From Correlation to the Interventional Density

YRD 2026

Aminata Ndiaye

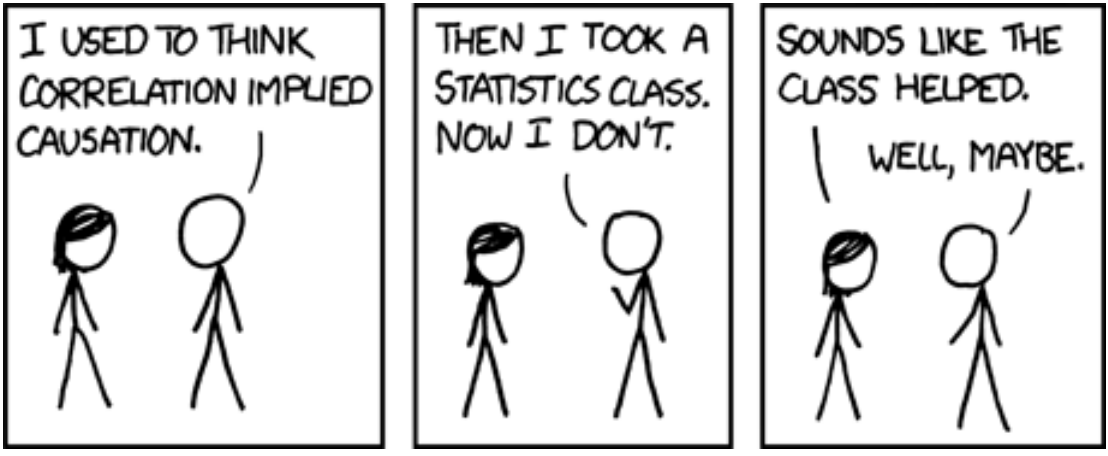
June 2026

CEREMADE– Paris Dauphine Université PSL

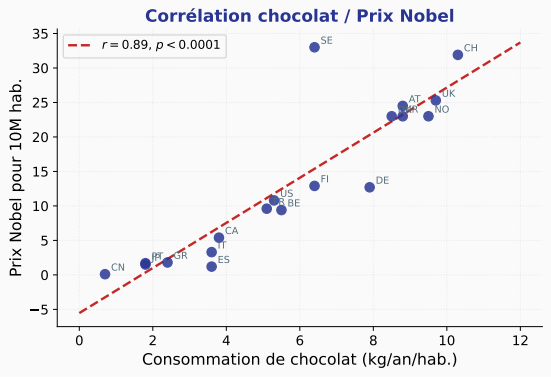
Outline

Introduction

Correlation $\neq$ Causation



A surprising result in the *NEJM*



Source

Messerli (2012)

Chocolate Consumption, Cognitive Function, and Nobel Laureates
N. Engl. J. Med., 367:1562, 2012.

Correlation: $r = 0.79$, $p < 0.0001$

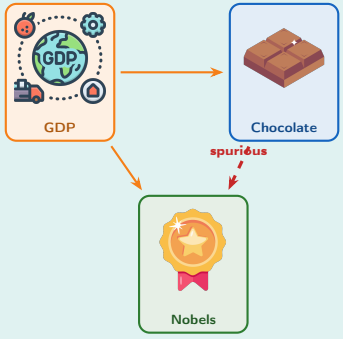
Question

Should we force scientists to eat more chocolate to win Nobel Prizes?

Correlation $\neq$ Causation.

Why the correlation is spurious

The confounder: GDP per capita



Wealthy countries simultaneously invest in science *and* consume more chocolate. No causal link.

Two very different questions

Observational question:
Among countries with high chocolate consumption, how many Nobel prizes do they win?

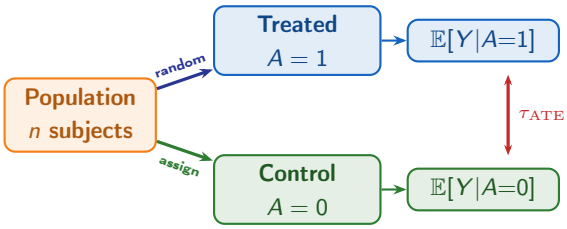
$$\mathbb{E}[\text{Nobel} \mid \text{Choc} = c] \quad \textit{(biased)}$$

Causal question:
If we *forced* a country to consume c kg, how many Nobels would it win?

$$\mathbb{E}[\text{Nobel} \mid \text{do}(\text{Choc} = c)] \quad \textit{(causal)}$$

These two quantities are **very different**.

Randomised Controlled Trials (RCTs)



Why randomisation works

Random assignment breaks $X \to A$: $A \perp\!\!\!\perp X$, so no confounding.

$$\tau_{\mathrm{ATE}} = \mathbb{E}[Y^{(1)}] - \mathbb{E}[Y^{(0)}] \approx \bar{Y}_1 - \bar{Y}_0$$

Valid **without any model** on X.

Limits of RCTs

- Expensive, sometimes unethical
- Often impossible (economics, history, epidemiology, policy...)

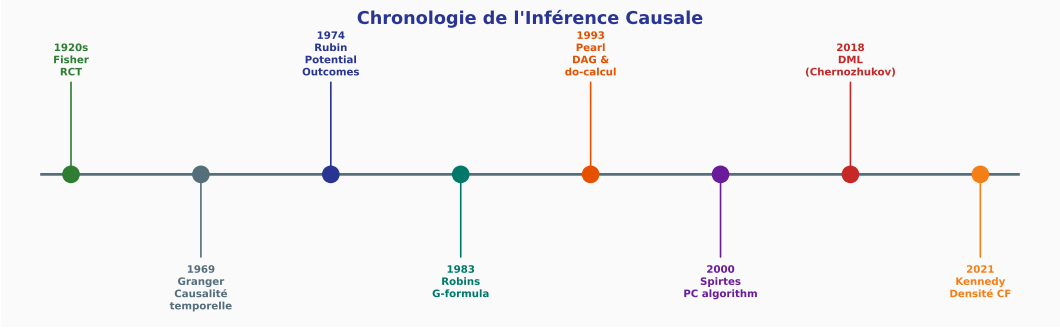
The observational challenge

In practice we observe $(X_i, A_i, Y_i)_{i=1}^n$ where subjects *self-select* into treatment.

Sicker patients receive more treatment $\Rightarrow$ naive comparison is **biased**.

Goal: recover the causal effect from observational data.

Four Paradigms for Modelling Causality



Each paradigm provides a different mathematical language for the same fundamental question: *what would happen if we intervened?* We now present each with its key formalism.

Paradigm 1: Granger Causality (1969)

Setting: time series

Stationary series (X_t), (Y_t).

Definition (Granger, 1969):
X *Granger-causes* Y if past X improves prediction of Y beyond its own past:

$$\mathbb{E}[Y_t \mid \mathcal{F}_{t-1}] \neq \mathbb{E}[Y_t \mid \mathcal{F}_{t-1}^{-X}]$$

Concrete example

Does money supply X_t Granger-cause GDP Y_t?

1. Regress Y_t on lags of Y_t only
2. Add lags of X_t to the regression
3. F-test: do the X-lags improve fit?

Limitation

This is **predictive causality**, not structural.

A common hidden cause Z_t driving both X_t and Y_t with different lags makes X *appear* causal:it is not.

$\Rightarrow$ correlation in time, not true intervention.

Paradigm 2: Potential Outcomes (Rubin, 1974)

Framework

For each unit i and treatment a, define $Y_i^{(a)}$: the outcome that *would have been* observed had unit i received treatment a (Rubin, 1974).

Fundamental problem (Holland, 1986):

$$Y_i = A_i Y_i^{(1)} + (1 - A_i) Y_i^{(0)}$$

Only *one* potential outcome is ever observed.

Key estimands:

$\tau_{\text{ATE}} = \mathbb{E}[Y^{(1)} - Y^{(0)}]$

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

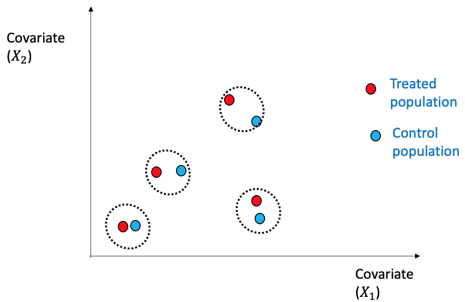


Figure 1: Matching

Paradigm 3: Structural Causal Models & DAGs (Pearl)

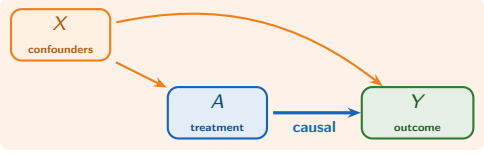
Structural Causal Model (SCM)

A set of structural equations (Pearl, 2009):

$$X_j = f_j(\mathrm{pa}(X_j), \varepsilon_j), \quad \varepsilon_j \perp\!\!\!\perp \varepsilon_k$$

$\mathrm{pa}(X_j)$: direct causal parents of X_j.

Intervention $\mathrm{do}(A = a)$:
replace A's equation by $A \equiv a$ (cut all incoming edges).



Backdoor identification

If X blocks all backdoor paths from A to Y and contains no descendants of A:

$$P(Y \mid \mathrm{do}(A = a)) = \int p(y|A = a, X = x)\, dP(x)$$

This is the **backdoor adjustment formula**.

The $\mathrm{do}(\cdot)$ operator (Pearl, 2009) makes the intervention *explicit* in the notation.

Limitation

The DAG must be **known** or learned. Causal discovery from data (Spirtes et al., 2001) (PC, FCI, GES) is identifiable only under strong assumptions and hard in high dimension.

The Backdoor Formula: Matching Intuition

From local to global: aggregation

Matching gives us local effects $\hat{\tau}(x)$. To get the ATE, we must aggregate them using the **true distribution of** X:

$$\tau_{\text{ATE}} = \int \hat{\tau}(x)\, dP(x) = \mathbb{E}_X\big[\hat{\tau}(X)\big]$$

This is exactly the **backdoor formula**:

$$\boxed{\tau_{\text{ATE}} = \int\int y \left[\eta_1(y|x) - \eta_0(y|x)\right] dy\, dP(x)}$$

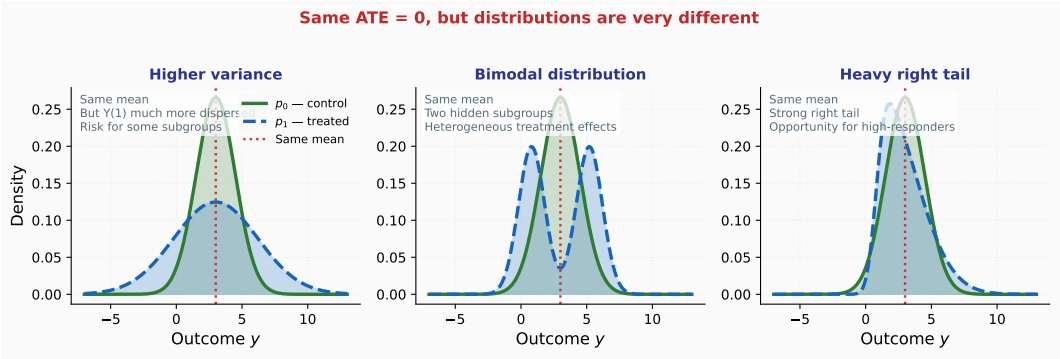
where $\eta_a(y|x) = p(y \mid A = a, X = x)$.

Three-step recipe

1. Estimate local densities $\eta_a(y|x)$ for each a.
2. Compute local effects at each x.
3. Average over the empirical distribution of X.

Why Estimate the Full Interventional Density?

The ATE Is Not Enough



Higher variance

Same mean, but $Y^{(1)}$ is much more dispersed. Some subgroups are heavily exposed to extreme outcomes.

Bimodal

Two hidden subgroups respond very differently. The treatment might be excellent for one group and harmful for another.

Heavy tail

A heavy right tail signals high-responders. Targeting them could multiply the benefit.

The Backdoor Formula for the Full Density

For the ATE (scalar)

The backdoor formula gives the ATE:

$$\tau_{\mathrm{ATE}} = \underbrace{\int y\, p_1(y)\, dy}_{\mathbb{E}[Y^{(1)}]} - \underbrace{\int y\, p_0(y)\, dy}_{\mathbb{E}[Y^{(0)}]}$$

with the **interventional density**:

$$p_a(y) = \int \eta_a(y \mid x)\, dP(x)$$

Estimating the ATE reduces to estimating the *mean* of p_a.

For the full density

The **same formula** directly delivers the full interventional density:

$$\boxed{p_t(y) = \int \eta(y \mid t, x)\, dP(x)}$$

$\Rightarrow$ The ATE is just *one summary* of p_t. The density carries everything.

The Plug-in Bias

Identification via the Backdoor Formula

Setting: continuous Y, A, X

$O = (X, A, Y), \quad X \in \mathbb{R}^d,\ A \in \mathbb{R},\ Y \in \mathbb{R}.$

Causal target: interventional mean at dose a:

$$\psi(a) = \mathbb{E}\big[Y \mid \mathrm{do}(A = a)\big]$$

Identification assumptions:

1. Consistency: $A = a \Rightarrow Y = Y(a)$
2. Unconfoundedness: $Y(a) \perp\!\!\!\perp A \mid X$
3. Positivity: $f_{A|X}(a \mid x) > 0$ a.s.

Backdoor formula and plug-in

Blocking $A \leftarrow X \rightarrow Y$ via X:

$$\psi(a) = \mathbb{E}_X\big[\mathbb{E}[Y \mid A = a, X]\big] = \mathbb{E}_X[m(a, X)]$$

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

Plug-in: $\hat{\psi}^{\mathrm{PI}}(a) = \mathbb{P}_n[\hat{m}(a, X)] = \frac{1}{n}\sum_{i=1}^{n} \hat{m}(a, X_i)$

IPTW representation

Let $\pi(a \mid x) = f_{A|X}(a \mid x)$ be the **generalised propensity score**. By iterated expectations:

$$\psi(a) = \mathbb{E}\left[\frac{f_A(a)}{\pi(a \mid X)}\, Y\right]$$

IPTW: $\hat{\psi}^{\mathrm{IPTW}}(a) = \mathbb{P}_n\left[\dfrac{\hat{f}_A(a)}{\hat{\pi}(a \mid X)}\, Y\right]$

Why Plug-in is Biased at First Order

Error decomposition

$$\hat{\psi}^{\mathrm{PI}}(a) - \psi(a) = \underbrace{(\mathbb{P}_n - P)[\hat{m}(a, X)]}_{\text{empirical fluctuation}} + \underbrace{P[\hat{m}(a, X) - m(a, X)]}_{\text{first-order bias}}$$

- **Empirical fluctuation:** $O_P(n^{-1/2})$ — harmless.
- **First-order bias:** rate $= \|\hat{m} - m\|$. With flexible ML this is **much larger than** $n^{-1/2}$:

$$\|\hat{m} - m\| = O_P\left(n^{-\frac{s}{2s+d}}\right)$$

s = smoothness, d = dimension of (a, x).

Goal: remove the bias

We want $\hat{\psi}^{\mathrm{DR}}$ such that

$$\hat{\psi}^{\mathrm{DR}} - \psi(a) = O_P(n^{-1/2})$$

even when $\hat{m}$ converges slowly, by estimating and subtracting the first-order bias via an **influence-function correction**.

Key quantity

The **influence function** (IF) satisfies:

$$\psi(\tilde{P}) - \psi(P) \approx \mathbb{E}_P\big[\mathrm{IF}(O; \tilde{P})\big]$$

Estimate it with $\hat{P}$ and subtract $\Rightarrow$ one-step estimator.

Deriving AIPTW as a One-Step Correction

AIPTW / one-step estimator

$$\hat{\psi}^{\mathrm{AIPTW}}(a) = \mathbb{P}_n\left[\hat{m}(a, X) + \frac{\hat{f}_A(a)}{\hat{\pi}(a \mid X)}(Y - \hat{m}(a, X))\right]$$

$$= \hat{\psi}^{\mathrm{PI}}(a) + \underbrace{\mathbb{P}_n[\hat{\phi}_a]}_{\text{bias correction}}$$

Error and double robustness

$$\hat{\psi}^{\mathrm{AIPTW}}(a) - \psi(a) = (\mathbb{P}_n - P)\phi_a + R_2$$

$$R_2 = O\big(\|\hat{\pi} - \pi\| \cdot \|\hat{m} - m\|\big)$$

$R_2 \to 0$ if *either* $\hat{m}$ or $\hat{\pi}$ is consistent **(double robustness)**.

Debiased Estimator: From Mean to Density

Natural extension: density functional

Let $p_a(y) = \int \eta(y \mid a, x)\, dP(x)$ be the **interventional density** of Y under $\mathrm{do}(A = a)$.

For any smooth h, define $\psi = \int h(p_a(y))\, dy$.

The **same one-step recipe** gives:

$$\hat{\psi}^{\mathrm{DR}} = \underbrace{\int h(\hat{p}_a(y))\, dy}_{\hat{\psi}^{\mathrm{PI}}} + \mathbb{P}_n \left[\frac{\hat{f}_A(a)}{\hat{\pi}(a \mid X)} \left(h'(\hat{p}_a(Y)) - \hat{\mu}_{h'}(X) \right) + \hat{\mu}_{h'}(X) \right]$$

where $\hat{\mu}_{h'}(x) = \int h'(\hat{p}_a(y))\, \hat{\eta}(y \mid a, x)\, dy$.

Conclusion

Conclusion

Main ideas

1. The **interventional density** $p_a(y)$ carries strictly more than the ATE: shape, tails, heterogeneity.
2. The **backdoor formula** $p_a(y) = \int \eta(y \mid a, x)\, dP(x)$ identifies p_a from observational data.
3. The **plug-in** $\mathbb{P}_n[\hat{\eta}(y \mid a, X)]$ is biased at first order with flexible ML.
4. A one-step **AIPW correction** restores efficiency and double robustness.

Take-away formula

$$\hat{p}_t^{\mathrm{DR}}(y) = \underbrace{\mathbb{P}_n[\hat{\eta}(y \mid t, X)]}_{\text{plug-in}} + \underbrace{\mathbb{P}_n[\hat{\phi}_t]}_{\text{one-step correction}}$$

Double robust: if *either* $\hat{\eta}$ or $\hat{\pi}$ correct.

Thank you!

Questions welcome

$$p_t(y) = \int \eta(y \mid t, x)\, dP(x)$$

References

Clive W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, 1969. doi: 10.2307/1912791.

Paul W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81 (396):945–960, 1986. doi: 10.2307/2289064.

Franz H. Messerli. Chocolate consumption, cognitive function, and nobel laureates. *New England Journal of Medicine*, 367(16):1562–1564, 2012. doi: 10.1056/NEJMon1211064.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition, 2009.

Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974. doi: 10.1037/h0037350.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, 2 edition, 2001.

The MCD Method

Marginal Contrastive Discrimination

Our proposal — **Meziani, Ndiaye & Riu (2026)**

MCD is a method that reframes conditional density estimation as a generalised contrastive learning task, enabling the use of supervised binary classification.

Step 1 — Marginal contrast function

$$= \frac{r\, f_{X,Y}(x,y)}{r\, f_{X,Y}(x,y) + (1-r)\, f_X(x) f_Y(y)}, \qquad r \in (0,1)$$

Step 2 — Recovering the conditional density

$$f_{Y|X=x}(y) \;=\; f_Y(y) \;\cdot\; \frac{}{1-} \;\cdot\; \frac{1-r}{r}$$

Probabilistic representation

$= \Pr(Z = 1 \mid W = (x,y))$

$Z \sim \mathrm{Ber}(r),\;\; f_{W|Z=1} = f_{X,Y},\;\; f_{W|Z=0} = f_X f_Y$

Consequence

Estimating = discriminating joint samples vs. product-of-marginals

$\Rightarrow$ **any** classifier works!

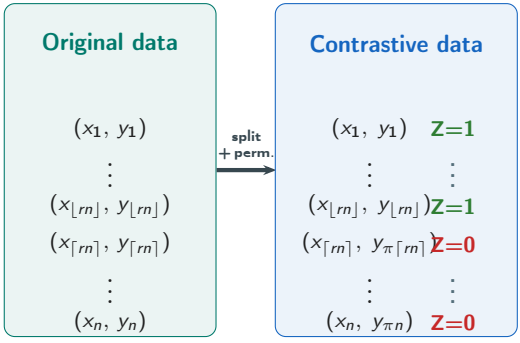
Building the Contrastive Dataset

Input: i.i.d. sample $\mathcal{D}_n^{X,Y} = \{(x_i, y_i)\}_{i=1}^n$ from $f_{X,Y}$.

Construction

1. Split into two halves $\mathcal{D}_1$, $\mathcal{D}_2$
2. Assign $Z{=}1$ to pairs in $\mathcal{D}_1$ (joint)
3. **Permute** the y_i's in $\mathcal{D}_2$, assign $Z{=}0$
4. Concatenate $\Rightarrow \mathcal{D}_N^{W,Z}$

Train **any** classifier on $\mathcal{D}_N^{W,Z}$ to get $\hat{q}_r(x, y)$.



Algorithm Estimating the Interventional Density

Input: $\mathcal{D} = \{(x_i, t_i, y_i)\}_{i=1}^{n}$, split ratio α, treatment value t.

1. **Split the data**

$$\mathcal{D} = \underbrace{\mathcal{D}_{\mathrm{DE}}}_{\lfloor \alpha n \rfloor \text{ samples}} \cup \underbrace{\mathcal{D}_{\mathrm{MC}}}_{n - \lfloor \alpha n \rfloor \text{ samples}}$$

2. **Learn the conditional density (MCD plug-in)**

   Fit $_{Y|T=t,X=x}(y)$ on $\mathcal{D}_{\mathrm{DE}}$ via contrastive dataset + classifier.

3. **Monte Carlo estimation**

$$\boxed{Y|\mathrm{do}(T=t)(y) \;=\; \frac{1}{|\mathcal{D}_{\mathrm{MC}}|} \sum_{x_j \in \mathcal{D}_{\mathrm{MC}}} Y|T=t,X=x_j(y)}$$

Experimental Setup

Data Generating Process

Joint Gaussian covariates and treatment:

$(T, X) \sim \mathcal{N}(\mu, \Sigma)$,

$Y = b^\top T + a^\top X + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2)$

$Y \mid \mathrm{do}(T{=}t) \sim \mathcal{N}\big(b^\top t + a^\top \mu_X,\ a^\top \Sigma_X a + \sigma^2\big)$

$\Rightarrow$ **Closed-form ground truth** for exact evaluation.

Protocol

Sample size	$n = 1\,000$
Replications	100
Train split	$\alpha = 0.8$
Hyperparameters	10-fold cross-validation
Metric	Wasserstein-1 distance

Methods Compared

Method	Description
MCD:MLP	Ours — MLP
MCD:CAT	Ours — CatBoost
NF	Normalizing Flows (Rezende & Mohamed 2016)
RFCDE	Random Forest CDE (Pospisil & Lee 2018)

Results : Accuracy & Computation Time

Mean (s.d.) over 100 replications. **Green** = best; underline = 2nd best.

$(\dim(X), \dim(T))$	Metric	MCD:MLP	MCD:CAT	NF	RFCDE
$(10, 3)$	$W_1 \downarrow$	**0.28 (0.03)**	0.36 (0.02)	0.36 (0.07)	0.42 (0.05)
	Time (s)$\downarrow$	14.91 (3.71)	**3.89 (0.02)**	33.36 (15.05)	5.41 (0.19)
$(100, 10)$	$W_1 \downarrow$	**0.79 (0.03)**	1.04 (0.02)	1.04 (0.02)	1.06 (0.02)
	Time (s)$\downarrow$	13.05 (4.85)	**5.50 (0.04)**	58.47 (37.42)	9.83 (1.02)

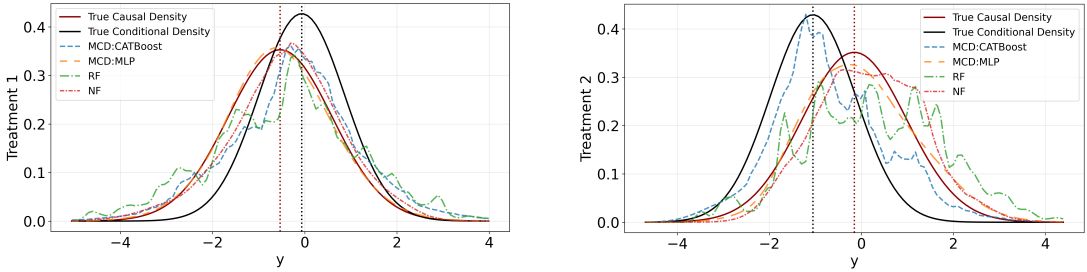
Accuracy

MCD:MLP achieves the **best** W_1 in all configurations, especially at $(100, 10)$.

Speed

MCD:CATBoost is the fastest accurate method. NF is $\approx 4\times$ slower with much higher variance.

Results : Estimated vs. True Causal Densities



MCD:MLP tracks the true causal density most closely in both panels.