Counterfactual Generative Models for Time-Varying Treatments

Shenghao Wu[1], Wenbin Zhou[1], Minshuo Chen[2] and Shixiang Zhu[1]

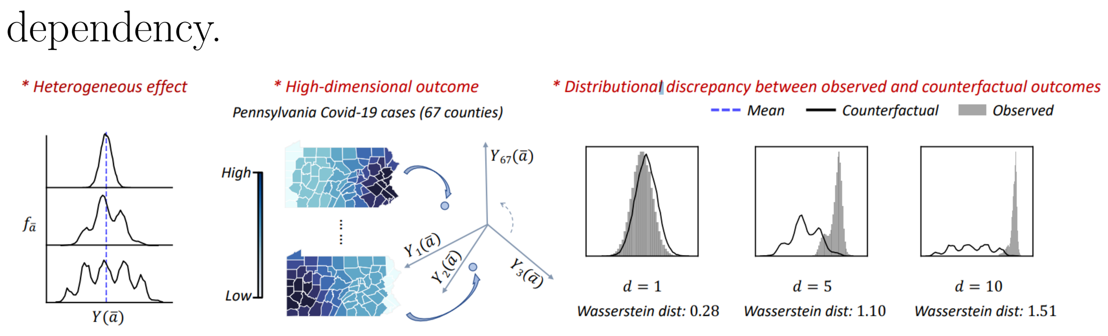
[1]*Carnegie Mellon University and* [2]*Princeton University*

Introduction

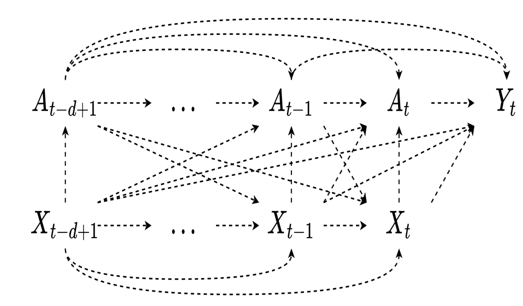
Estimating the counterfactual outcome of treatment is essential for decision-making in public health and clinical science, among others.

Challenges: There are **three** unaddressed main challenges for counterfactual inference considering time-varying treatments.

- **Heterogeneity:** The mean is incapable of describing the heterogenous effect in counterfactual distribution.
- **High-dimensionality:** Estimation accuracy of high-dimensional counterfactual outcomes quickly degrades.
- **Distributional discrepancy:** Greater distributional mismatch is observed for longer treatment history dependency.



Setting: At time t, denote the outcome variable as Y_t, denote the d-length history of treatments and covariates as $\overline{A}_t = (A_{t-d+1}, \dots, A_t)$ and $\overline{X}_t = (X_{t-d+1}, \dots, X_t)$. Lowercase letters represent their realizations. Distributions are denoted as f. The causal DAG is assumed to be as the figure to the right.



Goal: We aim to learn a generator function that produces samples of the outcome variable y given time-varying treatment $\overline{a}$,

$$g_\theta(z, \overline{a}) : \mathbb{R}^r \times \mathcal{A}^d \to \mathcal{Y}.$$

The generator can be learned by maximizing the following (intractable) likelihood function

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \mathbb{E}_{\overline{A}} \left[\mathbb{E}_{y \sim f_{\overline{a}}} \log f_\theta(\cdot | \overline{a}) \right].$$

Learning Objective

Lemma: Under unconfoundedness and positivity,

$$f_{\overline{a}}(y) = \int \frac{1}{\prod_{\tau=t-d+1}^{t} f(a_\tau | \overline{a}_{\tau-1}, \overline{x}_\tau)} f(y, \overline{a}, \overline{x}) \, d\overline{x}.$$

Proposition: The generative learning objective can be approximated by:

$$\mathbb{E}_{\overline{A}} \left[\mathbb{E}_{y \sim f_{\overline{a}}} \log f_\theta(y|\overline{a}) \right] \approx \frac{1}{N} \sum_{(y, \overline{a}, \overline{x}) \in \mathcal{D}} w_\phi(\overline{a}, \overline{x}) \log f_\theta(y|\overline{a}),$$

where N represents the sample size, and $w_\phi(\overline{a}, \overline{x})$ denotes the subject-specific IPTW, parameterized by $\phi \in \Phi$, which takes the form:

$$w_\phi(\overline{a}, \overline{x}) = \frac{1}{\prod_{\tau=t-d+1}^{t} f_\phi(a_\tau | \overline{a}_{\tau-1}, \overline{x}_\tau)}.$$

Model Architecture

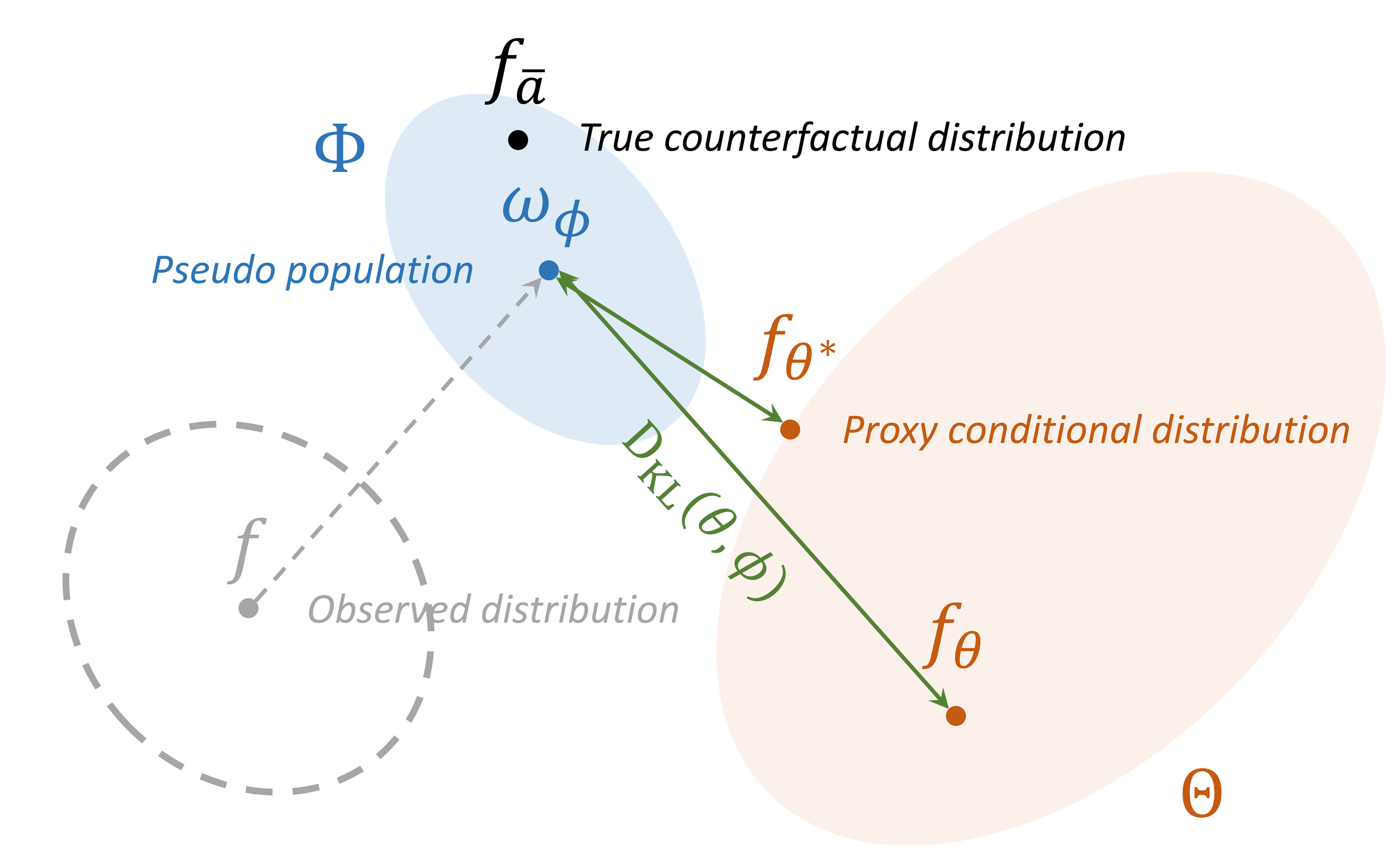


Figure 1: The learning objective is to minimize the KL-divergence between the true **counterfactual distribution** $f_{\overline{a}}$ and a proxy conditional distribution $f_\theta(\cdot|a)$

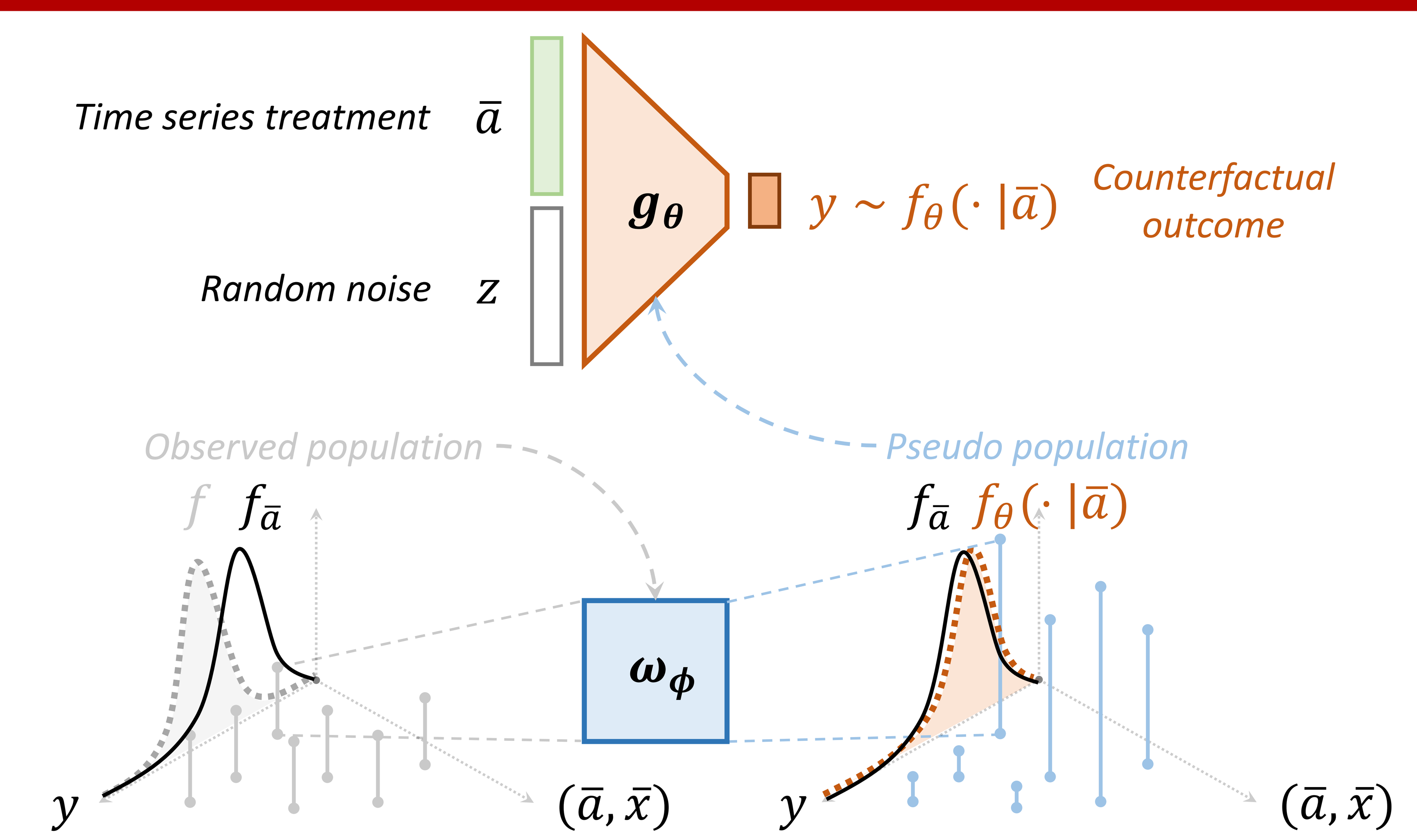


Figure 2: The generator g_θ produces **samples of the outcome variable** y given time varying treatment $\overline{a}$. The generated samples conform to the learned proxy distribution.

Experiments

Our framework is flexible to deploy with likelihood-based generative learning algorithms such as:

- **Classifier-free guided diffusion model:**

$$\log f_\theta(\cdot|\overline{a}) \geq -\mathbb{E}_{s \sim [1,S], y \sim f(y|\overline{a}), \epsilon_s} \| \epsilon_s - \epsilon_\theta(\sqrt{\bar{\lambda}_s} y + \sqrt{1 - \bar{\lambda}_s} \epsilon_s, s, \overline{a}) \|^2.$$

- **Conditional variational autoencoder:**

$$\log f_\theta(\cdot|\overline{a}) \geq -D_{\mathrm{KL}}\left(q(z|y, \overline{a}) || p_\theta(z|\overline{a})\right) + \mathbb{E}_{q_\theta(z|y,\overline{a})} \left[\log p_\theta(y|z, \overline{a})\right].$$

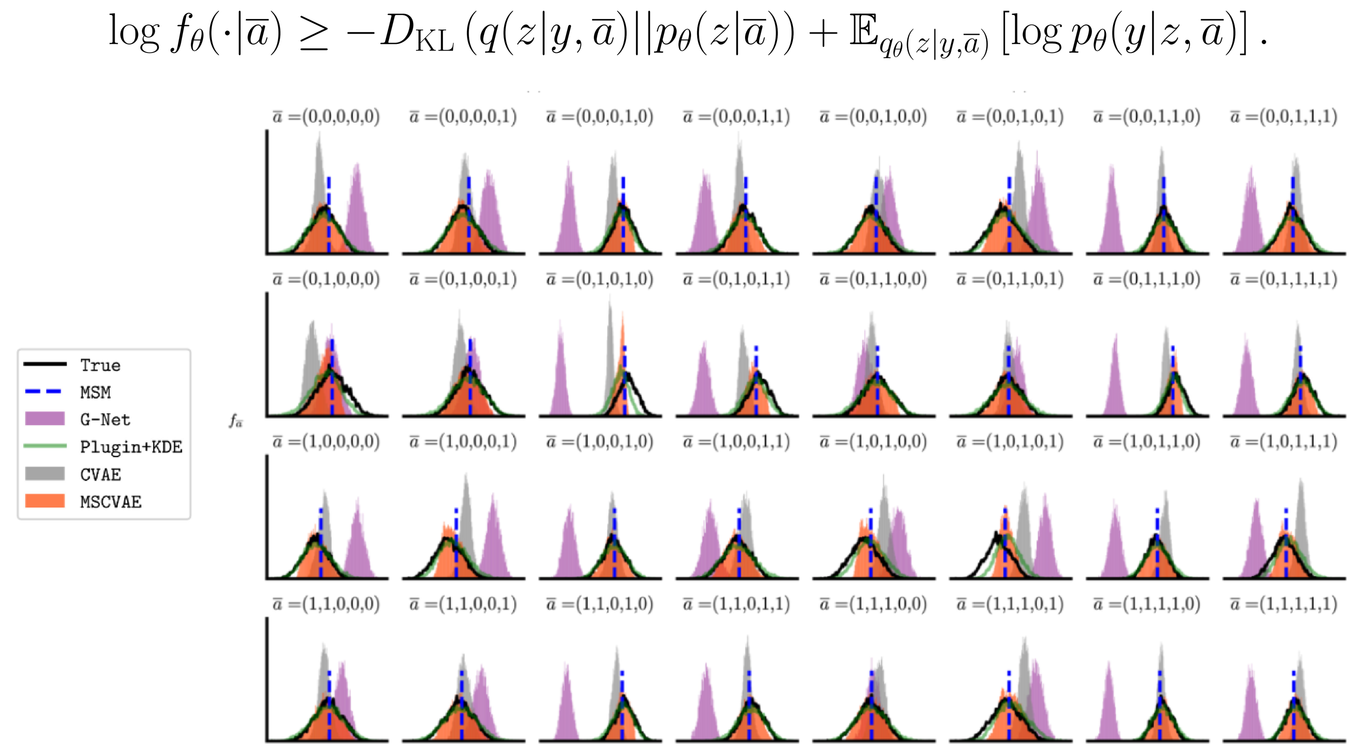


Figure: The estimated and true counterfactual distribution with history length of $d = 5$ on the fully synthetic dataset ($m = 1$)

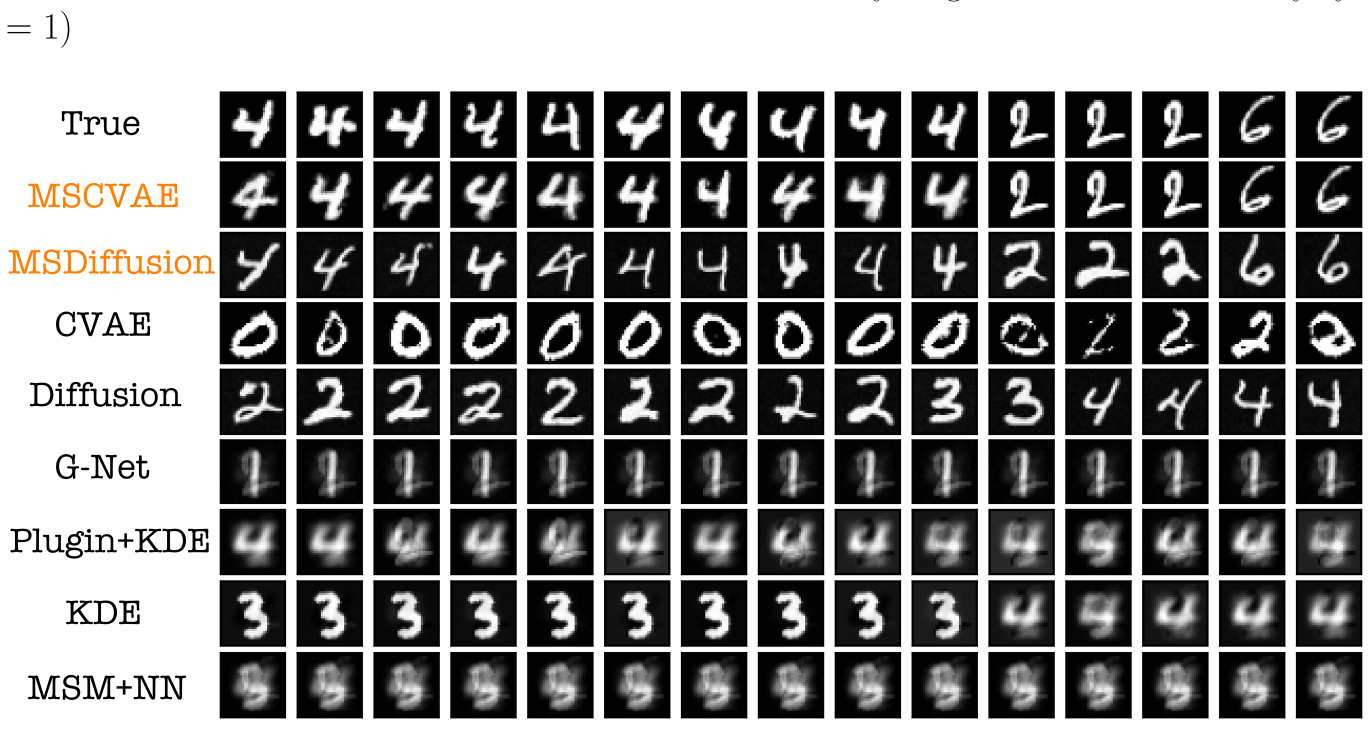


Figure: Results on the semi-synthetic TV-MNIST datasets ($m = 784$) generated under $\overline{a} = (1, 1, 1)$.

	$d = 1$		$d = 3$		$d = 5$	
Methods	Mean ↓	Wasserstein ↓	Mean ↓	Wasserstein ↓	Mean ↓	Wasserstein ↓
MSM+NN	**0.001** (**0.002**)	0.601 (0.603)	0.070 (0.159)	0.689 (0.718)	0.198 (**0.563**)	0.600 (0.737)
KDE	0.246 (0.267)	0.244 (0.268)	0.520 (1.080)	0.538 (1.080)	0.538 (1.419)	0.539 (1.419)
Plugin+KDE	0.010 (0.014)	**0.034** (**0.036**)	**0.045** (0.168)	0.132 (**0.168**)	**0.147** (0.598)	0.182 (**0.598**)
CRN	0.228 (0.280)	0.289 (0.331)	0.913 (1.753)	1.014 (1.757)	1.713 (4.080)	1.775 (4.080)
G-Net	0.211 (0.258)	0.572 (0.582)	1.167 (2.173)	1.284 (2.173)	2.314 (5.263)	2.354 (5.263)
CVAE	0.250 (0.287)	0.253 (0.288)	0.517 (1.061)	0.553 (1.061)	0.539 (1.430)	0.613 (1.430)
MSCVAE	0.006 (0.006)	0.055 (0.056)	0.046 (**0.150**)	**0.105** (0.216)	0.150 (0.633)	**0.173** (0.633)
MSDiffusion	0.029 (0.052)	0.056 (0.065)	0.086 (0.234)	0.135 (0.234)	0.207 (0.845)	0.259 (0.845)

Table: Quantitative performance on fully-synthetic data

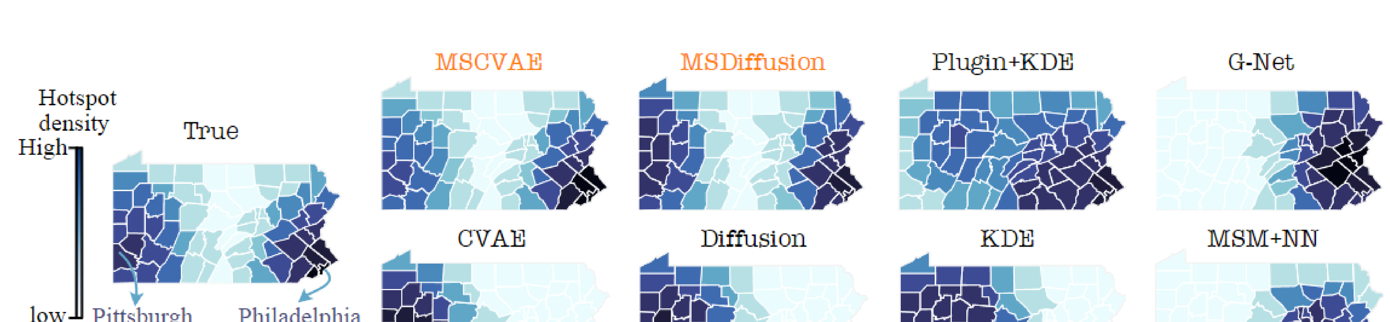
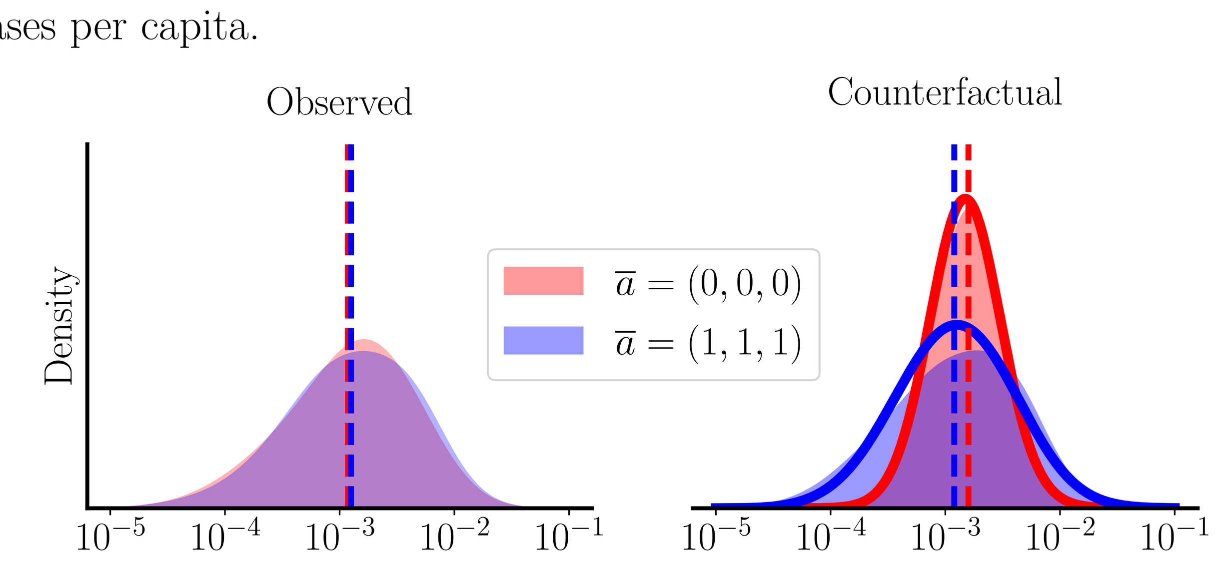


Figure: Results on the semi-synthetic Pennsylvania COVID-19 mask datasets ($m = 67$) under the treatment combination $\overline{a} = (1, 1, 1)$.

Application to COVID-19 Analysis

Description: 5 features of 3219 U.S. counties are collected in 2020-2021 spanning across 49 weeks. We aim to make counterfactual predictions regarding how mask policies affect COVID-19 number of cases per capita.



Insight: Imposing mask mandate can decrease the **mean** of the distribution, but increases its **variance** in the same time. This implies that while mask mandate tend to help control virus spread, a thorough examination of the specific circumstances is highly recommended for mask-policymakers to avoid any unintended consequences.