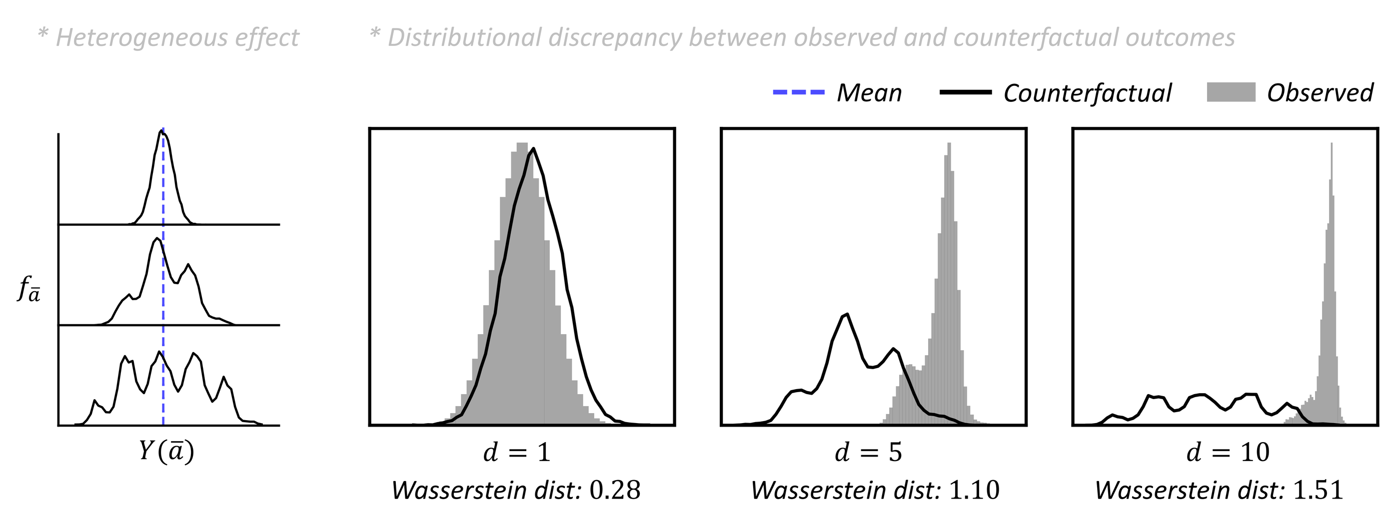
Counterfactual Generative Models for Time-Varying Treatments

Shenghao Wu[1], **Wenbin Zhou**[1], Minshuo Chen[2] and Shixiang Zhu[1]
[1] Carnegie Mellon University, [2] Princeton University

Introduction

Motivating example: Extreme differences beyond a simple shift in the average cannot be captured by the widely adopted average treatment effect outcome metric.

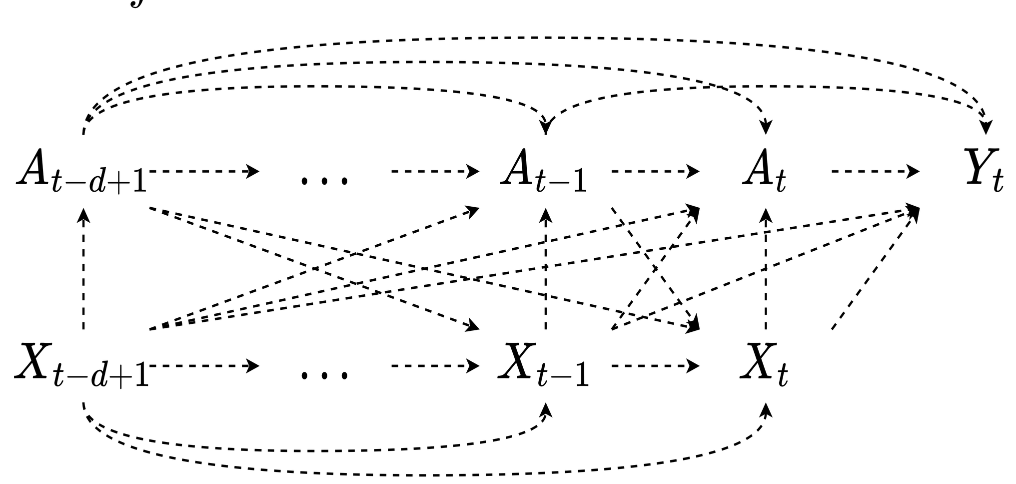


Goal: Our goal is to systematically address the following three practical challenges for data-driven decision making in one versatile and model-agnostic framework.

- **Counterfactual Inference:** The goal is to infer what would have happened if were to act in a way *not* observed in previous results.
- **Temporal Setting:** Collected data is blurred with treatments and confounders that has *time-dependent* structures.
- **Distribution Learning:** People care about the entire counterfactual *distribution* of the outcome variable.

Preliminaries

Notation: At time t, denote the outcome variable as Y_t, denote the d-length history of treatments and covariates as $\overline{A}_t = (A_{t-d+1}, \ldots, A_t)$ and $\overline{X}_t = (X_{t-d+1}, \ldots, X_t)$. Lowercase letters represents their realizations. We use f to denote distribution.

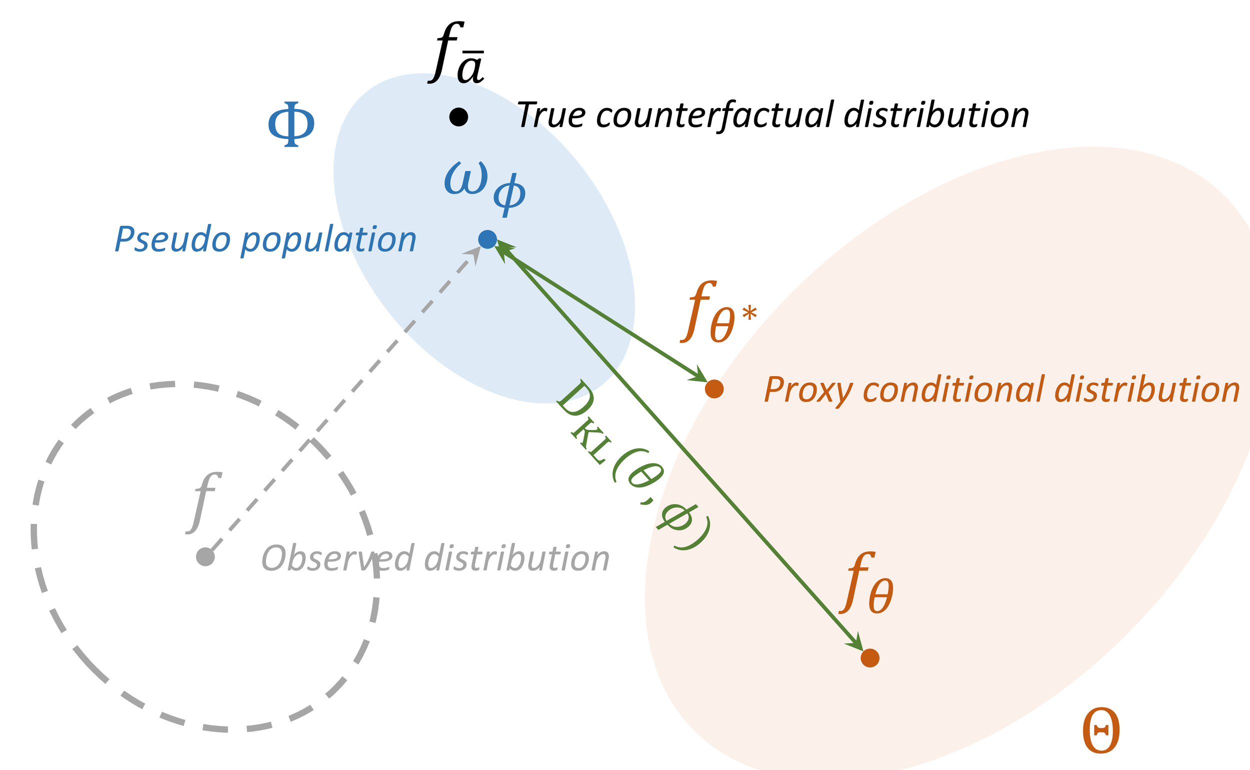


Important Lemma: Under some standard assumptions, we have

$$f_{\bar{a}}(y) = \int \frac{\mathbb{1}\{\overline{A} = \bar{a}\}}{\prod_{\tau=t-d}^{t} f\left(A_\tau | \overline{A}_{\tau-1}, \overline{X}_\tau\right)} f\left(y, \overline{A}, \overline{X}\right) d\overline{A} d\overline{X},$$

Proposed Method

Learning Objective: We aim to minimize the Kullback–Leibler (KL) divergence between a proxy conditional distribution $f_\theta(\cdot|a)$ and $f_{\bar{a}}$.



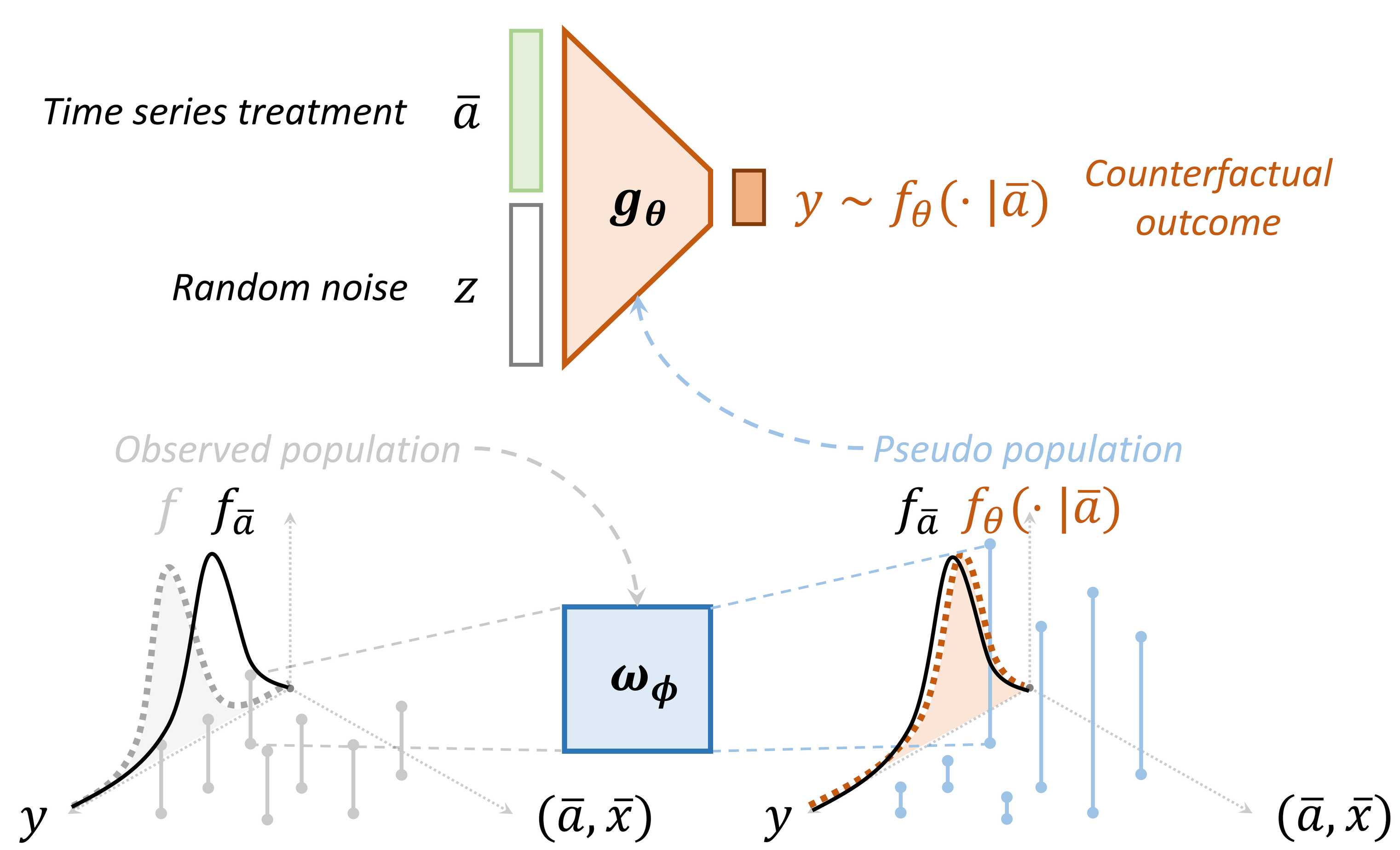
Loss Function: This generative learning objective can be approximated by maximizing the log-likelihood:

$$\mathbb{E}_{y \sim f_{\bar{a}}} \log f_\theta(y|\bar{a}) \approx \sum_{(y,\bar{a},\bar{x}) \in \mathcal{D}} w_\phi(\bar{a}, \bar{x}) \log f_\theta(y|\bar{a}),$$

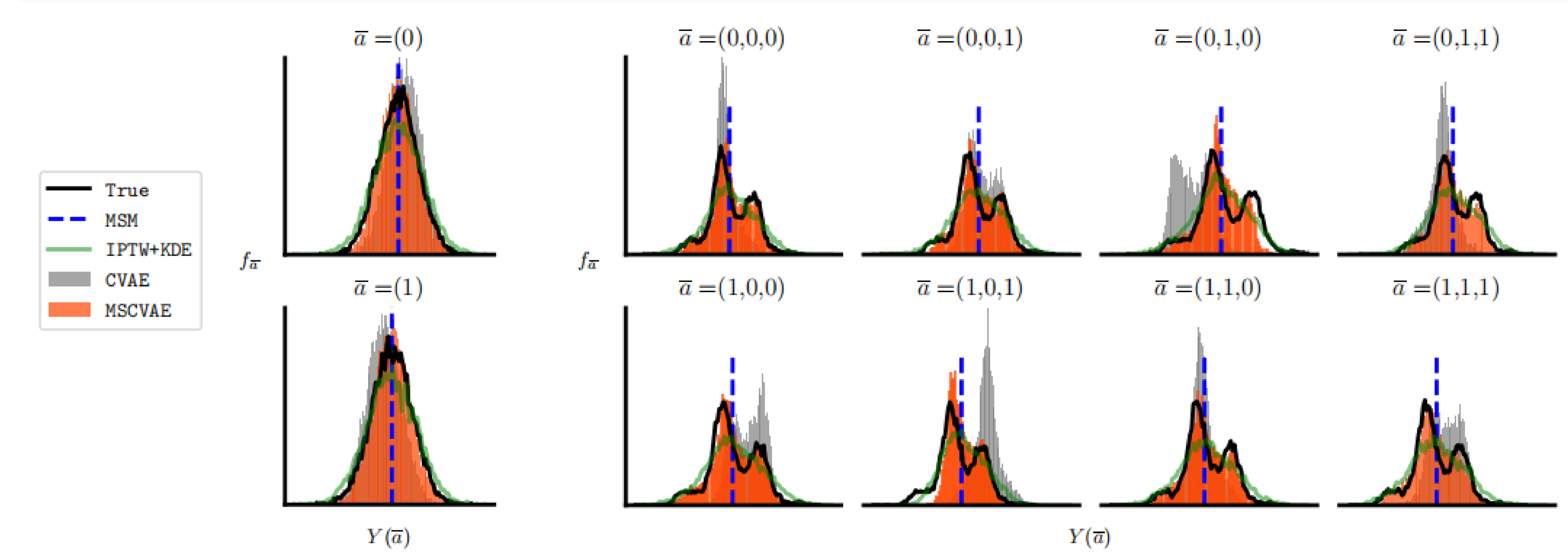
where $w_\phi(\bar{a}, \bar{x})$ denotes the subject-specific IPTW, parameterized by $\phi \in \Phi$, which takes the form:

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

Model Architecture: Our proposed model, `MSCVAE`, adopts a standard encoder-decoder structure.



Synthetic Experiment



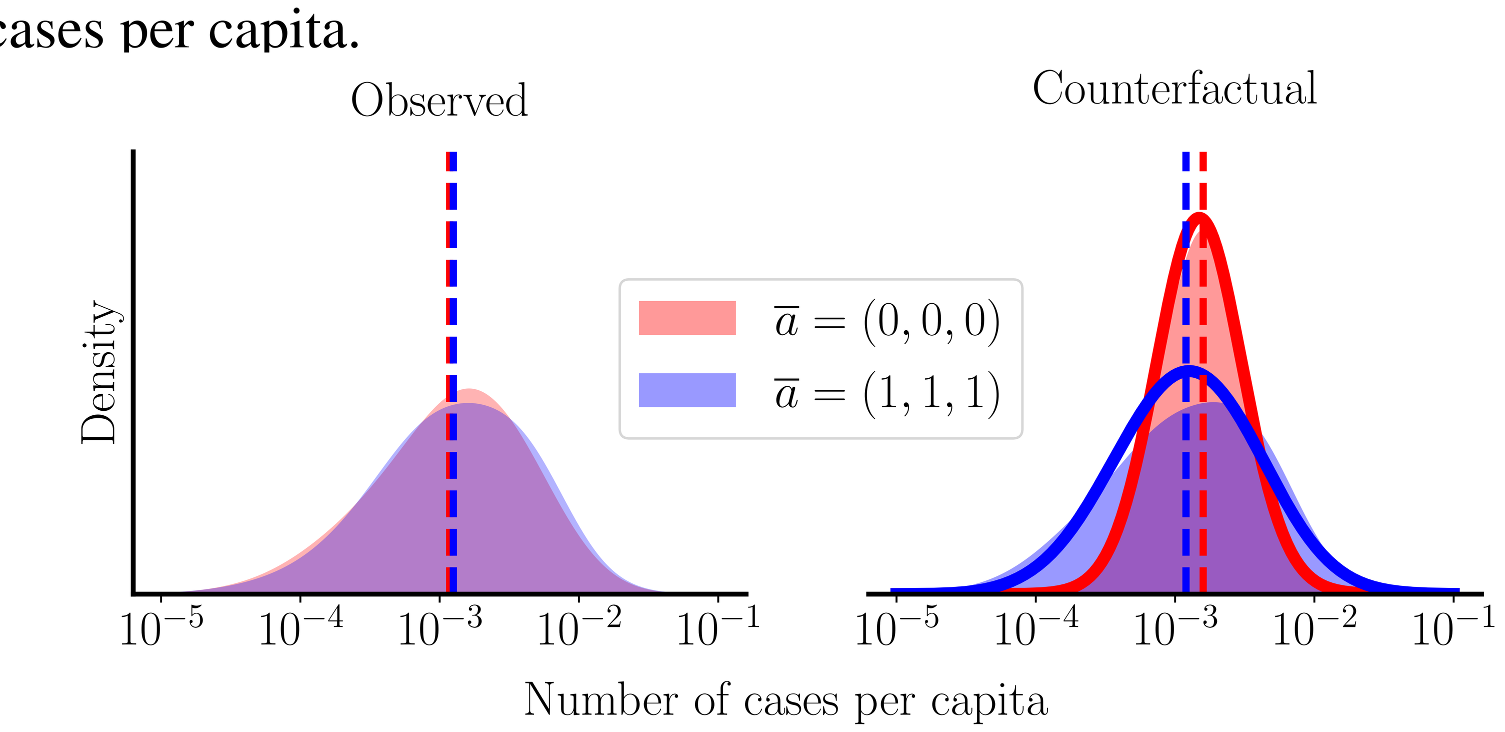
(a) $d = 1$ (b) $d = 3$

	$d=1$		$d=3$		$d=5$	
Methods	Mean	Wasserstein	Mean	Wasserstein	Mean	Wasserstein
`Linear MSM`	**0.003**	NA	0.055	NA	0.186	NA
`KDE`	0.246	0.433	0.528	0.579	0.536	0.601
`IPTW+KDE`	0.010	0.127	0.048	0.133	**0.146**	0.181
`CVAE`	0.263	0.264	0.524	0.559	0.537	0.612
`MSCVAE`	0.008	**0.053**	**0.043**	**0.107**	0.147	**0.171**

Conclusion: `MSCVAE` outperforms other baselines on synthetic data.

COVID-19 Data Experiment

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.