)} \circ g)$

Theorem 3 (Minimal Complexity Cocycle)

The parameterization $(f_1^\star, f_0^\star) = (T_{1,0}, id)$ *induces the smallest* $\mathbb{G}_f$

$\Longrightarrow$ $\mathrm{P}_{Y|0}$ is an optimal base distribution!

What is the optimal base distribution?

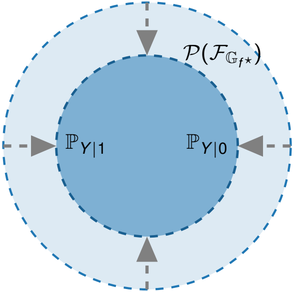
Define smallest group of transformations containing (f_0, f_1): $\mathbb{G}_f := \langle f_0, f_1 \rangle$

Size of $\mathbb{G}_f$: can depend on *parameterization* $(f_1^{(g)}, f_0^{(g)}) = (f^{(1)} \circ g, f^{(0)} \circ g)$

Theorem 3 (Minimal Complexity Cocycle)

The parameterization $(f_1^\star, f_0^\star) = (T_{1,0}, id)$ induces the smallest $\mathbb{G}_f$

$\Longrightarrow$ $\mathrm{P}_{Y|0}$ is an optimal base distribution!



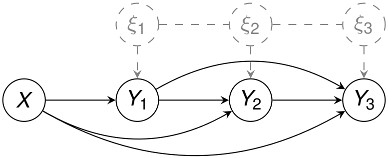
Example: $Y(x) = f_x(\xi) = x + g(\xi)$

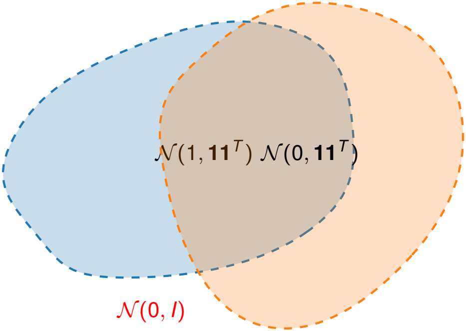
- Since $\mathbb{P}_\xi = \mathbb{P}_{Y|0}$:

$$f_x^\star(\xi) = x + \xi$$

- Any $\hat{\mathbb{P}}_\xi$ not a shift of $\mathbb{P}_{Y|0}$ will increase the size of $\mathbb{G}_f$

Counterfactual cocycles do *not* assume independent errors!





Example: Single latent cause

- $Y = \mathbf{1}X + \xi, \quad \xi \sim \mathcal{N}(0, \mathbf{1}\mathbf{1}^T)$
- If choosing $\hat{\mathbb{P}}_\xi = \mathcal{N}(0, I)$, then even

$$\mathcal{F} = \{f : \mathbb{Y} \to \mathbb{Y} \text{ is bijective}\}$$

... is mis-specified for f_0, f_1...

Cocycle Estimation

Cocycle Estimation

Disrecpancy in (DA):

$$\tilde{\ell}(T) = \mathbb{E}_{x,x' \sim P_X} D(\mathbb{P}_{Y|X=x}, (T_{x,x'})_{\#}\mathbb{P}_{Y|X=x'})^2$$

Estimating cocycles by minimising a distributional distance

Disrecpancy in (DA):

$$\tilde{\ell}(T) = \mathbb{E}_{X,X' \sim P_X} D(\mathbb{P}_{Y|X=X}, (T_{X,X'})_{\#}\mathbb{P}_{Y|X=X'})^2$$

Issues:

1. Can't generally compute D in closed form since $\mathbb{P}_{Y|X}$ is unknown
2. No obvious empirical analogue $\tilde{\ell}_n(T) \to_p \tilde{\ell}(T)$.

Estimating cocycles by minimising a distributional distance

Disrecpancy in (DA):

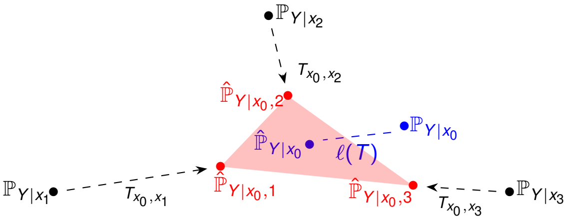
$$\tilde{\ell}(T) = \mathbb{E}_{X,X' \sim P_X} D(\mathbb{P}_{Y|X=X}, (T_{X,X'})_{\#}\mathbb{P}_{Y|X=X'})^2$$

Issues:

1. Can't generally compute D in closed form since $\mathbb{P}_{Y|X}$ is unknown
2. No obvious empirical analogue $\tilde{\ell}_n(T) \to_p \tilde{\ell}(T)$.

Idea: move expectation over X' inside the metric

$$\ell(T) = \mathbb{E}_{X \sim P_X} D(\mathbb{P}_{Y|X=X}, \mathbb{E}_{X' \sim P_X}[(T_{X,X'})_{\#}\mathbb{P}_{Y|X=X'}])^2$$



CMMD: Take $D =$ as Maximum Mean Discrepancy:

$$\ell(T) = \mathbb{E}\left\|\psi(Y) - \mathbb{E}[\psi(T_{X,X'}(Y'))|X]\right\|_{\mathcal{H}}^2 + \text{constant}$$

Empirical analogue V-statistic :

$$\ell_n(T) = \frac{1}{n}\sum_i^n \left\|\psi(Y^{(i)}) - \frac{1}{n}\sum_j^n \psi(T_{X^{(i)},X^{(j)}}(Y^{(j)}))\right\|_{\mathcal{H}}^2$$

CMMD: Take $D =$ as Maximum Mean Discrepancy:

$$\ell(T) = \mathbb{E}\left\|\psi(Y) - \mathbb{E}[\psi(T_{X,X'}(Y'))|X]\right\|_{\mathcal{H}}^2 + \text{constant}$$

Empirical analogue V-statistic :

$$\ell_n(T) = \frac{1}{n}\sum_i^n \left\|\psi(Y^{(i)}) - \frac{1}{n}\sum_j^n \psi(T_{X^{(i)},X^{(j)}}(Y^{(j)}))\right\|_{\mathcal{H}}^2$$

Gives us consistency and $\sqrt{n}$-consistency under general conditions!

In context of BCM $Y = f(X, \xi)$, consistency of $\hat{T}$ does not depend on properties of ξ!

Extension to Larger Systems and Confounding

Current RCT setting implies:

$$Y = f(X, \xi), \quad X \perp\!\!\!\perp \xi, \quad f(X, \bullet) \text{ injective}$$

Limitations

- Injectivity will break if too many independent causes $\xi_1, \ldots, \xi_m$
- Independence will break under confounding

Extension to Larger Systems and Confounding

Current RCT setting implies:

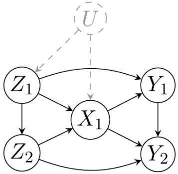
$$Y = f(X, \xi), \quad X \perp\!\!\!\perp \xi, \quad f(X, \bullet) \text{ injective}$$

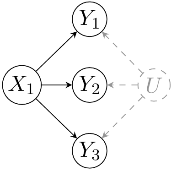
Limitations

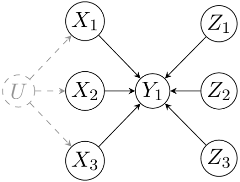
- Injectivity will break if too many independent causes $\xi_1, \ldots, \xi_m$
- Independence will break under confounding

Extension: Assume set of measured causes Z of outcomes that satisfy:

$$Z \prec X \prec Y$$







Extension to Larger Systems and Confounding

Current RCT setting implies:

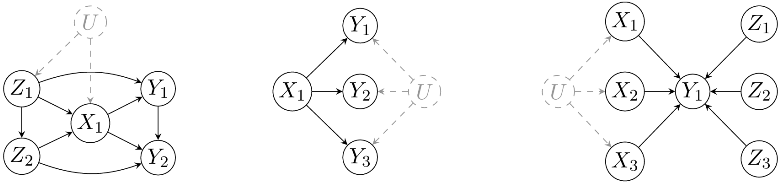
$$Y = f(X, \xi), \quad X \perp\!\!\!\perp \xi, \quad f(X, \bullet) \text{ injective}$$

Limitations

- Injectivity will break if too many independent causes $\xi_1, \ldots, \xi_m$
- Independence will break under confounding

Extension: Assume set of measured causes Z of outcomes that satisfy:

$$Z \prec X \prec Y$$

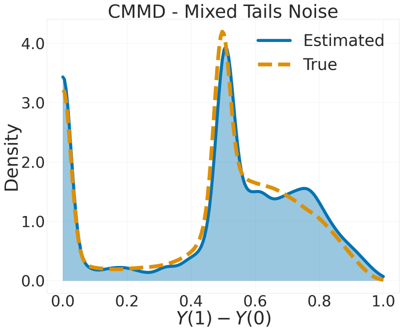


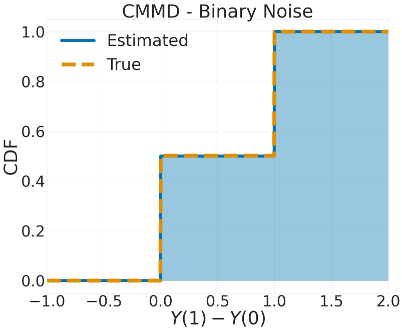
Idea: Do everything on counterfactuals $\{Y(x, z)\}_{(x,z) \in \mathbb{X} \times \mathbb{Z}}$ that relate to $Y(x)$ via nested consistency property: $Y(x) := Y(x, Z)$.

$$Y(x) = T_{(x,Z),(x',Z)}(Y(x'))$$

Experiments

Demonstration on the Toy Example





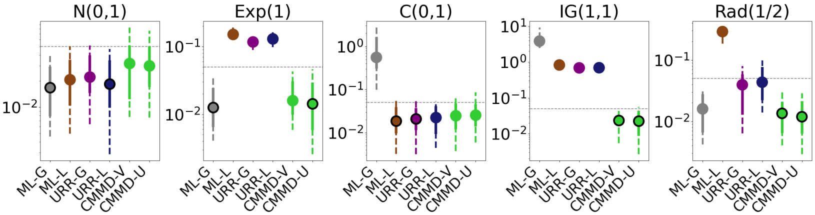
Noise Ablation vs flow-based SCMs

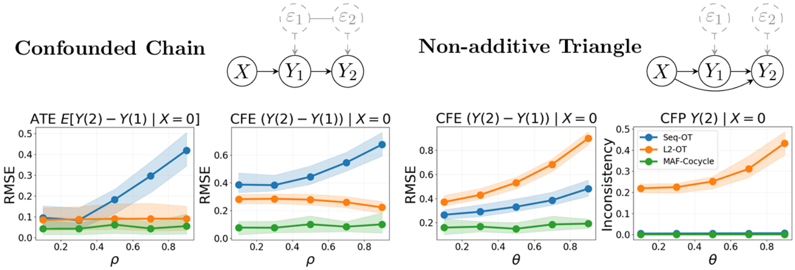
$$Y = X + \xi$$

Method	$\mathbb{P}_\xi = \mathsf{N}(0,1)$	$\mathbb{P}_\xi = \mathsf{Ga}(1,1)$	$\mathbb{P}_\xi = \mathsf{t}_1(0,1)$	$\mathbb{P}_\xi = \mathsf{IG}(1,1)$	$\mathbb{P}_\xi = \mathsf{Rad}(1/2)$
			Interventional KS		
ML-G	$\mathbf{0.015 \pm 0.007}$	0.081 ± 0.041	0.121 ± 0.079	0.208 ± 0.162	0.405 ± 0.069
ML-L	0.055 ± 0.010	0.074 ± 0.036	0.110 ± 0.069	0.120 ± 0.090	0.415 ± 0.067
ML-T	0.018 ± 0.008	0.079 ± 0.040	0.029 ± 0.007	0.075 ± 0.031	0.412 ± 0.062
CMMD-V	0.028 ± 0.009	$\mathbf{0.027 \pm 0.008}$	$\mathbf{0.027 \pm 0.009}$	$\mathbf{0.027 \pm 0.008}$	$\mathbf{0.271 \pm 0.031}$
CMMD-U	$\mathbf{0.010 \pm 0.004}$	$\mathbf{0.012 \pm 0.005}$	$\mathbf{0.008 \pm 0.003}$	$\mathbf{0.009 \pm 0.004}$	$\mathbf{0.268 \pm 0.008}$
			Counterfactual RMSE		
ML-G	$\mathbf{0.035 \pm 0.035}$	0.277 ± 0.118	113.667 ± 341.875	114.946 ± 147.841	0.326 ± 0.258
ML-L	0.036 ± 0.028	0.258 ± 0.073	97.745 ± 351.815	112.300 ± 165.014	0.480 ± 0.294
ML-T	$\mathbf{0.033 \pm 0.031}$	0.270 ± 0.097	0.044 ± 0.053	29.872 ± 41.628	0.391 ± 0.307
CMMD-V	$\mathbf{0.035 \pm 0.026}$	$\mathbf{0.020 \pm 0.015}$	$\mathbf{0.040 \pm 0.031}$	$\mathbf{0.028 \pm 0.024}$	$\mathbf{0.017 \pm 0.019}$
CMMD-U	0.040 ± 0.027	$\mathbf{0.022 \pm 0.016}$	$\mathbf{0.033 \pm 0.027}$	$\mathbf{0.027 \pm 0.023}$	$\mathbf{0.014 \pm 0.011}$
			True Architecture Selection %		
ML-G	96%	14%	0%	2%	2%
ML-L	**100%**	2%	4%	0%	0%
ML-T	98%	8%	94%	0%	0%
CMMD-V	**100%**	**100%**	**100%**	**100%**	**98%**
CMMD-U	**100%**	**100%**	**100%**	**100%**	**100%**

Noise Ablation vs flow-based SCMs

$$Y = X + \xi$$




Confounded Chain
Non-additive Triangle
ATE $E[Y(2)-Y(1) \mid X=0]$
CFE $(Y(2)-Y(1)) \mid X=0$
CFE $(Y(2)-Y(1)) \mid X=0$
CFP $Y(2) \mid X=0$
RMSE
Inconsistency
ρ
θ
Seq-OT
L2-OT
MAF-Cocycle

Performance on SCM Benchmarks

Table 4: Mean ± SD of KS_{int} and KS_{CF} on the *linear* SCMs.

Method	2var (lin)		triangle (lin)		fork (lin)		5chain (lin)	
	KS_{int}	KS_{CF}	KS_{int}	KS_{CF}	KS_{int}	KS_{CF}	KS_{int}	KS_{CF}
BGM	0.13 ± 0.07	0.19 ± 0.12	0.37 ± 0.05	0.06 ± 0.01	0.05 ± 0.07	0.61 ± 0.07	0.14 ± 0.05	0.06 ± 0.01
CausalNF	0.31 ± 0.11	0.24 ± 0.10	0.44 ± 0.05	0.06 ± 0.02	0.04 ± 0.02	0.66 ± 0.09	0.14 ± 0.02	0.06 ± 0.01
CAREFL	0.40 ± 0.04	0.15 ± 0.14	0.43 ± 0.05	0.06 ± 0.01	0.19 ± 0.01	0.58 ± 0.10	0.13 ± 0.02	0.06 ± 0.01
CocycleNF	$\mathbf{0.03 \pm 0.02}$	$\mathbf{0.04 \pm 0.04}$	$\mathbf{0.23 \pm 0.19}$	$\mathbf{0.02 \pm 0.01}$	$\mathbf{0.02 \pm 0.01}$	$\mathbf{0.19 \pm 0.23}$	$\mathbf{0.02 \pm 0.01}$	$\mathbf{0.03 \pm 0.01}$

Table 5: Mean ± SD of KS_{int} and KS_{CF} on the *nonlinear* SCMs.

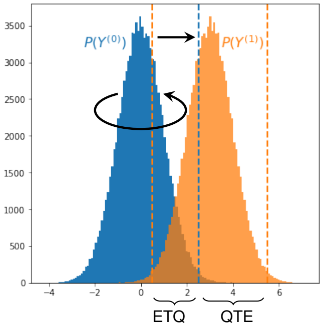
Method	2var (nonlin)		triangle (nonlin)		fork (nonlin)		5chain (nonlin)	
	KS_{int}	KS_{CF}	KS_{int}	KS_{CF}	KS_{int}	KS_{CF}	KS_{int}	KS_{CF}
BGM	0.12 ± 0.07	0.27 ± 0.13	0.41 ± 0.07	$\mathbf{0.09 \pm 0.05}$	$\mathbf{0.04 \pm 0.01}$	0.59 ± 0.08	0.07 ± 0.03	0.41 ± 0.12
CausalNF	0.30 ± 0.11	0.26 ± 0.08	0.47 ± 0.04	0.23 ± 0.08	0.07 ± 0.06	0.61 ± 0.09	0.06 ± 0.06	0.24 ± 0.16
CAREFL	0.44 ± 0.04	0.22 ± 0.15	0.47 ± 0.06	0.22 ± 0.09	0.19 ± 0.01	0.51 ± 0.21	0.17 ± 0.04	0.60 ± 0.25
CocycleNF	$\mathbf{0.04 \pm 0.02}$	$\mathbf{0.13 \pm 0.05}$	$\mathbf{0.28 \pm 0.13}$	0.20 ± 0.16	0.08 ± 0.05	$\mathbf{0.20 \pm 0.24}$	$\mathbf{0.05 \pm 0.02}$	$\mathbf{0.20 \pm 0.09}$

Counterfactual quantile effects on a real dataset

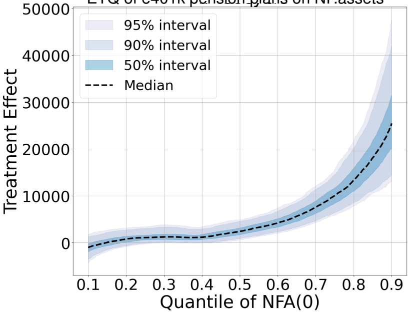
Quantile treatment effect: $\text{QTE}(\tau) = Q_{Y^{(1)}}(\tau) - Q_{Y^{(0)}}(\tau)$

Effect of Treatment on Quantile: $\text{ETQ}(\tau) = \mathbb{E}[Y^{(1)} - Y^{(0)} | Y^{(0)} = Q_{Y^{(0)}}(\tau)]$

Example difference between ETQ,QTE



ETQ of e401k pension plans on NF.assets



Summary

1. Counterfactual Cocycles as framework for Admissible counterfactual transports
2. Every counterfactual cocycle can be represented via left-invertible functions
3. Equivalent to injective SCMs, but cocycle is noise invariant + minimally complex
4. Robust cocycle estimator with consistency guarantees independent of true noise
5. Flexible parameterizations using flow-based toolkit
6. State-of-the-art performance on various benchmarks

Future Directions

1. ~~Causal discovery?~~[20]
2. More general forms of confounding?
3. Non-iid settings?
4. Stochastic cocycles?
5. Efficiency theory?

[20] Xi, J., Dance, H., Orbanz, P. & Bloem-Reddy, B. *'Distinguishing Cause From Effect with Causal Velocity Models'* (ICML25).

Thank you!

Paper Link: https://arxiv.org/abs/2405.13844

Github Repo: hwdance/Cocycles

To contact: hugh.dance.15@ucl.ac.uk, benbr@stat.ubc.ca