Scalable Counterfactual Distribution Estimation in Multivariate Causal Models

Thong Pham THONG-PHAM@BIWAKO.SHIGA-U.AC.JP
Faculty of Data Science, Shiga University / RIKEN AIP
Shohei Shimizu SHOHEI-SHIMIZU@BIWAKO.SHIGA-U.AC.JP
Faculty of Data Science, Shiga University / RIKEN AIP
Hideitsu Hino HINO@ISM.AC.JP
The Institute of Statistical Mathematics / RIKEN AIP
Tam Le TAM@ISM.AC.JP
The Institute of Statistical Mathematics / RIKEN AIP

Editors: Francesco Locatello and Vanessa Didelez

Abstract

We consider the problem of estimating the counterfactual joint distribution of multiple quantities of interests (e.g., outcomes) in a multivariate causal model extended from the classical difference-in-difference design. Existing methods for this task either ignore the correlation structures among dimensions of the multivariate outcome by considering univariate causal models on each dimension separately and hence produce incorrect counterfactual distributions, or poorly scale even for moderate-size datasets when directly dealing with such a multivariate causal model. We propose a method that alleviates both issues simultaneously by leveraging a robust latent one-dimensional subspace of the original high-dimension space and exploiting the efficient estimation from the univariate causal model on such space. Since the construction of the one-dimensional subspace uses information from all the dimensions, our method can capture the correlation structures and produce good estimates of the counterfactual distribution. We demonstrate the advantages of our approach over existing methods on both synthetic and real-world data.

Keywords: multivariate counterfactual distribution, optimal transport, difference in difference

1. Introduction

Causal inference has received explosive interest in the last decades, due to the need to extract causal knowledge from data in various research fields, such as statistics (Pearl, 2009), sociology (Gangl, 2010), biomedical informatics (Kleinberg and Hripcsak, 2011), public health (Glass et al., 2013), and machine learning (Schölkopf, 2022). One of the most popular causal inference models in practice is the difference in difference (DiD) model, which dates back to the works of John Snow in the 1850s (Snow, 1854, 1855; Donald and Lang, 2007; Lechner, 2011; Roth et al., 2023). In this model, we observe a quantity of interest (i.e., outcome) from two different groups, i.e., the treatment group and the control group, at two different time steps, i.e., before and after the intervention (i.e., treatment) event. More precisely, the intervention is only applied on the treatment group while there is no intervention applied on the control group. We assume that if there was no intervention in the treatment group, its outcome variable would evolve the same way as that of the control group. This is the so-called "*parallel trend*" assumption. The classical DiD model, however, requires that the

means of the outcome variable in the two groups must evolve the same way when there is no intervention. This may limit its application in practice. To extend the DiD model, several proposals have been developed in the literature (Abadie, 2005; Athey and Imbens, 2006; Blundell and Costa Dias, 2009; Sofer et al., 2016), of which notably is the Changes-in-Changes (CiC) model (Athey and Imbens, 2006).

The CiC model generalizes the DiD model to include parallel trends that can act on the whole distribution of the outcome, i.e., in the absence of intervention, the means of the outcome in control and treatment groups are allowed to evolve in different ways, as long as the two outcome *distributions* evolve in the same way. This allows identifications of more complex treatment effects that require information from the whole outcome distributions, not just their first moments (Lechner, 2011). The standard CiC model, however, is only designed for univariate outcomes. In order to extend the CiC model for a multivariate outcome variable, a naive approach is to tensorize univariate CiC models, i.e., considering independently a univariate CiC model for each coordinate. Nevertheless, this naive approach fails to capture correlations among coordinates of the outcome and thus is incapable of modelling complex, multivariate parallel trends.

Recently, by leveraging the optimal transport (OT) theory (Villani, 2003, 2008), Torous et al. (2021) proved that the counterfactual outcome distribution of the treatment group in the CiC model (i.e., the outcome distribution of the treatment group *without* receiving intervention) at the post-intervention time stamp can be estimated by exploiting the *optimal transport map* which pushforwards the outcome distribution of the control group at the pre-intervention time stamp to that at the post-intervention time stamp. Consequently, it is natural to extend the CiC model for univariate quantity of interest into that for multivariate one through the lens of OT, since this would take into account the dependence structure of the dimensions and be able to model complex parallel trends.

However, OT suffers a few drawbacks. It has a high computational complexity, which is super cubic with respect to the number of supports of the input distribution. A popular approach is to rely on the entropic regularization for OT, a.k.a., Sinkhorn (Cuturi, 2013), to reduce its computational complexity to quadratic. Yet, Sinkhorn yields a dense estimator for the optimal transport plan, which is not a desirable property for counterfactual estimation in the CiC model. Additionally, OT has a high sample complexity, i.e., $\mathcal{O}(n^{-1/d})$ where n is the number of samples and d is the dimension of samples in the probability measures.

In this work, in order to exploit the efficient computation of the CiC model for the univariate quantity of interest, and alleviate the above-mentioned challenges of the OT approach for the multivariate CiC model, we propose to leverage the max-min robust OT approach (Paty and Cuturi, 2019; Deshpande et al., 2019; Le et al., 2024). In particular, we propose to lift the univariate CiC model to that for a multivariate quantity of interest by seeking a *robust* latent univariate subspace. Unlike the naive tensorization approach, our approach can incorporate the correlations of coordinates. Moreover, unlike the standard OT approach as in (Torous et al., 2021), our estimator can preserve the efficient computation as in the univariate CiC model since the optimal transport plan is estimated on the robust latent one-dimensional subspace instead of its original high-dimensional space.

Intuitively, our approach follows the *max-min robust OT* approach (Paty and Cuturi, 2019; Deshpande et al., 2019; Le et al., 2024), which steams from the robust optimization (Ben-Tal et al., 2009; Bertsimas et al., 2011) where there are uncertainty non-stochastic parameters. The robust optimization has many roots in applied sciences, e.g., in robust control (Keel et al., 1988), machine learning (Morimoto and Doya, 2001; Xu et al., 2009; Panaganti et al., 2022). In the context of OT, several advantages of the max-min (and its relaxation min-max) robust OT have been reported.

For example, (i) it makes the OT approach robust to noise (Paty and Cuturi, 2019; Dhouib et al., 2020; Le et al., 2024); and (ii) it also helps to reduce the sample complexity (Paty and Cuturi, 2019; Deshpande et al., 2019). At a high level, our contributions are two-fold as follows:

- (i) We propose a max-min robust OT approach for the multivariate CiC model. The proposed approach not only inherits properties of the OT approach for the CiC model but also preserves the efficient computation as in the univariate CiC model.
- (ii) We evaluate our approach on both synthesized and real data to illustrate the advantages of the proposed method.

The paper is organized as follows: we review the multivariate CiC model and existing methods for estimating the counterfactual distribution in this model in Section 2. Then, we discuss our proposed method in Section 3 and demonstrate its benefits through synthetic data in Section 4. We apply it to the classical dataset of Card and Krueger (1993) in Section 5 before giving concluding remarks in Section 6. Furthermore, we have released code for our proposed approach.[1]

Notations. We use the superscripts $\mathtt{C}$ and $\mathtt{T}$ to indicate the control group and treatment group, respectively. We drop those superscripts when either the context is clear or it is not necessary to distinguish these two groups.

2. The Causal Model

In this section, we describe the CiC causal model for multiple quantities of interests (Torous et al., 2021), which is an extension based on OT theory from the original, univariate model (Athey and Imbens, 2006). Some additional discussions can be found in Appendix D.

We use a stochastic process to model the quantity of interests (i.e., outcomes) before the intervention, i.e., at the time stamp $t = 0$, and after the intervention, i.e., at the time stamp $t = 1$. We denote them as $\{Y_t\}_{t=0,1}$ where Y_t is in the $\mathbb{R}^d$ space. For the original CiC causal model, we have $d = 1$ (Athey and Imbens, 2006). We let μ_t be the distribution of Y_t for $t = 0, 1$. Without loss of generality, we assume that Y_t is generated from a latent variable $U_t \in \mathbb{R}^d$ which may change over time, but the distribution ν of the latent variable U_t is not changed over time, i.e., time-invariant $U_t \sim \nu$ for $t = 0, 1$. Intuitively, U_t may be regarded as (unobserved) intrinsic features of a sample. In the CiC model, we observe two groups:

- (i) the control group: the stochastic process $\{Y_t^{\mathtt{C}}\}$ is solely affected by the *natural drift*. The evolution from $\{Y_0^{\mathtt{C}}\}$ at $t = 0$ to $\{Y_1^{\mathtt{C}}\}$ at $t = 1$ is independent of the treatment effects (of intervention).
- (ii) the treatment group: the stochastic process $\{Y_t^{\mathtt{T}}\}$ is affected by both the *natural drift* and the *treatment effects*.

For the CiC causal model, the goal of our causal inference is to deconvolve the *natural drift* and the *treatment effects* in the treatment group. For example, we would like to estimate the counterfactual distribution of the control group at post-intervention under *only natural drift effect* (i.e., without the treatment effects). By doing so, we can estimate the *treatment effects* of the intervention for the considered groups in application domains.

1. https://github.com/thongphamthe/scalable-counterfactual

2.1. The Natural Drift Model

Natural drift is best explained in the stochastic process $\{Y_t^{\mathtt{C}}\}$ for $t = \{0, 1\}$ since this process involves solely the natural drift and is not affected by the treatment effect. The change of $\{Y_t^{\mathtt{C}}\}$ from the pre-intervention ($t = 0$) to the post-intervention ($t = 1$) is modeled by assuming the existence of two *production functions* $h_t : \mathbb{R}^d \mapsto \mathbb{R}^d$ with $t \in \{0, 1\}$ such that

$$Y_t^{\mathtt{C}} = h_t\left(U_t^{\mathtt{C}}\right).$$

Consequently, we have $\mu_t^{\mathtt{C}} = (h_t)_\sharp \nu^{\mathtt{C}}$ where we introduce a new notation $\sharp$ as the *pushforward* operator which is defined as for any measurable set $A \subseteq \mathbb{R}^d$, $\nu(A) = \mu(h_t^{-1}(A))$ (Peyré and Cuturi, 2019, Def. 2.1). In other words, the distribution of the quantity of interests in the control group at the time stamp t (i.e., μ_t^C) is the pushforward of the distribution of the latent variable $U_t^{\mathtt{C}}$ of the control group at the time stamp t (i.e., $\nu^{\mathtt{C}}$) by the production function h_t.

Natural drift map. Assume that the production function h_0 is invertible, we have $\nu^{\mathtt{C}} = (h_0^{-1})_\sharp \mu_0^C$. Additionally, the distribution $\nu^{\mathtt{C}}$ of $U_t^{\mathtt{C}}$ is time-invariant. Thus, we have

$$\mu_1^{\mathtt{C}} = (h_1 \circ h_0^{-1})_\sharp \mu_0^{\mathtt{C}}. \tag{1}$$

Equivalently, $\mu_1^{\mathtt{C}}$ is the pushforward of $\mu_0^{\mathtt{C}}$ by the *natural drift map*

$$\mathtt{f} = h_1 \circ h_0^{-1}. \tag{2}$$

Natural drift in the treatment group. We first introduce the concept of a *counterfactual distribution* of the outcome in this group, which is the outcome variable of the treatment group at post-intervention $t = 1$ under the purely hypothetical situation that the treatment was never applied. Denote $\{Y_1^{\mathtt{T}^*}\}$ as this hypothetical stochastic process for the outcome and its distribution as $\mu_1^{\mathtt{T}^*}$. It is assumed that the change from $\{Y_0^{\mathtt{T}}\}$ to $\{Y_1^{\mathtt{T}^*}\}$ is governed by the same production functions h_0 and h_1 used in modeling natural drift in the control group, namely $Y_0^{\mathtt{T}} = h_0(U_0^T)$ and $Y_1^{\mathtt{T}^*} = h_1(U_1^T)$. This assumption generalizes the parallel trend assumption in the classical DiD model. Then, similarly to Equation (1), we have

$$\mu_1^{\mathtt{T}^*} = (h_1 \circ h_0^{-1})_\sharp \mu_0^{\mathtt{T}} = (\mathtt{f})_\sharp \mu_0^{\mathtt{T}}. \tag{3}$$

Thus, the counterfactual distribution $\mu_1^{\mathtt{T}^*}$ is the pushforward of $\mu_0^{\mathtt{T}}$ by the natural drift map $\mathtt{f}$. Moreover, Equations (1) and (3) suggest the following two-step method to estimate $\mu_1^{\mathtt{T}^*}$, the counterfactual distribution of outcomes in the treatment group under only effects of the natural drift:

- (i) We first estimate the natural drift map $\mathtt{f}$ from observed samples in the stochastic processes $\{Y_0^{\mathtt{C}}\}, \{Y_1^{\mathtt{C}}\}$ of the distributions $\mu_0^{\mathtt{C}}, \mu_1^{\mathtt{C}}$ respectively. How to perform this estimation will be discussed in the next section.
- (ii) We then use the estimated natural drift map $\mathtt{f}$ as the pushforward function for the stochastic process $\{Y_0^{\mathtt{T}}\}$ with distribution $\mu_0^{\mathtt{T}}$ to obtain an estimate for the distribution $\mu_1^{\mathtt{T}^*}$.

The schematic summary of the causal model is shown in Fig. 1.

If we further assume $U_0^{\mathtt{C}} = U_1^{\mathtt{C}}$, then we have $Y_1^{\mathtt{C}} = \mathtt{f}\left(Y_0^{\mathtt{C}}\right)$, i.e., the natural drift map can be estimated by regress the control group at $t = 1$ on the control group at $t = 0$. However, it requires coupled observations $(Y_0^{\mathtt{C}}, Y_1^{\mathtt{C}})$, which may be not available in practical applications, e.g., single-cell RNA-Seq data.

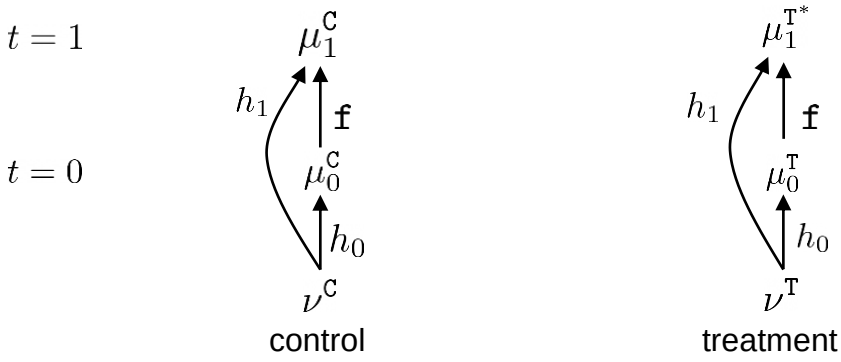

Figure 1: **Schematic of the multivariate CiC model**. The latent distributions $\nu^{\mathtt{C}}$ and $\nu^{\mathtt{T}}$ are time-invariant. At each time t, the production function h_t is applied to each latent distribution to produce the corresponding outcome distribution in the absence of intervention. Consequently, the natural drift map $\mathtt{f} = h_1 \circ h_0^{-1}$ dictates the evolutions of the outcome distributions in both control and treatment groups in the absence of intervention. This generalizes the parallel trend assumption in the classical DiD model. For estimating treatment effects on the treatment group, our goal is to estimate $\mu_1^{\mathtt{T}^*}$, the counterfactual distribution of outcomes in the treatment group at $t = 1$. One can estimate $\mathtt{f}$ from the observable empirical versions of $\mu_0^{\mathtt{C}}$ and $\mu_1^{\mathtt{C}}$ then use this estimation of $\mathtt{f}$ to push-forward the observable empirical version of $\mu_0^{\mathtt{T}}$ to obtain an estimation of $\mu_1^{\mathtt{T}^*}$.

Remark 1 *For $x \in \mathbb{R}^d$, denote x^i $(i = 1, \cdots, d)$ as the i-th coordinate of x. The multivariate CiC model is equivalent to the naive tensorization of univariate CiC models when (i) each production function $h_t : \mathbb{R}^d \mapsto \mathbb{R}^d$ $(t \in \{0, 1\})$ can be decomposed as $h_t(x) = [h_t^1(x^1) \cdots h_t^d(x^d)]^T$ for univariate functions $h_t^i : \mathbb{R} \mapsto \mathbb{R}$, and (ii) each coordinate of the latent variable (in both control and treatment groups) is independent. Therefore, it is difficult for the tensorization of univariate CiC models to express complex, multivariate natural drifts.*

The CiC causal model and optimal transport. For the uncoupled observations $Y_0^{\mathtt{C}}, Y_1^{\mathtt{C}}$, given $d = 1$, the original CiC estimator (Athey and Imbens, 2006) assumes that h_0 and h_1 are *monotone increasing*. Let $F_t^{\mathtt{C}}$ be the cumulative distribution function (cdf) of $Y_t^{\mathtt{C}}$ for $t = 0, 1$. Then, the unique monotone increasing *natural drift map* is given by $(F_1^{\mathtt{C}})^{-1} \circ F_0^{\mathtt{C}}$ such that $\mu_1^{\mathtt{C}} = \left((F_1^{\mathtt{C}})^{-1} \circ F_0^{\mathtt{C}}\right)_\sharp \mu_0^{\mathtt{C}}$. Interestingly, the natural drift map $\left((F_1^{\mathtt{C}})^{-1} \circ F_0^{\mathtt{C}}\right)$ is also the optimal transport map between $\mu_0^{\mathtt{C}}$ and $\mu_1^{\mathtt{C}}$. Therefore, the OT theory provides a natural framework to extend the Changes-in-Changes causal model for multivariate outcomes (Torous et al., 2021).

2.2. The OT approach for estimating the counterfactual distribution in multivariate CiC

As briefly discussed in Section 2.1, for the univariate outcome, the OT map is the natural drift map estimation for CiC causal model. Torous et al. (2021) leveraged the OT theory to extend it for the multivariate quantity of interests. We will discuss some mathematical background of OT for estimating $\mathtt{f}$ in high dimensional setting for the CiC causal model.

Let μ, ν be two probability measures supported on $\mathbb{R}^d$, the OT problem between μ an ν with squared Euclidean ground cost is defined as

$$\text{OT}(\mu, \nu) = \min_T \int_{\mathbb{R}^d} \|x - T(x)\|_2^2 \, \mathrm{d}\mu \quad \text{s.t.} \quad (T)_\sharp \mu = \nu, \tag{4}$$

where T is known as the transport plan.

For univariate model ($d = 1$). Given a probability measure μ supported on $\mathbb{R}$, we define its cumulative distribution function (cdf) F_μ as

$$F_\mu(x) = \mu((-\infty, x]). \tag{5}$$

Note that the cdf is not always invertible, since it is not strictly increasing. The pseudo-inverse of cdf $F^{-1} : [0, 1] \mapsto \mathbb{R}$ is given by

$$F^{-1}(x) = \inf \{t \in \mathbb{R} \mid F(t) \geq x\}. \tag{6}$$

When $d = 1$, the OT admits a closed-form expression for the optimal map T as follows:

$$T = (F_\nu)^{-1} \circ F_\mu. \tag{7}$$

Note that T is the unique increasing map such that $(T)_\sharp \mu = \nu$.[2] Therefore, if $\mu = \mu_0^{\mathsf{C}}$, $\nu = \mu_1^{\mathsf{C}}$, and $\mathtt{f}$ is increasing, e.g., by choosing monotone h_0 and h_1, then $\mathtt{f}$ is the OT map between μ_0^{C} and μ_1^{C}.

For multivariate model ($d > 1$). The natural drift map $\mathtt{f}$ can be estimated via the OT map between μ_0^{C} and μ_1^{C} (Torous et al., 2021).[3] However, for high-dimensional space, OT suffers a few drawbacks: (i) computational complexity, i.e., super cubic $\mathcal{O}(n^3 \log(n))$ where n is the number of supports of input measures; (ii) high sample complexity, i.e., $\mathcal{O}(n^{-1/d})$, which requires too many samples to precisely estimate the OT between two continuous distributions (e.g., $\mu_0^{\mathsf{C}}, \mu_1^{\mathsf{C}}$).

A naive approach to estimate the natural drift map in the multivariate CiC causal model is to decompose the model for each dimension. Specifically, one treats the multivariate CiC causal model with d-dimensional outcomes ($d > 1$) as d independent univariate CiC causal models, a.k.a., the tensorization of univariate CiC causal models. It is then efficient to estimate univariate OT maps for these univariate CiC causal models via their closed-form expressions of the corresponding univariate OT problems as in the original CiC causal model (Athey and Imbens, 2006) instead of solving the high-complexity full OT problem for measures supported in high-dimensional spaces (Torous et al., 2021). One then tensorizes all the estimated univariate OT maps to create an estimation of the multivariate OT map. However, this approach might fail to capture the dependence structure among dimensions of the multivariate OT map, i.e., the natural drift map $\mathtt{f}$, and thus when one uses this tensorized map to pushforward samples of μ_0^{T}, one might produce a counterfactual distribution with a wrong dependence structure (e.g., see Fig. 2).

In this work, we propose an efficient approach to leverage the advantages of the univariate CiC causal model for the multivariate CiC by seeking a robust latent 1-dimensional space for OT estimation. More precisely, our approach is inspired by the subspace robust OT (Paty and Cuturi, 2019) whose authors proposed to estimate the OT map in low-dimensional subspace to reduce the sample complexity for the OT problem, and further increase the robustness of OT estimation with respect to noise.

2. The optimal condition for OT (i.e., existence of the optimal map T) was described in (Gangbo, 1999) for $d = 1$.
3. The optimal condition for OT (with $d > 1$) was derived in (Brenier, 1991).

3. The proposed method based on robust OT over latent one-dimensional subspaces

In order to efficiently leverage the advantages of the univariate CiC causal model and mitigate issues in the naive tensorization for the multivariate CiC, in this work, we propose to seek a robust latent 1-dimensional subspace as a surrogate to estimate the univariate OT map to bridge the univariate and multivariate CiC causal models. This approach is also known as *max-min robust variant of OT* (Paty and Cuturi, 2019; Deshpande et al., 2019; Le et al., 2024).

Let $\mathcal{G}_1$ be the Grassmannian of the 1-dimensional subspaces of $\mathbb{R}^d$, defined as

$$\mathcal{G}_1 = \{E \subset \mathbb{R}^d \mid \dim(E) = 1\},$$

where $\dim(E)$ is the dimension of the space E. Given two measures μ, ν supported on the space $\mathbb{R}^d$, the max-min robust OT (Paty and Cuturi, 2019) considers the maximal OT distance over all possible 1-dimensional projections of input probability measures. Denote P_E as the projector on the 1-dimensional space E, the robust OT is defined as

$$\widetilde{\text{ROT}}(\mu, \nu) = \sup_{E \in \mathcal{G}_1} \text{OT}(P_{E\sharp}\mu, P_{E\sharp}\nu). \tag{8}$$

One way to parameterize the projection on the 1-dimensional space P_E and the Grassmannian manifold $\mathcal{G}_1$ is to utilize a projected direction vector ω on the sphere centering at the origin with a radius 1 as follows:

$$\overline{\text{ROT}}(\mu, \nu) = \max_{\omega \in \mathbb{R}^d \mid \|\omega\|_2 = 1} \text{OT}(P_{\omega\sharp}\mu, P_{\omega\sharp}\nu), \tag{9}$$

where P_ω denotes the projector on the direction ω. For a support $x \in \mathbb{R}^d$, its projection by the projected direction ω is computed by $\langle x, \omega \rangle$. However, the problem $\overline{\text{ROT}}$ is non-convex, which is usually approximated by the first-order method in practical applications (Deshpande et al., 2019).[4]

Much as the sliced-Wasserstein (SW) (Rabin et al., 2011)[5] which is usually approximated by averaging over a few random directions in practical applications instead of integrating over all possible directions on the sphere, and in order to optimize the robust OT efficiently, we also seek a direction over a subset of projected directions Ω for the robust OT, defined as follows:

$$\text{ROT}(\mu, \nu) = \max_{\omega \in \Omega} \text{OT}(P_{\omega\sharp}\mu, P_{\omega\sharp}\nu), \tag{10}$$

where we construct Ω by randomly sampling k directions similar as SW. We observe that a small k, e.g., in the range of 10 to 50, provides a good balance between computation speed and accuracy. For estimating the natural drift map $\mathtt{f}$, μ and ν are chosen to be the empirical versions of μ_0^{C} and μ_1^{C} respectively. A detailed description of our algorithm is provided in Appendix F.

While our approach inherits the fast computation of the CiC estimator for univariate causal models and thus can scale up for large-scale causal inference applications, it also gives good performances in estimating the natural drift map $\mathtt{f}$. This is in line with previous observations from various applications of sliced-based OT (Rabin et al., 2011; Deshpande et al., 2019; Le et al., 2019; Nguyen et al., 2021).

4. $\overline{\text{ROT}}$ is also known as the max-sliced Wasserstein (Deshpande et al., 2019).
5. SW projects supports into 1-dimensional space and exploits the closed-form expression of the univariate OT.

Discussions. Our approach is inspired by the recent success of using sliced-Wasserstein (SW) (i.e., based on one-dimensional OT) in applications, e.g., computer vision (Kolouri et al., 2016; Lee et al., 2019; Deshpande et al., 2019; Wu et al., 2019; Nguyen et al., 2021; Naderializadeh et al., 2021), deep learning (Kolouri et al., 2019; Lee et al., 2019; Wu et al., 2019), domain adaptation (Lee et al., 2019), and machine learning (Naderializadeh et al., 2021; Rakotomamonjy and Liva, 2021; Liutkus et al., 2019). Much as SW, the one-dimensional projection in our robust OT does not offer interesting properties from a distortion perspective, but remains useful as a surrogate function to optimize the transport map in applications. From many empirical evidences in our experiments, we believe that one-dimensional projection with large distortion, and further refined under the uncertainty sets in our robust OT, remains useful as a surrogate in causal applications. Furthermore, it also follows the recent realization that solving exactly the OT problem leads to overfitting (Peyré and Cuturi, 2019, §8.4), (Le et al., 2019). Excessive efforts to approximate the ground-truth cost for OT (e.g., under noisy supports or outliers) would be self-defeating since it leads to overfitting within the optimization of OT problem itself.

Although it is possible to use more dimensions in the latent space as in Subspace Robust Wasserstein (Paty and Cuturi, 2019), our approach can be regarded as a trade-off to obtain scalable approach, since only 1-dimensional OT yields a closed-form solution, in general, this property does not hold when the dimensional is greater than one. Also, notice that OT and SW are equivalent metrics (Bonnotte, 2013, Theorem 5.1.5) (Carrière et al., 2017, Theorem 3.3). Moreover, SW is indeed a distance for input measures on original high-dimensional space (Bonnotte, 2013, Proposition 5.1.2). Therefore, leveraging the closed-form solution of OT in 1-dimension space is well ground-based in OT theory.

Remark 2 *A technical remark is that it may not exist the optimal transport map T for the OT problem in Equation (4) (i.e., the Monge formulation of OT problem). In practical applications, one can relax such the Monge problem of OT by the Kantorovich formulation of OT problem (Peyré and Cuturi, 2019), in which the optimal transport plan always exists.*[6] *Consequently, one can utilize such optimal transport plan and leverage the barycentric map (Bonneel et al., 2016) for the OT map in the Monge problem of OT.*

4. Synthetic experiments

All experiments were carried out on commodity hardware and can be reproduced by the released code.[7] From here, we will denote the naive tensorization of the univariate CiC method for estimating multidimensional counterfactual distributions simply as the CiC method.

4.1. Illustrative examples

We demonstrate our method through two 2D examples for ease of visualization. For the latent distributions ν^{C} and ν^{T}, we choose bivariate Gamma distributions in the first example and 2D Gaussian mixtures in the second example. These settings lead to an unimodal counterfactual for the Gamma case, and a multimodal counterfactual distribution for the Gaussian mixture case, as can be seen in the ground truth panels of Fig. 2. More details on experiment settings can be found in Appendix A.

6. We give a review about the Kantorovich formulation of OT problem in Appendix C.
7. https://github.com/thongphamthe/scalable-counterfactual

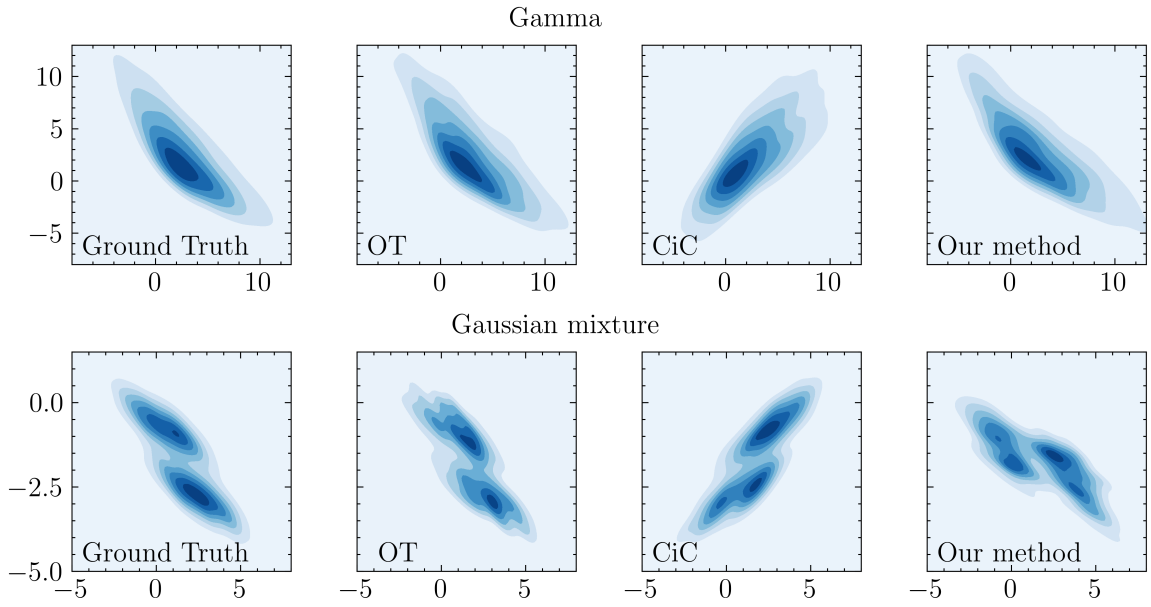

Figure 2: **Counterfactual distribution estimated by each method**. While CiC failed to capture the dependence structures between dimensions of the counterfactual distributions in both examples, our method succeeded.

The proposed method produced a counterfactual distribution with the correct correlation structure, while CiC failed to do so. This demonstrates the 1-dimensional subspace created by our method can indeed provide the type of information that CiC cannot capture. To systematically evaluate the performance of each method, we generate 10 datasets and then measure the running time as well as the OT distance between the estimated counterfactual distribution of each method and the empirical version of the true counterfactual distribution $\mu_1^{\mathrm{T}^*}$. This empirical distribution, denoted as $\widetilde{\mu_1^{\mathrm{T}^*}}$, is regarded as the ground truth in our experiments. The averaged values are reported in Fig. 2.

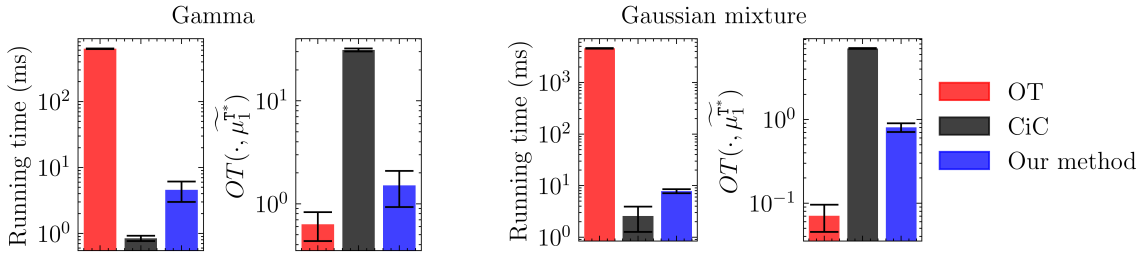

Figure 3: **Averaged running time and OT distance to ground truth**. The OT distance is measured between the estimation result of each method and the ground truth $\widetilde{\mu_1^{\mathrm{T}^*}}$, which is the empirical version of the true counterfactual distribution. Values are calculated over ten datasets. Our method outperforms the OT approach in terms of speed while being significantly better than the CiC approach in terms of accuracy measured by OT distance to ground truth.

The CiC estimator has the worst averaged OT distance to ground truth among the three methods. This might be due to its failure to capture the correlation structure between dimensions. The proposed method is about one order smaller in OT distance to ground truth than the CiC while being about two orders faster than the OT approach.

4.2. Varying the number of samples n

In this experiment, we check the findings in the illustrative examples by running the same experiment setting, with d fixed at 2, for various values of n. For each value of n, we generate 10 datasets and report the averaged running time as well as the averaged OT distance between the estimated counterfactual distribution of each method and the ground truth $\widetilde{\mu_1^{\mathrm{T}*}}$. We also add Sinkhorn as a baseline. The results are shown in Fig. 4.

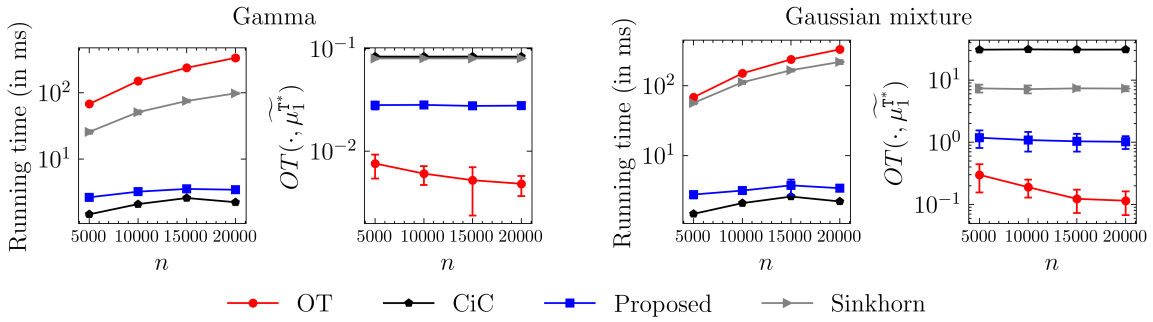

Figure 4: **Effect of varying the number of samples n for two types of latent distributions**. In both cases, the proposed method is much faster than both the OT and Sinkhorn approaches, while being much more accurate than the CiC.

The running time of our method is close to that of the CiC and is faster than OT and Sinkhorn for all values of n. This is in line with the theoretical worst-case running time of each method. Regarding OT distance to ground truth, while being worse than OT, our method outperforms CiC for all values of n. This suggests that, while a large n might help CiC in estimating the marginals of the counterfactual distribution in each dimension, the advantages of the 1-dimensional subspace constructed by our method do not diminish.

In comparison with Sinkhorn, our method is both faster and more accurate. We caution that the performance of Sinkhorn depends heavily on the strength of the entropic regularization term, and choosing a suitable value for the entropic regularization hyperparameter is non-trivial. However, as stated earlier, since the worst-case running time of Sinkhorn is quadratic, it is reasonable to expect it to be generally slower than our method. Additional results that include Sinkhorn with different hyperparameters are shown in Appendix B, and additional results with a baseline using Principal Component Analysis (PCA) are shown in Appendix E.

4.3. Varying the dimension d

In order to preserve the computational efficiency of CiC, the robust subspace in our method has to be one-dimensional. We investigate whether this subspace can still capture meaningful information in high-dimensional cases, by varying d while keeping the number of samples n fixed at 5000. In this experiment, the latent distribution is multivariate Gamma. For each d, we generate 10 datasets and

report the averaged running time as well as the averaged OT distance to ground truth. More details on experiment settings can be found in Appendix A. The results are shown in Fig. 5. Additional results that include Sinkhorn with different hyperparameters are shown in Appendix B. Additional results with a baseline using PCA are shown in Appendix E.

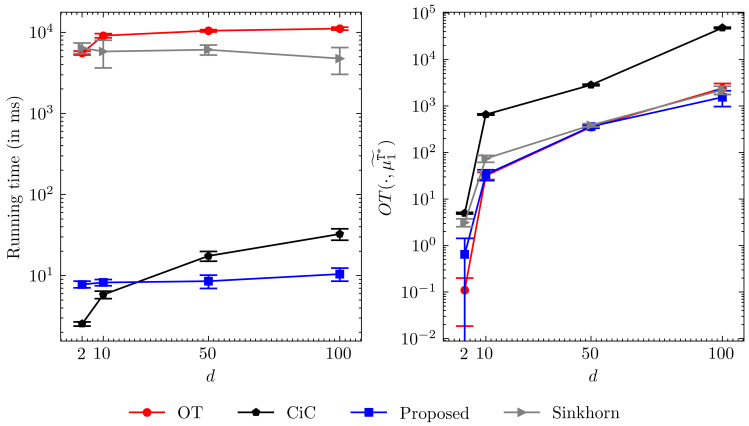

Figure 5: **Varying the dimension d while fixing number of samples** $n = 5000$. The latent distribution is multivariate Gamma. When d is high, our method is among the bests in both running time and accuracy. This suggests that the robust 1-dimensional subspace of our method can capture meaningful information even in high-dimensional situations.

It is interesting to observe that when d is high, our method is the fastest method, i.e., even faster than CiC. This is in line with the theoretical worst-case running time of CiC being $\mathcal{O}(dn \log n)$ and ours being $\mathcal{O}(n \log n + dn)$ when k is fixed. As d increases, the superiority of OT over our method in terms of OT distance to the ground truth diminishes and even reverses: our method is as accurate as, if not better than, OT when $d = 100$. Since the performance of CiC is still much worse, this reversal comes from the robust one-dimensional subspace. This is consistent with previous observations that found robustifications can alleviate the poor sample complexity of OT (Paty and Cuturi, 2019).

4.4. Varying the number of projections k

In all previous experiments, the number of projections used in constructing the robust 1-dimensional subspace has been fixed at $k = 10$. Using the same setting as the experiment above, for each d, we look at one dataset and inspect how varying k affects the quality of the method. One trade-off to expect is that increasing k might improve the accuracy of the estimation, e.g., finding a better subspace or reducing the variance between each run, at the cost of longer running time. We also investigate a closely related approach to estimate the counterfactual distribution by using the $\overline{\text{ROT}}$ objective function in Equation (9) optimized by Adam (Kingma and Ba, 2015) with various numbers of iterations. The results are shown in Fig. 6.

We found that increasing k to the hundreds indeed improves the variance and accuracy of the result of our method, as can be seen from the mean and variance of OT distance to ground truth when $d = 100$. However, these improvements are marginal and come at the cost of a linearly longer running time. This is the reason we suggest choosing a small value for k, e.g., in the range of $[10, 50]$, as a reasonable balance region between accuracy and running time. We also observe that the $\overline{\text{ROT}}$ approach with Adam offered no improvement in terms of variance and accuracy of

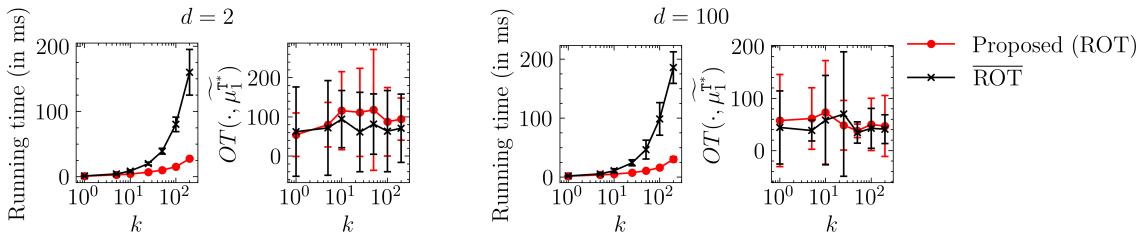

Figure 6: **Performances of our proposed method and an approach employed the $\overline{\text{ROT}}$ function when varying k.** Here k has two different meanings depending on which method is considered. In our proposed method that uses the ROT function in Eq. (10), k is the number of projections used to construct the robust one-dimensional subspace. In the approach that employs the $\overline{\text{ROT}}$ function in Eq. (9), k is the number of iterations of Adam first-order method. Our approach is as good as $\overline{\text{ROT}}$ in terms of accuracy while being much faster. Increasing the number of projections in our method to the hundreds marginally improves the accuracy and variance of the results, while requiring more time.

the results, while significantly running longer. This ineffectiveness of Adam might be due to the non-convexity and non-smooth of $\overline{\text{ROT}}$.

5. A real dataset example

We demonstrate the working of our proposed method on the classical data of Card and Krueger (1993) (CK). On April 1, 1992, New Jersey's minimum wage rose from \$4.25 to \$5.05 per hour, while Pennsylvania's did not. This provided an opportunity to estimate the causal impact of the rise on employment in fast-food restaurants in New Jersey, by analyzing employment data of New Jersey, i.e., the treatment group, and Pennsylvania, i.e., the control group, before and after the rise. In CK and subsequent re-examinations (Lu and Rosenbaum, 2004; Card and Krueger, 2000), the number of full-time employees (FT) and the number of part-time employees (PT) in each restaurant were converted to a single number, the full-time equivalent employees (FTE), which is defined as $FTE = FT + 0.5 \times PT$. The causal impact of the rise on FTE was then investigated.

This conversion may lose fine details in the characteristics of restaurants. For the same FTE, there might be a restaurant with high FT and low PT and another one with low FT and high PT, depending on each restaurant's characteristics. These restaurants might respond differently to the increase in minimum wages. Therefore, analyzing only the univariate FTE risks confounding those different trends. Simultaneous analysis of FT and PT can be expected to better capture different trends in the responses of restaurants by dissecting the causal impacts of increasing minimum wage on FT and PT. The estimation results of the 2-dimensional counterfactual distribution of FT and PT in New Jersey after the rise are shown in Fig. 7. More details on the dataset can be found in Appendix A.

We compare our method with CiC by measuring how close the estimation results of our method and CiC to the estimation result of OT, which can be regarded as the standard in terms of accuracy when one does not know the ground truth. A quick visual inspection of Fig. 7 reveals that our result captures well the relationship between PT and FT as found in the result of OT, while CiC struggles in the region where both PT and FT are large. Since there is randomness in our method,

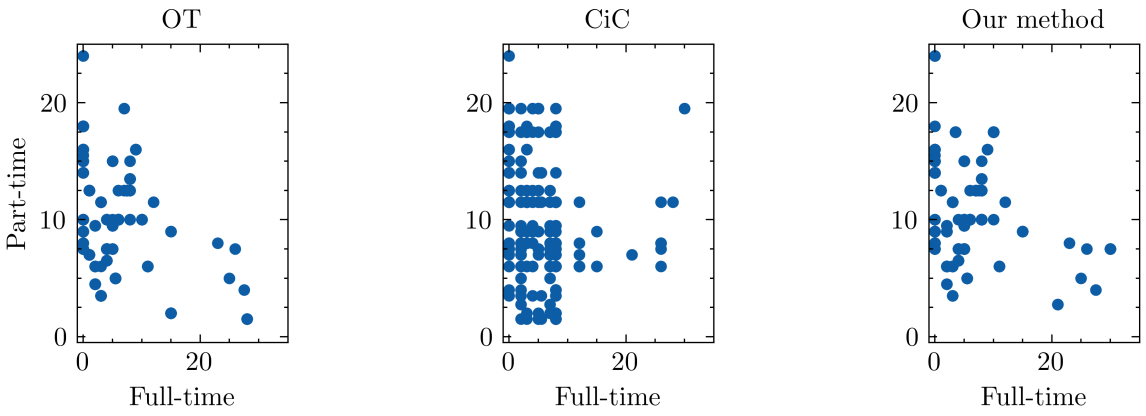

Figure 7: **Counterfactual distribution of the numbers of full-time (FT) and part-time (PT) employees in restaurants in New Jersey**. The data is from Card and Krueger (1993). The panel of our method shows the result of one typical run. Our method captures well the correlation between FT and PT in the result of OT, while CiC struggles in the region of high FT and PT. The OT distance between the result of CiC and the result of OT is 72.26. For our method, this number averaged over 1000 runs is 68.66 ± 1.42.

we measure the OT distance between our method and the result of OT, averaging over 1000 runs. The OT distance between the result of CiC and the result of OT is 72.26, while that of our method is 68.66 ± 1.42, where the confidence interval is two standard deviations. These numbers reinforce the aforementioned visual impressions and offer statistical evidence to support the conclusion that our method captured better the relationship between FT and PT than CiC.

6. Concluding remarks

We proposed a method for estimating the counterfactual distribution in multidimensional CiC models. Our method, like CiC, enjoys the computational efficiency of one-dimensional optimal transports while utilizing correlation information that is ignored under CiC. Through synthetic and real-dataset experiments, our method is shown to consistently outperform CiC in terms of accuracy, while running at a fraction of the time of the multidimensional OT approach. In future works, we plan to explore the robustness of our proposed method in the presence of outliers or noises, as well as in other causal settings, such as the triple difference model (Gruber, 1994; Olden and Møen, 2022). Additionally, it may be interesting to leverage more general local structures for scalable OT, e.g., tree and graph over supports of measures (Le et al., 2019; Le and Nguyen, 2021; Le et al., 2022, 2023), rather than the one-dimensional structure for causal inference problems.

Acknowledgments

We thank anonymous reviewers and area chairs for their comments. This work was partially supported by Shiga University Competitive Research Fund, JST CREST JPMJCR22D2, JPMJCR2015, JST MIRAI program JPMJMI21G2, ISM joint-research 2023-ISMCRP-2010, and JSPS KAKENHI Grant number 23K11243.

References

Alberto Abadie. Semiparametric difference-in-differences estimators. *The Review of Economic Studies*, 72(1):1–19, 2005. ISSN 00346527, 1467937X.

Sina Akbari, Luca Ganassali, and Negar Kiyavash. Learning causal graphs via monotone triangular transport maps. *arXiv preprint*, 2023. doi: 10.48550/arXiv.2305.18210.

Matthew Ashman, Chao Ma, Agrin Hilmkil, Joel Jennings, and Cheng Zhang. Causal reasoning in the presence of latent confounders via neural ADMG learning. In *International Conference on Learning Representations*, 2023.

Susan Athey and Guido W. Imbens. Identification and inference in nonlinear difference-in-differences models. *Econometrica*, 74(2):431–497, 2006. ISSN 00129682, 14680262.

Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust optimization*, volume 28. Princeton university press, 2009.

Dimitris Bertsimas, David B Brown, and Constantine Caramanis. Theory and applications of robust optimization. *SIAM review*, 53(3):464–501, 2011.

Richard Blundell and Monica Costa Dias. Alternative approaches to evaluation in empirical microeconomics. *Journal of Human Resources*, 44(3), 2009.

Stéphane Bonhomme and Ulrich Sauder. Recovering distributions in difference-in-differences models: A comparison of selective and comprehensive schooling. *The Review of Economics and Statistics*, 93(2):479–494, 2011. ISSN 00346535, 15309142.

Nicolas Bonneel, Gabriel Peyré, and Marco Cuturi. Wasserstein barycentric coordinates: histogram regression using optimal transport. *ACM Trans. Graph.*, 35(4):71–1, 2016.

Nicolas Bonnotte. *Unidimensional and Evolution Methods for Optimal Transportation*. Theses, Université Paris Sud - Paris XI ; Scuola normale superiore (Pise, Italie), December 2013. URL https://theses.hal.science/tel-00946781.

Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics*, 44(4):375–417, 1991.

Brantly Callaway and Tong Li. Quantile treatment effects in difference in differences models with panel data. *Quantitative Economics*, 10(4):1579–1618, 2019.

David Card and Alan B Krueger. Minimum wages and employment: A case study of the fast food industry in New Jersey and Pennsylvania. Working Paper 4509, National Bureau of Economic Research, October 1993. URL http://www.nber.org/papers/w4509.

David Card and Alan B. Krueger. Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania: Reply. *The American Economic Review*, 90(5): 1397–1420, 2000. ISSN 00028282.

Mathieu Carrière, Marco Cuturi, and Steve Oudot. Sliced Wasserstein kernel for persistence diagrams. In *International Conference on Machine Learning*, volume 70, pages 664–673, 2017.

M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, pages 2292–2300, 2013.

Ishan Deshpande, Yuan-Ting Hu, Ruoyu Sun, Ayis Pyrros, Nasir Siddiqui, Sanmi Koyejo, Zhizhen Zhao, David Forsyth, and Alexander G Schwing. Max-sliced Wasserstein distance and its use for gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10648–10656, 2019.

Sofien Dhouib, Ievgen Redko, Tanguy Kerdoncuff, Rémi Emonet, and Marc Sebban. A Swiss army knife for minimax optimal transport. In *International Conference on Machine Learning*, pages 2504–2513. PMLR, 2020.

Stephen G Donald and Kevin Lang. Inference with difference-in-differences and other panel data. *The Review of Economics and Statistics*, 89(2):221–233, 05 2007. ISSN 0034-6535.

Wilfrid Gangbo. The Monge mass transfer problem and its applications. *Contemporary Mathematics*, 226:79–104, 1999.

Markus Gangl. Causal inference in sociological research. *Annual Review of Sociology*, 36(1):21–47, 2010.

Thomas A. Glass, Steven N. Goodman, Miguel A. Hernán, and Jonathan M. Samet. Causal inference in public health. *Annual Review of Public Health*, 34(1):61–75, 2013.

Jonathan Gruber. The incidence of mandated maternity benefits. *The American Economic Review*, 84(3):622–641, 1994. ISSN 00028282.

Inwoo Hwang, Yunhyeok Kwak, Yeon-Ji Song, Byoung-Tak Zhang, and Sanghack Lee. On discovery of local independence over continuous variables via neural contextual decomposition. In *2nd Conference on Causal Learning and Reasoning*, 2023.

Alexander Immer, Christoph Schultheiss, Julia E Vogt, Bernhard Schölkopf, Peter Bühlmann, and Alexander Marx. On the identifiability and estimation of causal location-scale noise models. In *International Conference on Machine Learning*, pages 14316–14332, 2023.

Lee H Keel, SP Bhattacharyya, and Jo W Howze. Robust control with structure perturbations. *IEEE Transactions on Automatic Control*, 33(1):68–78, 1988.

Edward H. Kennedy, Sivaraman Balakrishnan, and Larry Wasserman. Semiparametric counterfactual density estimation. *arXiv preprint*, 2021. doi: 10.48550/arXiv.2102.12034.

Kwangho Kim, Jisu Kim, and Edward H. Kennedy. Causal effects based on distributional distances. *arXiv preprint*, 2021. doi: 10.48550/arXiv.1806.02935.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Klaus-Rudolf Kladny, Julius von Kügelgen, Bernhard Schölkopf, and Michael Muehlebach. Deep backtracking counterfactuals for causally compliant explanations. *arXiv preprint*, 2023. doi: 10.48550/arXiv.2310.07665.

Samantha Kleinberg and George Hripcsak. A review of causal inference for biomedical informatics. *Journal of Biomedical Informatics*, 44(6):1102–1112, 2011. ISSN 1532-0464.

Soheil Kolouri, Yang Zou, and Gustavo K. Rohde. Sliced Wasserstein kernels for probability distributions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5258–5267, 2016.

Soheil Kolouri, Phillip E. Pope, Charles E. Martin, and Gustavo K. Rohde. Sliced Wasserstein auto-encoders. In *International Conference on Learning Representations*, 2019.

Tam Le and Truyen Nguyen. Entropy partial transport with tree metrics: Theory and practice. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 130 of *Proceedings of Machine Learning Research*, pages 3835–3843. PMLR, 2021.

Tam Le, Makoto Yamada, Kenji Fukumizu, and Marco Cuturi. Tree-sliced variants of Wasserstein distances. In *Advances in Neural Information Processing Systems*, pages 12283–12294, 2019.

Tam Le, Truyen Nguyen, Dinh Phung, and Viet Anh Nguyen. Sobolev transport: A scalable metric for probability measures with graph metrics. In *International Conference on Artificial Intelligence and Statistics*, pages 9844–9868. PMLR, 2022.

Tam Le, Truyen Nguyen, and Kenji Fukumizu. Scalable unbalanced Sobolev transport for measures on a graph. In *International Conference on Artificial Intelligence and Statistics*, pages 8521–8560. PMLR, 2023.

Tam Le, Truyen Nguyen, and Kenji Fukumizu. Optimal transport for measures with noisy tree metric. In *International Conference on Artificial Intelligence and Statistics*, 2024.

Michael Lechner. The estimation of causal effects by difference-in-difference methods. *Foundations and Trends® in Econometrics*, 4(3):165–224, 2011. ISSN 1551-3076.

Chen-Yu Lee, Tanmay Batra, Mohammad Haris Baig, and Daniel Ulbricht. Sliced Wasserstein discrepancy for unsupervised domain adaptation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10277–10287, 2019.

Antoine Liutkus, Umut Simsekli, Szymon Majewski, Alain Durmus, and Fabian-Robert Stöter. Sliced-Wasserstein flows: Nonparametric generative modeling via optimal transport and diffusions. In *International Conference on Machine Learning*, pages 4104–4113, 2019.

Bo Lu and Paul R Rosenbaum. Optimal pair matching with two control groups. *Journal of Computational and Graphical Statistics*, 13(2):422–434, 2004.

Diego Martinez-Taboada and Edward Kennedy. Counterfactual density estimation using kernel Stein discrepancies. In *International Conference on Learning Representations*, 2024.

Robert J. McCann. Existence and uniqueness of monotone measure-preserving maps. *Duke Mathematical Journal*, 80(2):309 – 323, 1995.

Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Normalizing flows for interventional density estimation. In *International Conference on Machine Learning*, pages 24361–24397, 2023.

Jun Morimoto and Kenji Doya. Robust reinforcement learning. *Advances in Neural Information Processing Systems*, pages 1061–1067, 2001.

Navid Naderializadeh, Joseph F Comer, Reed Andrews, Heiko Hoffmann, and Soheil Kolouri. Pooling by sliced-Wasserstein embedding. In *Advances in Neural Information Processing Systems*, volume 34, pages 3389–3400, 2021.

Trung Nguyen, Quang-Hieu Pham, Tam Le, Tung Pham, Nhat Ho, and Binh-Son Hua. Point-set distances for learning representations of 3D point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10478–10487, 2021.

Andreas Olden and Jarle Møen. The triple difference estimator. *The Econometrics Journal*, 25(3): 531–553, 03 2022. ISSN 1368-4221.

Kishan Panaganti, Zaiyan Xu, Dileep Kalathil, and Mohammad Ghavamzadeh. Robust reinforcement learning using offline data. In *Advances in Neural Information Processing Systems*, volume 35, pages 32211–32224, 2022.

François-Pierre Paty and Marco Cuturi. Subspace robust Wasserstein distances. In *International Conference on Machine Learning*, pages 5072–5081, 2019.

Judea Pearl. Causal inference in statistics: An overview. *Statistics Surveys*, 3:96 – 146, 2009.

Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Bernot. Wasserstein barycenter and its application to texture mixing. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 435–446, 2011.

Alain Rakotomamonjy and Ralaivola Liva. Differentially private sliced Wasserstein distance. In *International Conference on Machine Learning*, pages 8810–8820, 2021.

James M Robins and Andrea Rotnitzky. Inference for semiparametric models: Some questions and an answer - comments. *Statistica Sinica*, 11(4):920–936, 2001.

Jonathan Roth and Pedro H. C. Sant'Anna. When is parallel trends sensitive to functional form? *Econometrica*, 91(2):737–747, 2023.

Jonathan Roth, Pedro H.C. Sant'Anna, Alyssa Bilinski, and John Poe. What's trending in difference-in-differences? A synthesis of the recent econometrics literature. *Journal of Econometrics*, 235 (2):2218–2244, 2023. ISSN 0304-4076.

Filippo Santambrogio. *Optimal transport for applied mathematicians*. Birkäuser, 2015.

Andreas W.M. Sauter, Erman Acar, and Vincent Francois-Lavet. A meta-reinforcement learning algorithm for causal discovery. In *2nd Conference on Causal Learning and Reasoning*, 2023.

Bernhard Schölkopf. Causality for machine learning. *Probabilistic and Causal Inference: The Works of Judea Pearl*, page 765–804, 2022.

John Snow. The cholera near Golden-square, and at Deptford. *Medical Times and Gazette*, 9: 321–322, 1854.

John Snow. On the mode of communication of cholera. page 162.1, 1855.

Tamar Sofer, David B. Richardson, Elena Colicino, Joel Schwartz, and Eric J. Tchetgen Tchetgen. On negative outcome control of unobserved confounding as a generalization of difference-in-differences. *Statistical Science*, 31(3):348 – 361, 2016.

Masayuki Takayama, Tadahisa Okuda, Thong Pham, Tatsuyoshi Ikenoue, Shingo Fukuma, Shohei Shimizu, and Akiyoshi Sannai. Integrating large language models in causal discovery: A statistical causal approach. *arXiv preprint*, 2024. doi: 10.48550/arXiv.2402.01454.

William Torous, Florian Gunsilius, and Philippe Rigollet. An optimal transport approach to causal inference. *arXiv preprint*, 2021. doi: 10.48550/ArXiv.2108.05858.

Ruibo Tu, Hedvig Kjellstrom, Kun Zhang, and Cheng Zhang. Optimal transport for causal discovery. In *International Conference on Learning Representations*, 2022.

Cédric Villani. *Topics in Optimal Transportation*. American Mathematical Society, 2003.

Cédric Villani. *Optimal transport: Old and New*. Springer, 2008.

Ted Westling and Marco Carone. A unified study of nonparametric inference for monotone functions. *The Annals of Statistics*, 48(2):1001 – 1024, 2020.

Jonas Bernhard Wildberger, Siyuan Guo, Arnab Bhattacharyya, and Bernhard Schölkopf. On the interventional Kullback-Leibler divergence. In *2nd Conference on Causal Learning and Reasoning*, 2023.

Jiqing Wu, Zhiwu Huang, Dinesh Acharya, Wen Li, Janine Thoma, Danda Pani Paudel, and Luc Van Gool. Sliced Wasserstein generative models. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3708–3717, 2019.

Huan Xu, Constantine Caramanis, and Shie Mannor. Robustness and regularization of support vector machines. *Journal of Machine Learning Research*, 10(7), 2009.

Appendix A. Details on experiment settings

A.1. The illustrative examples

The latent distributions in two examples are as follows. In the bivariate Gamma example, the first dimension of ν^{C} is a Gamma distribution with shape 2 and scale 3, while the second dimension is Gamma with shape 3 and scale 2. For ν^{T}, the first and second dimensions are reversed. In the Gaussian mixture example, the first and second dimensions of ν^{C} are $0.5\mathcal{N}(1,1) + 0.5\mathcal{N}(5,1)$ and $0.5\mathcal{N}(2,1) + 0.5\mathcal{N}(4,1)$, respectively. For ν^{T}, the first and second dimensions are reversed.

The production functions h_0 and h_1 are $h_0(u) = \left(\begin{smallmatrix} 1 & 0.5 \\ 0.5 & 1 \end{smallmatrix}\right)u$ and $h_1(u) = \left(\begin{smallmatrix} 1 & -0.5 \\ -0.5 & 1 \end{smallmatrix}\right)u$ for a length-2 vector u. These functions are co-monotone, and thus the natural drift map $\mathtt{f}$ is identifiable (Torous et al., 2021).

A.2. Settings for the experiments with varying d

We discuss the settings for the production functions h_0 and h_1. When $d \geq 2$, for the natural drift map $\mathtt{f}$ to be identifiable, the functions h_0 and h_1 need to be co-monotone (Torous et al., 2021), i.e.,

$$\langle h_0(x) - h_0(y), h_1(x) - h_1(y) \rangle \geq 0, \quad \forall x, y \in \mathbb{R}^d.$$

When $h_0(x) = \mathbf{H}_0 x$ and $h_1(x) = \mathbf{H}_1 x$ for $\mathbb{R}^{d \times d}$ matrices $\mathbf{H}_0$ and $\mathbf{H}_1$, this condition is satisfied if $\mathbf{H}_0^T \mathbf{H}_1$ is positive semi-definite. We generate one matrix $\mathbf{H}_0$ as a $d \times d$ matrix where each off-diagonal entry is uniformly distributed in $(0, 1)$ and the diagonal entries are 1. Note that $\mathbf{H}_0$ generated this way is almost surely invertible. We then generate a diagonal matrix $\mathbf{B}$ where each diagonal entry is uniformly distributed in $(0, 1)$. We then let $\mathbf{H}_1 = (\mathbf{H}_0^{-1})^T \mathbf{B}$. This will ensure that $\mathbf{H}_0^T \mathbf{H}_1$ is equal to $\mathbf{B}$ and thus positive semi-definite. We then fix the pair ($\mathbf{H}_0$, $\mathbf{H}_1$) and then generate datasets using this pair.

The latent distributions are as follows. For the control group, each dimension independently follows a Gamma distribution with shape 2 and scale 3. For the treatment group, each dimension independently follows a Gamma distribution with shape 3 and scale 2.

A.3. CK data

The dataset is available on https://davidcard.berkeley.edu/data_sets/njmin.zip. We also include the following covariates into the analysis: HRSOPEN, OPEN, NMGRS, NREGS, INCTIME, PSODA, and PENTREE, and estimate the 9-dimensional counterfactual distribution. We process the data by removing samples that contain missing values at any covariates. The final numbers of samples after this pre-processing are 57 for the control group and 220 for the treatment group.

Appendix B. Additional results with different hyperparameters of Sinkhorn

In Sinkhorn, an entropic regularization term $\lambda\Omega(T)$, where $\Omega(T)$ is the entropy of the transportation plan T, is added to the objective function of OT. The hyperparameter $\lambda \geq 0$ controls the strength of the regularization. The results reported in Figs. 4 and 5 are obtained using $\lambda = 30$. We report in the following figures two more cases when $\lambda = 10$ and $\lambda = 90$.

Appendix C. Brief review

In this section, we further give brief review for some technical details on optimal transport (OT) which are used in our work.

Kantorovich formulation of OT. Given two probability distributions μ, ν with a cost function c, the Kantorovich formulation of OT is as follow:

$$\mathrm{OT}(\mu, \nu) = \inf_{\pi \in \Pi(\mu, \nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \pi(x, y) c(x, y) d\mu(x) d\nu(y),$$

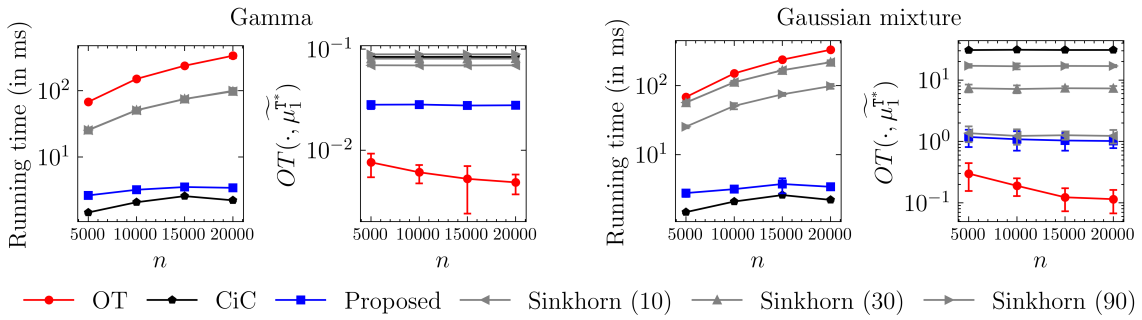

Figure B.1: Additional results of Sinkhorn for the case of varying the number of samples n while fixing the dimension $d = 2$.

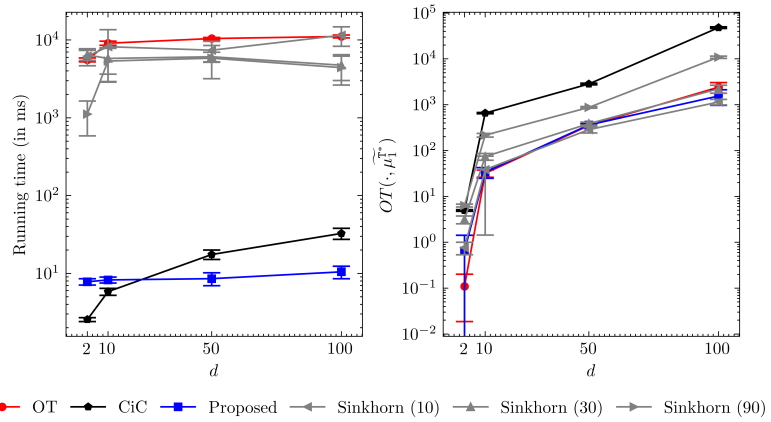

Figure B.2: Additional results of Sinkhorn for the case of varying the dimension d while fixing number of samples $n = 5000$.

where π is known as the transport plan, and $\Pi(\mu, \nu)$ is the set of all probability distributions on the product space $\mathbb{R}^d \times \mathbb{R}^d$ such that its first and second marginals equal to μ, ν respectively.

Optimal condition for OT with $d = 1$. If μ, ν are two probability measures supported on $\mathbb{R}$, and μ is atomless (i.e., μ is absolutely continuous with respect to the Lebesgue measure), then there exists at least a transport map T such that $T_\sharp \mu = \nu$ (Santambrogio, 2015, Lemma 1.27). With quadratic cost, the transport map will be the derivative of a convex function, i.e., a nondecreasing map (see (Santambrogio, 2015, Remark 1.23) and (Gangbo, 1999, §2) for further details).

Optimal condition for OT with $d > 1$. It is also known as Brenier theorem (Brenier, 1991). Given two probability measures μ, ν supported on the $\mathbb{R}^d$ space such that μ is absolutely continuous with respect to the Lebesgue measure. Then, for all possible map $\widetilde{T} : \mathbb{R}^d \mapsto \mathbb{R}^d$ such that $\widetilde{T}_\sharp \mu = \nu$, there is an unique Brenier map T which is the gradient of a convex function, and T is the optimal transport in the following sense: the Kantorich formulation of OT between μ and ν admits a unique optimal transportation plan π^* such that $(x, y) \sim \pi^*$ if and only if $x \sim \mu$ and $y = T(x)$, μ-a.s.

For further details, please see (Santambrogio, 2015, Theorem 1.17) and (Brenier, 1991; McCann, 1995).

Appendix D. Further discussions

About the assumptions of the CiC model. The assumptions of the CiC model rule out some settings such as group-time random measurement errors (Athey and Imbens, 2006, Section 3.1) or the case when the outcome distribution (under no treatment) is a mixture between a time-dependent distribution and a group-dependent distribution (Roth and Sant'Anna, 2023, Remark 4). Some components of the natural drift assumption might be tested if one observes more than one control group or more than one baseline time (Athey and Imbens, 2006, Section 6.3).

About the max-min robust OT. Max-sliced Wasserstein (Deshpande et al., 2019) is similar to our proposed method, i.e., they consider the sphere for the uncertainty set Ω of projections. However, when Ω is a sphere, the problem becomes non-convex and non-smooth. The entropic regularized OT (Cuturi, 2013) is a popular approach to reduce the computation of OT into quadratic computation complexity, but it comes with a trade-off for a dense optimal transport plan.

Some related causal models and methods. Beside the CiC model, another notable extension of the classical DiD model is provided in Abadie (2005). Some other assumptions to model the natural drift have been proposed in, for example, Callaway and Li (2019), Roth and Sant'Anna (2023), and Bonhomme and Sauder (2011). When one can only collect data at a post-treatment time, there are numerous existing methods for estimating the multivariate counterfactual distribution in a single time-step setting. These methods employ various approaches, such as kernel smoothing (Robins and Rotnitzky, 2001; Kim et al., 2021), finite-dimensional modelling with f-divergences or Lp norms (Westling and Carone, 2020; Kennedy et al., 2021; Melnychuk et al., 2023), and potentially unnormalized density function estimation with kernel Stein discrepancies (Martinez-Taboada and Kennedy, 2024). It is intriguing to investigate how these approaches can be modified to take into account some kind of natural drifts when one can observe the data at more than one time point.

About the use of machine learning methods in causal inference and causal discovery Recently, Tu et al. (2022) provided the first method that uses optimal transport for solving causal discovery, i.e., the task of uncovering the graph that represents the dependence relationships between variables. Akbari et al. (2023) generalized the method to higher dimensions and different noise settings. Takayama et al. (2024) provided a framework to utilize large language models in causal discovery algorithms. Other machine learning tools and frameworks have also been used in causal inference/causal discovery, such as neural networks (Hwang et al., 2023; Ashman et al., 2023; Immer et al., 2023; Kladny et al., 2023), the Kullback-Leibler divergence (Wildberger et al., 2023), and reinforcement learning and meta learning (Sauter et al., 2023).

Appendix E. Additional experiments for a PCA-based baseline

We explore another baseline that uses the first principal component of PCA as the sole candidate direction for the uncertainty set in our robust OT. The results for the experiments in Fig. 4 are shown in Tabs. E.1 and E.2. The results for the experiments in Fig. 5 are shown in Tab. E.3.

Table E.1: Results for the case of Gaussian mixtures in Fig. 4. The mean and standard deviation over 10 repetitions are reported.

Sample size	5000	10000	15000	20000
Running time of PCA-based method	4.71 (0.34)	8.98 (0.66)	13.24 (2.52)	15.10 (0.52)
Running time of proposed method	7.07 (0.29)	10.35 (0.50)	12.32 (0.75)	11.73 (0.43)
$OT(\cdot, \widetilde{\mu_1^{T^*}})$ of PCA-based method	0.70 (0.04)	0.70 (0.03)	0.68 (0.03)	0.68 (0.03)
$OT(\cdot, \widetilde{\mu_1^{T^*}})$ of proposed method	0.79 (0.07)	0.80 (0.05)	0.76 (0.05)	0.77 (0.04)

Table E.2: Results for the case of Gamma distributions in Fig. 4. The mean and standard deviation over 10 repetitions are reported.

Sample size n	5000	10000	15000	20000
Running time of PCA-based method	4.71 (0.2)	10.26 (1.5)	12.44 (1.03)	14.92 (0.42)
Running time of proposed method	7.63 (0.92)	10.02 (0.38)	14.29 (3.36)	11.69 (0.98)
$OT(\cdot, \widetilde{\mu_1^{T^*}})$ of PCA-based method	1.09 (0.21)	1.05 (0.11)	1.02 (0.12)	1.05 (0.16)
$OT(\cdot, \widetilde{\mu_1^{T^*}})$ of proposed method	1.18 (0.19)	1.09 (0.19)	1.04 (0.17)	1.02 (0.12)

Table E.3: Results for Fig. 5, i.e., when d is varying. The mean and standard deviation over 10 repetitions are reported. This experiment shows that while our method and the PCA-based method give comparable results, our method is much faster when the dimension is high.

Dimension d	2	10	50	100
Running time of PCA-based method	27.2 (10.77)	23.82 (3.38)	111.94 (9.02)	188.24 (18.84)
Running time of proposed method	7.83 (0.36)	8.19 (0.36)	8.49 (0.83)	10.24 (0.73)
$OT(\cdot, \widetilde{\mu_1^{T^*}})$ of PCA-based method	0.46 (0.16)	31.95 (2.48)	341.04 (8.70)	1394.59 (73.12)
$OT(\cdot, \widetilde{\mu_1^{T^*}})$ of proposed method	0.65 (0.39)	33.72 (4.36)	362.7 (14.43)	1549.3 (287.29)

These results show that it is sufficing to use random directions as in our robust OT which gives comparable results, but lower computational cost (e.g., finding the first principal component for high-dimensional spaces increases the computational cost, but its benefits are marginal.)

Appendix F. The algorithm for our proposed approach

Our approach consists of four steps as follows.

- **compute the push-forward for input measures on each direction in the uncertainty set**. Specifically, for each direction ω in the uncertainty set, we calculate the inner product $\langle x, \omega \rangle$ for all points x in the empirical distribution of the control group at time $t = 0, 1$.

- **leverage 1d-OT as a surrogate on the 1d latent space for OT**. For each direction ω in the uncertainty set, we solve the 1D OT problem with push-forward measure on ω of the outcome distribution in control group at $t = 0$ as source and push-forward measure on ω of the outcome distribution in control group at $t = 1$ as target. Thus for each direction ω, we will obtain a mapping and an OT distance.
- **find the direction which maximizes the OT on the 1d latent space**. We choose the direction ω^* that gives the largest value of the OT distances calculated above.
- **Use the optimal transport map on $\mathbb{R}^d \times \mathbb{R}^d$ to create the counterfactual distribution in treatment group**. We use the OT map corresponding to ω^* to push-forward each sample in the empirical outcome distribution in treatment group at time $t = 0$. The result is the counterfactual distribution in treament group at time $t = 1$.