Research Article

William Torous*, Florian Gunsilius, and Philippe Rigollet

An optimal transport approach to estimating causal effects via nonlinear difference-in-differences[#]

Abstract: We propose a nonlinear difference-in-differences (DiD) method to estimate multivariate counterfactual distributions in classical treatment and control study designs with observational data. Our approach sheds a new light on existing approaches like the changes-in-changes estimator and the classical semiparametric DiD estimator, and it also generalizes them to settings with multivariate heterogeneity in the outcomes. The main benefit of this extension is that it allows for arbitrary dependence between the coordinates of vector potential outcomes and includes higher-dimensional unobservables, something that existing methods cannot provide in general. We demonstrate its utility on both synthetic and real data. In particular, we revisit the classical Card & Krueger dataset, which reports fast food restaurant employment before and after a minimum wage increase. A reanalysis with our methodology suggests that these restaurants substitute full-time labor with part-time labor on aggregate in response to a minimum wage increase. This treatment effect requires estimation of the multivariate counterfactual distribution, an object beyond the scope of classical causal estimators previously applied to this data.

1 Introduction

Difference-in-differences (DiD) estimators are among the most widely used approaches to estimate causal effects of discrete treatment interventions in observational data [1–4]. The classical setting we considered in this article is one in which the researcher observes the outcomes of interest in a group of treated units and a corresponding control group of untreated units over time, i.e., a binary treatment setting.[1] At one fixed point in time, the treatment group undergoes an intervention and stays treated thereafter. The fundamental problem consists of isolating the causal effect of this intervention from the existing *natural trend* that the treatment

---

[#] A previous version was titled "An optimal transport approach to causal inference."

[1] The proposed method can straightforwardly be extended to finitely many treatments. We do not address the question of continuous treatments in this article.

---

* Corresponding author: William Torous, Department of Statistics, University of California, Berkeley, CA 94720, United States of America, e-mail: wtorous@berkeley.edu

Florian Gunsilius: Department of Economics, University of Michigan, Ann Arbor, MI 48109, United States of America

Philippe Rigollet: Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA 02139, United States of America

group undergoes irrespectively. The method of DiD achieves this under the assumption that the units in the control group have the same fundamental trend as the treated units would have without intervention. Under this assumption, the difference of the average outcome of the treated group between preintervention and postintervention periods net the difference of the control groups average outcomes preintervention and postintervention manages to isolate the true average causal effect of the intervention.

The arguably two most fundamental methods for estimating causal effects in this setting are the classical method of DiD [1,2,5] as described earlier and its generalization beyond average and aggregate effects, the changes-in-changes (CiC) estimator [6].

DiD is a linear approach designed to estimate average (or aggregate) causal effects [1,7] over all units within a group. This method can be proved to correctly identify counterfactuals under the "parallel trends" assumption [1,8,9] under which the *average* natural trend is assumed to be the same across both the control and treatment groups. This idea has been used in many areas of science where capturing the average treatment effect is sufficient to formulate informative causal conclusions. Recent applications include quantifying the effect of public health measures in response to COVID-19 [10] and estimating how irrigation farmers adapt their watering to fixed usage limits [11].

The classical linear DiD estimator only provides estimates of average and aggregate treatment effects. This can be limiting in many settings where treatment heterogeneity is important, i.e., where different units can react differently to the same level of intervention. The CiC estimator [6] addresses this issue for univariate outcomes. It extends the fundamental idea of the DiD estimator by estimating changes in the entire probability distribution of the respective units within a group. This permits the estimation of the entire counterfactual law of the treated units had they not received the treatment, hence allowing for a general form of heterogeneity in their response to treatment. The cost is a slightly stricter assumption on the evolution of the unobservable distribution as we specify below.

In its current form, the CiC estimator is only applicable in settings with univariate outcomes, as it relies heavily on the definition of quantile functions. This can be restrictive for modern applications where the outcomes of interest are often multivariate. Examples range from A/B testing in digital marketing campaigns where outcomes are a combination of features such as click-through rate, time-per-page, and so on, to measuring intervention effects such as gene knock down on a population of cells measured in high-dimensional gene space. In principle, the CiC estimator may be extended to higher dimensions through tensorization, which estimates treatment effects independently for each coordinate, but this solution fails to capture correlations that are often important in causal discovery. We demonstrate a simple linear setting with synthetic data where this is the case in Section 4.1.

Our contribution. We recast the CiC estimator using tools from the theory of optimal transportation. This perspective allows us to extend this methodology to handle multivariate observables after introducing novel structural assumptions on the natural trend underlying both treatment and control groups. In particular, we note that the causal model proposed in Athey and Imbens [6] readily implies that both populations, treatment and control, evolve between pretreatment and posttreatment periods via optimal transport maps (Theorem 2); our methodology relies on a natural higher-dimensional extension of this via optimal transport theory. In fact, we show that if the corresponding functions mapping the unobservables to the potential outcomes are *cyclically comonotone*, then the counterfactual distributions and causal effects can be estimated from data. In particular, this allows for the unobservable variables to be of the same dimension as the dimension of the outcome variables and not just univariate.

The counterfactual distributions can be estimated consistently from data using recent results from statistical optimal transport and implemented efficiently using recent advances in computational optimal transport [12]. We demonstrate the benefits of this extension by comparing it to the simple multivariate extension of the CiC estimator on both artificial and real data. As illustrated in Section 4.1, a tensorized version of CiC can be inconsistent and even estimate opposite correlation structures in the multivariate distribution of counterfactual outcomes. We also revisit the classical dataset of Card and Krueger [13], which has sparked an intense debate about the effects of raising the minimum wage on employment. Our ability to jointly handle the number of part-time and full-time workers as a bivariate outcome reveals an interesting substitution effect

from the perspective of labor economics: When the minimum wage is raised, restaurants increase full-time labor and decrease part-time labor, suggesting a substitution effect.

2 The causal model behind the CiC estimator

2.1 The CiC model

We recall in this section the general causal model on which the classical DiD and CiC estimators are built, and which our model extends. See Figure 1 for illustration.

We first formalize a stochastic model for an experimental design in which two groups are measured before and after some intervention. Sampled from a larger population, each unit i in a realized experiment is characterized by four vector potential outcomes, their level of treatment received, and a set of time points at which potential outcomes were observed. This model includes randomized control trials with noncompliance and general observational studies. Adapting the notation of Athey and Imbens [6], we define an indicator random variable G_i for a unit's adopted treatment arm as well as random vectors $Y_{i;C,0}$ and $Y_{i;C,1}$ to model unit i's potential control group preintervention and postintervention outcomes observable when $G_i = 0$; likewise $Y_{i;T,0}$ and $Y_{i;T,1}$ are the unit's observable potential outcomes when $G_i = 1$. These potential outcomes are assumed to lie in a subspace of $\mathbb{R}^d$, and each unit has indicator random variables $T_{i,0}$ and $T_{i,1}$ denoting whether an outcome was observed in each study period.

We assume that each potential outcome vector is generated by a deterministic function of a latent random vector also in $\mathbb{R}^d$. While any given unit's latent vector may change over time, the population distribution of continuous latent variables is time invariant. Furthermore, these latent distributions can differ arbitrarily between control and treatment arms. We denote this latent distribution ν for controls and $\nu^\star$ for treated. Therefore, we write for control units the latent random vectors $U_{i,t} \overset{d}{\sim} \nu$, $t = 0, 1$ and treated $U_{i,t} \overset{d}{\sim} \nu^\star$. We can think of $U_{i,t}$ as capturing all of the intrinsic and unobservable characteristics of unit i, such as an employee's skills and motivation, which influence the outcome of interest at time t, for instance, their salary. This assumption may be less appropriate in settings where external shocks could change the latent distributions over time. For instance, in a study measuring individuals' opinion of the economy, a latent distribution of intrinsic perception reasonably shifts over time. Our model can accommodate exogenous noise through the assumption that latent $U_{i,t}$ is the additive perturbation of a unit's time invariant baseline latent state μ_i by a stationary noise process $\varepsilon_{i,t}$. Correlation between these two terms is allowed as long as the distribution of
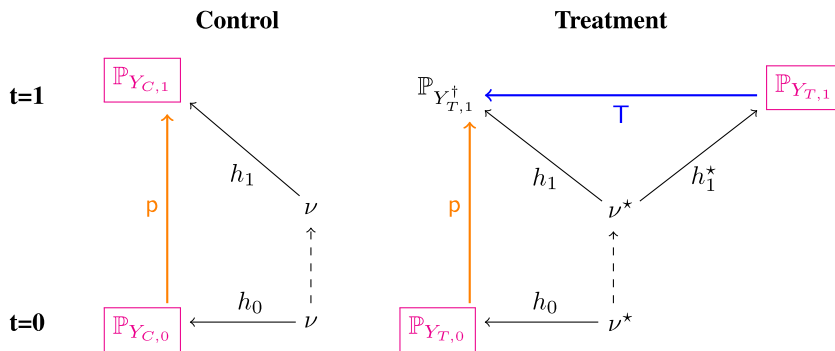

Figure 1: Illustration of various maps in the space of measures. An arrow indicates a pushforward map between two measures; for example $\mathbb{P}_{Y_{C,1}} = \mathsf{p}_\# \mathbb{P}_{Y_{C,0}}$. The maps h_j are the "production functions" linking the unobservables ν and ν^* to the potential outcomes. A dashed arrow indicates a map from a measure to itself. $\mathbb{P}_{Y^\dagger_{T,1}}$ is the counterfactual outcome measure of the treated units had they not received treatment. p is the natural trend map and T is map from an observed outcome to its counterfactual. The observable data is drawn from the four boxed measures.

realized latent variables is time invariant within each treatment arm. These latent variable assumptions are general enough to include many forms of mechanistic structure researchers may know about their population.

As Figure 1 illustrates, three global *production functions* h_0, h_1, and $h_1^\star$ characterize the evolution of both treatment arms over time. At $t = 0$, the treatment arms are assumed to have identical factors influencing the translation of a latent variable to observed potential outcome and thus share h_0. For controls, h_1 describes any changes due the natural trend over evolving time. For treated units, $h_1^\star$ captures both drift over time and the causal effect of interest. If $h_1 \circ (h_0^{-1})$ exists, it describes exactly how any given control group unit evolves over time with a fixed latent state. We can therefore obtain a counterfactual outcome distribution with this function under an assumption that the treatment arm units would have evolved in the absence of treatment just as controls do. It is notable that this proposed model does not rely on additional information from observed covariates. As such, it does not require classical unconfoundedness assumptions. However, the question about the assignment mechanism in these DiD setups is a delicate one, see, for instance, studies by Ghanem et al. [14] and Marx et al. [15].

As units in the same treatment arm share a latent variable distribution, we define random vectors with the population distribution such that $Y_{i;C,0} \overset{d}{\sim} Y_{C,0}$ for all units such that $G_i = 0$. The random vectors $Y_{C,1}$, $Y_{T,0}$, and $Y_{T,1}$ are defined similarly for observations adopting the appropriate treatment arms. Each population-distributed random variable of potential outcomes induces a measure or *law*, which we will utilize in application of results from optimal transport theory. We adopt the notation $\mathbb{P}_{Y_{G,T}}$ for this measure. The action of each production function can be described as a *pushforward* of measure. For example, we write $\mathbb{P}_{Y_{C,0}} = h_{0\#}\nu$ and $\mathbb{P}_{Y_{T,1}} = h_{1\#}^\star \nu^\star$. We make an assumption that both $\mathbb{P}_{Y_{C,0}}$ and $\mathbb{P}_{Y_{T,0}}$ are absolutely continuous with respect to Lebesgue measure. In Section 3.1, we discuss possible changes to our assumptions, which would allow our methodology to handle discrete outcomes at the cost of point identification of the counterfactual distribution. We note that the discrete case in general does not allow for point identification, which is already evident in the univariate setting, where Athey and Imbens [6] provide bounds on discrete counterfactuals. Our optimal transport framework illuminates this issue even further.

In addition, assuming that h_0 is invertible, we have $\nu = (h_0^{-1})_\# Y_{C,0}$ so that

$$\mathbb{P}_{Y_{C,1}} = (h_1 \circ h_0^{-1})_\# \quad \mathbb{P}_{Y_{C,0}} =: \mathsf{p}_\# \mathbb{P}_{Y_{C,0}}. \tag{1}$$

In the sequel, we will propose a class of production functions such that the multivariate counterfactual distribution for the treated, distributed as a random variable $Y_{T,1}^\dagger$, can be identified as follows:

$$\mathsf{p}_\# \mathbb{P}_{Y_{T,0}} := \mathbb{P}_{Y_{T,1}^\dagger}. \tag{2}$$

A consistent estimator for p and hence this counterfactual pushforward is presented utilizing optimal transport. Figure 1 shows that the production function $h_1^\star$, which generates postintervention outcomes for the treated, does not need to be specified to recover the map p and its counterfactual pushforward $Y_{T,1}^\dagger$. As discussed in Section 3.2, further assumptions on $h_1^\star$ enable the treatment effect map T to be consistently estimated and allow for the estimation of nonlinear treatment effects.

2.2 Modeling the control group time trend

Given two measures $\mathbb{P}_{Y_{C,0}}$ and $\mathbb{P}_{Y_{C,1}}$ which can be estimated from control group data preintervention and postintervention, there exist infinitely many maps p such that $\mathbb{P}_{Y_{C,1}} = \mathsf{p}_\# \mathbb{P}_{Y_{C,0}}$ (1) holds. To identify a unique such map p, one must make additional assumptions on the causal model.

Recall that the source of randomness in the proposed causal model are latent random vectors $U_{i,t}$, which can be thought of as summarizing a unit's intrinsic characteristics. If we not only assume that the distribution ν of control group latent vectors is time invariant but further that the latent variables are constant over time (i.e., $U_{i,0} = U_{i,1}$ for all control units i), then we have $Y_{i;C,1} = \mathsf{p}(Y_{i;C,0})$; this map can be estimated via multivariate regression from independent copies of the pair $(Y_{C,0}, Y_{C,1})$. In many practical applications, pairs of potential outcomes $(Y_{i;C,0}, Y_{i;C,1})$ are observed, but in general, the assumption that $U_{i,0} = U_{i,1}$ for all control units is more

tenuous. Moreover, this setup leaves out an important class of problems where the units in the control group can differ between $t = 0$ and $t = 1$. A canonical example arises in genomics where in single-cell RNA-Seq data each unit corresponds to a cell. Measuring the gene expression levels of a cell is a destructive process, and as a result, a given cell may be only measured once. Therefore, coupled potential outcomes of the form $(Y_{i;C,0}, Y_{i;T,1})$ are not available. Instead, one observes independent copies of $Y_{C,0}$ as well as independent copies of $Y_{C,1}$. While difficult, this problem can be solved at the cost of additional assumptions on the natural trend p. This problem is sometimes referred to as *uncoupled regression*, and it arises in various applications, far beyond the genomics example described earlier [see, e.g., 16–18].

In a setting with uncoupled scalar data, the map p is unique when it is monotone with known monotonicity (either increasing or decreasing). The CiC estimator enforces this by assuming both production functions, and therefore, $h_1 \circ h_0^{-1}$ are monotone increasing. In fact, the unique monotone increasing function p such that $\mathbb{P}_{Y_{C,1}} = \mathsf{p}_{\#}\mathbb{P}_{Y_{C,0}}$ is given by $F_1^{-1} \circ F_0$, where F_t denotes the cumulative distribution function (CDF) of $Y_{C,t}$ for $t = 0, 1$. Estimation of p from data and plugging that estimator into equation (2) to yield an estimated counterfactual measure is the key idea of the CiC estimator. As previously noted, this estimator can be extended to multivariate outcomes through tensorization, which applies it independently to each coordinate. Joint structure between the coordinates is incompatible with the scalar distribution and quantile functions used in the estimator, as well as with the assumption of scalar production functions.

Recently, in the univariate setting, other approaches for modeling the control group time trend have been proposed, which generalize the "parallel trends" assumption to allow for general heterogeneity. These models in general make stronger or less interpretable assumptions than the monotone production functions of Athey and Imbens [6]. Callaway and Li [19] directly assume that the copulas between $\mathbb{P}_{Y_{C,0}}$ and $\mathbb{P}_{Y_{T,0}}$ as well as $\mathbb{P}_{Y_{T,0}}$ and $\mathbb{P}_{Y_{T,1}^{\dagger}}$ are the same; Roth and Sant'Anna [8] assume that the pointwise differences between the corresponding cumulative distribution functions are equal, i.e., that $F_1(x) - F_0(x) = F_1^{\dagger}(x) - F_0^{\star}(x)$ for all $x \in \mathbb{R}$; Bonhomme and Sauder [20] restrict the heterogeneity of the model to be additively separable and assume that the pointwise differences between the corresponding logarithms of the characteristic functions are equal. In the next section, we introduce cyclical comonotonicity, a natural extension of the monotonicity requirement in the study by Athey and Imbens [6] to multivariate settings, and hence, the multivariate notion of parallel trends closest to the framework of that work.

In the following section, we show that under the CiC model, the univariate natural trend map p is unique, and therefore identifiable as an optimal transport map between $\mathbb{P}_{Y_{C,0}}$ and $\mathbb{P}_{Y_{C,1}}$. We leverage this insight to identify unique natural trend maps in higher dimensions under the same model. In particular, we propose an assumption of cyclically comonotonicity, a multivariate notion of monotonicity between the production functions, which extends the CiC estimator's assumption of monotone increasing production functions.

3 Higher dimensional CiC

3.1 An optimal transport extension of the CiC estimator

Identifiability of the causal model presented in the previous section hinges on the uniqueness of the monotone increasing map p such that $\mathsf{p}_{\#}\mathbb{P}_{Y_{C,0}} = \mathbb{P}_{Y_{C,1}}$. This uniqueness follows from a fundamental result in the theory of optimal transportation known as Brenier's theorem.

Theorem 1. [Brenier, 21] *Let P and Q be two probability measures defined over $\mathbb{R}^d$ such that P is absolutely continuous with respect to the Lebesgue measure. Then, among all maps $T : \mathbb{R}^d \to \mathbb{R}^d$ such that $Q = T_{\#}P$, there is a unique one, called the Brenier map denoted $\bar{T}$, which is the gradient of a convex function. Furthermore, $\bar{T}$ is an optimal transport map in the following sense.*

Let Γ denote the set of joint probability distributions of $(X, Y) \in \mathbb{R}^d \times \mathbb{R}^d$ such that $X \sim P$ and $Y \sim Q$, then the optimal transport problem

$$\inf_{\gamma \in \Gamma} \int \|x - y\|^2 \mathrm{d}\gamma(x, y) \tag{3}$$

admits a unique solution $\bar{\gamma}$ such that $(X, Y) \sim \bar{\gamma}$ if and only if $X \sim P$ and $Y = \bar{T}(X)$, P-almost surely.

For $d = 1$, a map T is the gradient of a convex function if and only if it is nondecreasing. The quantile map $F_{Y_{C,1}}^{-1} \circ F_{Y_{C,0}}$ has the appropriate domain and range measures and is nondecreasing by the properties of CDFs; therefore, it is the unique optimal transport solution.

Theorem 1 demonstrates that the structural properties of the univariate CiC model can be extended to higher dimensions by leveraging the theory of optimal transportation. In particular, the Brenier map is unique and hence identifiable between pairs of sufficiently regular measures in $\mathbb{R}^d$. We therefore introduce regularity assumptions on the production functions $h_{\cdot,\cdot}$ in the model, which imply that the optimal drift is the Brenier map between $\mathbb{P}_{Y_{C,0}}$ and $\mathbb{P}_{Y_{C,1}}$. The counterfactual distribution $\mathbb{P}_{Y_{T,1}^\dagger}$ in Figure 1 can be identified by first estimating the unique natural trend p between the observable measures $\mathbb{P}_{Y_{C,0}}$ and $\mathbb{P}_{Y_{C,1}}$ in the control group and then applying that transformation to the treatment group preintervention potential outcome measure $\mathbb{P}_{Y_{T,0}}$.

For our main result, we assume that all four the potential outcome measures $\mathbb{P}_{Y_{\cdot,\cdot}}$ and the counterfactual measure $\mathbb{P}_{Y_{T,1}^\dagger}$ are absolutely continuous with respect to the Lebesgue measure. This assumption for preintervention measures is needed to apply Theorem 1, and for the postintervention measures, it helps enable identification. To satisfy technical assumptions on the regularity of the optimal transport problem that often arise in classical results [22], we enforce through Assumption 1 that measures are supported on convex sets. Furthermore, if the support $K_0^\star$ of $\mathbb{P}_{Y_{T,0}}$ is not contained in the support K_0 of $\mathbb{P}_{Y_{C,0}}$, one cannot infer the counterfactual distribution $\mathbb{P}_{Y_{T,1}^\dagger}$ at the parts of $K_0^\star$ that fall outside K_0 without extrapolation via structural assumptions. This restriction on the supports is a standard assumption that prevents the need for extrapolation [6]. We summarize our structural assumptions as follows:

Assumption 1. The observable measures $\mathbb{P}_{Y_{C,t}}, \mathbb{P}_{Y_{T,t}}, t = 0, 1$, and the counterfactual measure $\mathbb{P}_{Y_{T,1}^\dagger}$ are supported on proper convex subsets K_t, $K_t^\star$, and $K_1^\dagger$ of $\mathbb{R}^d$ and are absolutely continuous with respect to Lebesgue measure. Moreover, $K_0^\star \subset K_0$.

When p is an optimal transport map, Theorem 1 implies that p is the gradient of a convex function. Recall that in higher dimensions, gradients of convex functions from $\mathbb{R}^d \to \mathbb{R}$ are *cyclically monotone* [23, Theorem 24.8]. A map $T : \mathbb{R}^d \to \mathbb{R}^d$ is cyclically monotone if for any positive integer m and any cycle $u_1, \ldots, u_m, u_{m+1} \equiv u_1$ in its domain, it holds

$$\sum_{i=1}^m \langle u_i, T(u_i) - T(u_{i+1}) \rangle \geq 0. \tag{4}$$

Therefore, the natural trend map is identifiable when it is cyclically monotone. For $m = 2$, equation (4) reduces to the simple notion of multivariate monotonicity defined as follows:

$$\langle u_1 - u_2, T(u_1) - T(u_2) \rangle \geq 0. \tag{5}$$

Furthermore, in the univariate case, these two definitions are equivalent; in the CiC model, this implies that the monotone map $h_1 \circ h_0^{-1}$ is also cyclically monotone, which explains why the assumption of monotonicity of the functions h in the unobservable is sufficient.

In the context of our latent model, the monotonicity assumption on the production functions for the CiC estimator implies that larger values of the unobservable latent variable correspond to strictly larger potential outcomes. This is a common assumption in economic models [24–26], but can be restrictive in some settings. For instance, consider a classroom experiment where the treatment variable is class size and the outcomes of interest are standardized test scores before the intervention and 5 years after. Monotone production functions

imply that individuals with higher unobserved ability U will receive strictly higher test scores. Specifically, the assumption of being able to rank individuals in terms of one univariate unobservable U is a strong restriction.

A careful inspection of the data generating process reveals that monotone assumptions on the production functions allow the latent variable to be entirely abstracted away. This actually follows from the weaker notion of comonotonicity such that $h_0, h_1 : \mathbb{R}^d \to \mathbb{R}^d$ are comonotone if $\langle h_0(x) - h_0(y), h_1(x) - h_1(y) \rangle \geq 0$ for all $x, y \in \mathbb{R}^d$. Note that with an identity production function, comonotonicity reduces to classical monotonicity. To get a sense of the strength of this relaxation, consider univariate differentiable functions. In this case, comonotonicity implies locally that their derivatives $h_0' \cdot h_1' \geq 0$ have the same sign and furthermore imposes additional global constraints. As a concrete example, if h_0 is a polynomial production function and h_1 its pointwise scaling by some $\gamma > 0$, then this pair of functions is comonotone because they have the same sign between all zeros, and hence, the same signed difference between any pair of points. This example emphasizes that h_0 and h_1 need not be individually monotone themselves.

We have now seen that that the monotone production functions assumption in Athey and Imbens [6] implies that the natural trend is cyclically monotone, which enables identification of the counterfactual. Furthermore, the production functions are comonotone, which allows the model to include general latent space distributions. However, these two functional assumptions only collapse to monotonicity in the scalar case. To extend their model to higher dimensions, we propose an assumption of cyclically comonotone production functions, which achieves both these desired properties.

Definition 1. Two production functions h_0 and h_1 are *cyclically comonotone* if for any positive integer m and any cycle $u_1, \ldots, u_m, u_{m+1} = u_1$ in their common domain, it holds

$$\sum_{i=1}^{m} \langle h_0(u_i), h_1(u_i) - h_1(u_{i+1}) \rangle \geq 0. \tag{6}$$

Just as cyclical monotonicity collapses to monotonicity when $m = 2$, cyclical comonotonicity collapses to comonotonicity in that case. Whenever h_0 has an inverse h_0^{-1}, condition (6) implies the map $\mathsf{p} = h_1 \circ h_0^{-1}$ is also cyclically monotone. This result is shown in the proof of Theorem 2, which we will state shortly. By Theorem 1, p is the unique Brenier map such that $\mathbb{P}_{Y_{C,1}} = \mathsf{p}_{\#}\mathbb{P}_{Y_{C,0}}$.

Cyclical comonotonicity is one of the several concepts to extend comonotonicity to higher dimensions. One set of extensions imposes a total ordering on the outcome space; however, these imply that the copula of outcomes preintervention and postintervention are identical, a strong assumption on the causal model. Note that the tensorized CiC assumes this shared copula structure. More recently, optimal transportation theory has been used to propose other multivariate notions of comonotonicity. A particular contribution is μ-comonotonicity from Ekeland et al. [27]. Similar to our model, the authors assume a latent variable and production functions mapping it to potential outcomes. Both production functions h_0 and h_1 are assumed to have optimal transport structure, as they lie in the graph of a gradient of a convex function. In general, the composition of optimal transport maps, such as $h_1 \circ h_0^{-1}$, is not an optimal transport map, except in special circumstances [28]. Puccetti and Scarsini [29] provide an overview of these ideas.

Cyclical monotonicity is a natural extension of monotonicity to multivariate settings with many uses in economic theory and mechanism design [30,31], econometrics [32,33], and statistics [34]. Cyclical comonotonicity is even weaker than this, as it does not imply that each production function itself is cyclically monotone, which makes it a natural and weak extension of the monotonicity assumption in the univariate setting. In particular, it is weaker than the assumption that individuals can be ranked in terms of a univariate unobservable variable (ability in the classroom example) and that this ranking persists in the observed outcome distribution. We argue that this makes it a more natural assumption than the monotonicity assumption of the CiC estimator in practical settings. Examples of cyclic monotonicity in causal inference have been highlighted by Chernozhukov et al. [33], Shi et al. [32], and Gunsilius [35], among many others. More fundamentally, Rochet [31] shows that the maximizer of quasi-linear utility functions is cyclically monotone. For a concise survey, we refer to the study by Galichon [36].

From a practical perspective, cyclical comonotonicity is an important weakening of cyclic monotonicity in that it allows for unobservables of general dimension in practice. Allowing for multivariate unobserved

heterogeneity in general causal models has been a long-standing challenge: in most models, the unobservable $U_{i,t}$ is required to be univariate [6,24,26]. By relying on the concept of cyclical comonotonicity, we can relax this assumption by only requiring that the unobservable is of the same dimension as the observable variable. Note that even this can be relaxed by using recent advances in optimal transport theory [37]. The cost of this relaxation is a more complicated model and definitions from geometric measure theory like the area and coarea formula, which is why we focus on the simpler setting of unobservables of the same dimension in this article.

3.1.1 Discrete outcome distributions

Mirroring the univariate setting [6], our methodology does not guarantee a point identified-counterfactual distribution when outcomes are discrete valued. Theorem 1 guarantees a unique control group natural trend with optimal transport structure when the measure of preintervention outcomes is absolutely continuous with respect to the Lebesgue measure. Indeed, unique optimal transport maps between discrete measures need not exist. For illustration, consider transporting a measure distributed on vertices of the unit square with equiprobability at each main diagonal vertex to a one with equiprobability at the antidiagonal vertices. With respect to squared distance, it is equivalent to transport the mass between vertical and horizontal pairs of vertices. Furthermore, it is admissible to split mass and transport 0.25 probability mass from each main diagonal vertex to each antidiagonal vertex. This solution with fractional structure is an example of a non-deterministic optimal transport *plan* (cf. deterministic optimal transport *maps*), which can be interpreted as a probabilistic assignment rule. Discrete-valued optimal transport problems often have a minimum cost achieved by multiple probabilistic plans but no deterministic maps. This reality implies that point-identification in the discrete case need not be possible in general without restrictive further assumptions.

    This optimal transport view hence also explains the lack of point-identification in the univariate setting, matching the conclusions of Athey and Imbens [6]. The extension of CiC to discrete outcomes in that setting provides bounds on the counterfactual outcome distribution due to the same challenges described earlier. Their estimator for that case is heavily dependent on notions of quantile functions without a direct multivariate extension. Identification only becomes possible with additional assumptions about conditional independence and exogenous covariates, which the authors admit are restrictive. In the discrete multivariate setting, similar partial identification results are possible through carefully application of recent developments in optimal transport theory. For example, Auricchio and Veneroni [38] bound the supremum cost of moving any single point mass in ℓ^p-cost transport problems. With one possible counterfactual distribution identified by a solution to the minimization, uniform bounds can then be found by applying the aforementioned bound on the possible pointwise deviation. Pointidentification can also be reestablished in a similar manner to the univariate setting by introducing additional information from covariates under strong assumptions. The focus of this article is to introduce an optimal transport framework for pointidentification, which is why we focus on the case of absolutely continuous measures throughout.

3.2 Identification of causal effects in the multivariate setting

On the basis of the extensions of the causal model introduced in the previous sections, we can formally state the identification result for the counterfactual distribution of the treated units had they not received treatment.

Theorem 2. *Consider the causal model depicted in* Figure 1. *Let Assumption* 1 *hold. Moreover, assume that the production function h_0 has a well-defined inverse and that h_0 and h_1 are cyclically comonotone in the sense of* (6). *Then there exists a unique map* $\mathsf{p} : K_0^\star \to K_1^\dagger$. *It is the Brenier map from* $\mathbb{P}_{Y_{C,0}}$ *to* $\mathbb{P}_{Y_{C,1}}$. *The counterfactual distribution* $\mathbb{P}_{Y_{T,1}^\dagger}$ *of the treated unit had it not received treatment is then identified via* $\mathbb{P}_{Y_{T,1}^\dagger} = \mathsf{p}_\# \mathbb{P}_{Y_{T,0}}$.

The proof of Theorem 2 is presented in Appendix A.1. Another potential quantity of interest is the actual counterfactual random variable $Y^{\dagger}_{T,1}$. Identifying this quantity requires more structure in the model. Indeed, in the setup of the causal model represented in Figure 1, we only have that $\mathbb{P}_{Y_{C,1}} = \mathsf{p}_{\#}\mathbb{P}_{Y_{C,0}}$ but not necessarily that $Y_{C,1} = \mathsf{p}(Y_{C,0})$ because we do not assume[2] that $U_{i,0} = U_{i,1}$ for all units. However, assuming that both $h^{\dagger}_1$ and h_1 are cyclically comonotone allows to identify $Y^{\dagger}_{T,1}$, which follows from the fact that $T : K^{\star}_1 \to K^{\dagger}_1$ will be identified for the fixed unobservable $U_{i,1}$ for treated units.

Corollary 1. *Consider the setting and assumptions from Theorem* 2. *If the production function h_1 has a well-defined inverse and $h^{\star}_1$ and h_1 are cyclically comonotone, then there exists a unique map $T : K^{\star}_1 \to K^{\dagger}_1$ such that $\mathbb{P}_{Y^{\dagger}_{T,1}} = T_{\#}\mathbb{P}_{Y_{T,1}}$. T is the Brenier map from $\mathbb{P}_{Y_{T,1}}$ to $\mathbb{P}_{Y^{\dagger}_{T,1}}$. $Y^{\dagger}_{T,1}$ is then identified via $Y^{\dagger}_{T,1} = T(Y_{T,1})$.*

The assumption that h_1 and $h^{\star}_1$ are cyclically comonotone can be hard to satisfy in practice, which is why the main parameter of interest is the counterfactual law $\mathbb{P}_{Y^{\dagger}_{T,1}}$ and not the random variable $Y^{\dagger}_{T,1}$. It is worth noting that having cyclically comonotone pairs $h^{\star}_1$ and h_1 as well as h_0 and h_1 does not necessarily imply $h^{\star}_1$ and h_0 are cyclically comonotone.

In general, the assumptions of our model match those in Athey and Imbens [6] when outcomes are scalar and extend their structure in higher dimensional cases. Our production function setup illustrated in Figure 1 exactly matches that of CiC. In the scalar case, our assumption of comonotone production functions reduces to monotone production functions; although this is weaker than CiC's requirement that h_0 and h_1 are strictly monotone, the requirement of Theorem 2 that h_0 has a well-defined inverse restricts that preintervention production function to be strictly monotone as well. In one dimension, a strictly monotone and a weakly monotone production function pair remains cyclically comonotone, so our model does provides a relaxation of the control group outcome generation process. Here, it is the assumption that h_0 has an inverse, which forces the production functions to be themselves monotone, as contrasted by our previous example of two positively proportional polynomials being comonotone. In higher dimensions, however, the invertibility of h_0 does not imply that both h_0 and h_1 are themselves cyclically monotone. If this were the case, then both production functions would be optimal transport maps, which can be seen by considering cyclical monotonicity as the special case of cyclical comonotonicity when h_0 is the identity. This is precisely the notion μ-comonotonicity introduced by Ekeland et al. [27].

For the latent distributions, our extension retains the time invariance within groups assumption of CiC. While we do not make the same assumption of nested latent supports $v^{\star} \subset v$ as Athey and Imbens [6], our assumption that $K^{\star}_0 \subset K_0$ is equivalent. To see why note that both treatment arm share the same preintervention production function h_0. Our assumption that the support sets are convex is an additional requirement not needed for CiC but necessary to apply theoretical results about multivariate optimal transport. Athey and Imbens [6] allow for an additional case with discrete latent distributions, but our assumption that h_0 is invertible with a continuous co-domain implies that the latent variable measures must be continuous.

3.2.1 Examples of functionals that can be analyzed with the proposed method

One of the advantages of the method is to deal with multivariate outcome distributions. Many functionals of interest in different areas of the (social) sciences are based on the information of the joint distribution of several outcomes of interest, from social inequality [39–41] to risk management [42,43] to decision theory [44], to name a few.

---

2 We note in passing that if we were to assume that $U_{i,0} = U_{i,1}$ for all units, we could, in fact, test the assumption that the natural trend p, and hence the production functions h_0, h_1, are monotone increasing from independent copies of paired data of the form (X_0, Y_1). We would do so by checking if there exists an increasing function p that indeed interpolates these points. In practice, such a function seldom exists and allowing a time variable $U_{i,t}$ accounts for deviations from this ideal situation.

More generally, the proposed estimator provides consistent estimates of functionals that fundamentally rely on the *copula structure* of the outcome distributions. Examples of this are multivariate stochastic dominance [45], multivariate risk [46], and functionals that encompass interrelation between random variables such as mutual information or correlation. We say that a distribution F dominates another distribution G in the α-th order, written as $F \succeq G$ if $D_\alpha^F(x) \leq D_\alpha^G(x)$, where

$$D_1^F(x) \equiv F(x) \quad \text{and} \quad D_\alpha = \int_{S(x)} D_{\alpha-1}(t)\mathrm{d}t,$$

with $S(x) \coloneqq \{y \in \mathbb{R}_+^d : y \leq x\}$, see the study by García-Gómez et al. [45] for details. An example of a multivariate risk measure is the multivariate lower-orthant value-at-risk at level $\alpha \in [0, 1]$ of a multivariate increasing function G, which is defined as follows [46]:

$$\underline{\mathrm{VaR}}_\alpha(G) = \partial\{x \in \mathbb{R}^d : G(x) \geq \alpha\},$$

where ∂A denotes the boundary of the set A. An analogous definition holds for the upper-orthant counterpart. The mutual information $I(X, Y)$ of two random variables X and Y is defined as follows [47]:

$$I(X, Y) = \mathrm{KL}(P_{X,Y}||P_X \otimes P_Y),$$

where KL is the Kullback–Leibler divergence and $P_X \otimes P_Y$ is the independence coupling.

The aforementioned examples are only a short list of potentially interesting functionals one can analyze using the proposed method. All of these functionals rely on the information of the corresponding copula structure, which cannot be identified with a tensorized version of the CiC estimator unless in the limited setting of completely independent coordinates described previously.

3.3 Estimation from observed data

In practice, we only observe observations from each of the four distributions $\mathbb{P}_{Y_{C,0}}$, $\mathbb{P}_{Y_{C,1}}$, $\mathbb{P}_{Y_{T,0}}$, and $\mathbb{P}_{Y_{T,1}}$. To distinguish between population measures and empirical measures, we define the corresponding empirical measures of $\mathbb{P}_{Y_{C,0}}$ as $\hat{\mu}_{0,n}$, of $\mathbb{P}_{Y_{C,1}}$ as $\hat{\mu}_{1,l}$, and of $\mathbb{P}_{Y_{T,0}}$ as $\hat{\mu}_{0,m}^\star$ with corresponding sample sizes n, l, m. We denote the estimator of the counterfactual $\mathbb{P}_{Y_{T,1}^\dagger}$ by $\hat{\mu}_{1,m}^\dagger$. Estimating Brenier maps from data is a subject of intense activity both from theoretical and practical angles [48–57]. Further details about estimating Brenier maps from data can be found in Appendix B. In this section, we state propositions about the consistency of the estimated counterfactual distribution and its rate of convergence in expected Wasserstein distance to the population distribution. The 2-Wasserstein distance between measures μ and ν, which we denote as $W_2(\mu, \nu)$, is the square root of the infimum transportation cost (3) between the two measures. Proofs of these propositions are deferred to Appendix A.

The observations form empirical measures $\hat{\mu}_{0,n}$, $\hat{\mu}_{0,m}^\star$, and $\hat{\mu}_{1,l}$ with n, m, and l denoting the respective sample sizes. These samples are assumed to be mutually independent. An estimator for the natural trend, $\hat{\rho}$ is a minimizer of $W_2(\hat{\mu}_{0,n}, \hat{\mu}_{1,l})$. A plug-in estimator for the counterfactual distribution is given by $\hat{\rho}_{\#\hat{\mu}_{0,m}^\star}$. However, $\hat{\rho}$ is nontrivially defined only on the support of $\hat{\mu}_{0,n}$. Therefore, in practice, we project $\hat{\mu}_{0,m}^\star$ onto the support of $\hat{\mu}_{0,n}$ through one-nearest neighbor Euclidean matching, itself a 2-Wasserstein optimal transport problem. The "rounded" measure of treatment outcomes preintervention is denoted $\tilde{\mu}_{0,m}^\star$. This leads to our proposed estimator for the counterfactual distribution: $\hat{\rho}_{\#\tilde{\mu}_{0,m}^\star} \eqqcolon \hat{\mu}_{1,m}^\dagger$.

Under the assumptions on population measures stated in Assumption 1, the Wasserstein distance between the estimated and population counterfactual distribution converges uniformly to zero. This enables a consistency statement about that distance converging to probability to zero, as well as a stronger statement about the Wasserstein distance of the estimated and population natural trend maps.

Proposition 1. *Under the conditions of Assumption* 1,

$$\limsup_{\min(m,n)\to\infty} \int_{K_0^\star} \|\hat{\mathsf{p}}(\tilde{x}) - \mathsf{p}(x)\| \mathbb{P}_{Y_{T,0}}(\mathrm{d}x) = 0.$$

Furthermore, the squared 2-Wasserstein distance of the empirical natural trend map, $W_2^2(\hat{\mu}_{0,m}^\star, \hat{\mu}_{1,m}^\dagger)$, converges in probability to the distance from preintervention samples to their population counterfactual, $W_2^2(\hat{\mu}_{0,m}^\star, \mathbb{P}_{Y_{T,1}^\dagger})$.

Under mild further assumptions, we show that the expected 2-Wasserstein distance between the population and estimated counterfactual measures converges to 0 at a $O(\min(m,n)^{\frac{1}{d}})$ rate.

Assumption 2. The convex support K_0 is compact, and in addition, the Radon–Nikodym derivative $f \coloneqq \frac{\mathrm{d}\mathbb{P}_{Y_{T,0}}}{\mathrm{d}\mathbb{P}_{Y_{C,0}}}$ of $\mu_0^\star$ with respect to μ_0 is bounded.

Combining these assumptions with those previously stated for consistency under Assumption 1 yields the following claim.

Proposition 2. *The expected 2-Wasserstein distance between $\hat{\mu}_{1,m}^\dagger$ and $\mathbb{P}_{Y_{T,1}^\dagger}$ converges to 0 at a $O(\min(m,n)^{\frac{1}{d}})$, that is,*

$$\mathbb{E}_{\mathbb{P}_{Y_{C,0}},\mathbb{P}_{Y_{T,0}},\mathbb{P}_{Y_{C,1}}} W_2(\hat{\mu}_{1,m}^\dagger, \mathbb{P}_{Y_{T,1}^\dagger}) = O(\min(m,n)^{\frac{1}{d}}).$$

Obtaining the large sample distribution of optimal transport maps is largely an open problem. However, rates of convergence are known [50,57], which enables inference on these functionals through subsampling methods. Under mild assumptions (e.g., those in [58–60]), subsampling theory provides confidence regions and hypothesis tests with strong asymptotic guarantees. Even if the rate of convergence for the functional is unknown, that rate may still be estimated.

4 Numerical experiment and application to the Card and Krueger dataset

In this section, we demonstrate the performance of the optimal transport-based estimator with two numerical experiments and apply it to a classical causal inference dataset.

4.1 A stylized example

We first present a simple simulation experiment, which demonstrates that improper modeling of dependency between coordinates of multivariate outcomes can lead to highly biased estimates. In particular, our proposed method manages to estimate the correct bivariate counterfactual distributions, while the multivariate extension of the CiC estimator through tensorization does not. Consider the following set of linear production functions in which map latent vectors in $\mathbb{R}^2$ to outcomes without intervention in $\mathbb{R}^2$:

$$h_0(u) = \begin{bmatrix} 1 & \alpha \\ \alpha & 1 \end{bmatrix} u \quad \text{and} \quad h_1(u) = \begin{bmatrix} 1 & -\alpha \\ -\alpha & 1 \end{bmatrix} u.$$

These production functions are a cyclically comonotone but not element-wise monotone as required for the CiC estimator. Choosing beta distributions as the laws of the latent variables, we generate $n = 3,000$ independent outcomes from the control population before and after the intervention, as well as from the
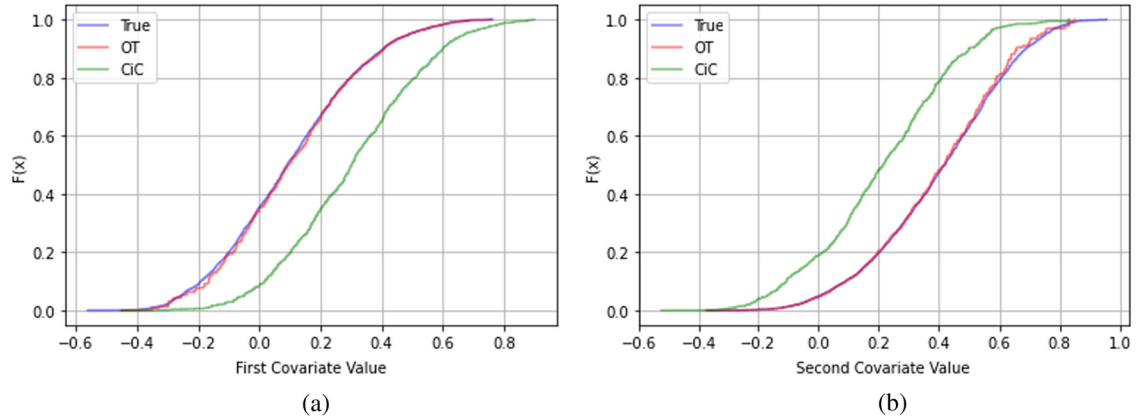
Figure 2: Recovery of counterfactual marginals by OT and CiC. (a) First marginal and (b) second marginal.
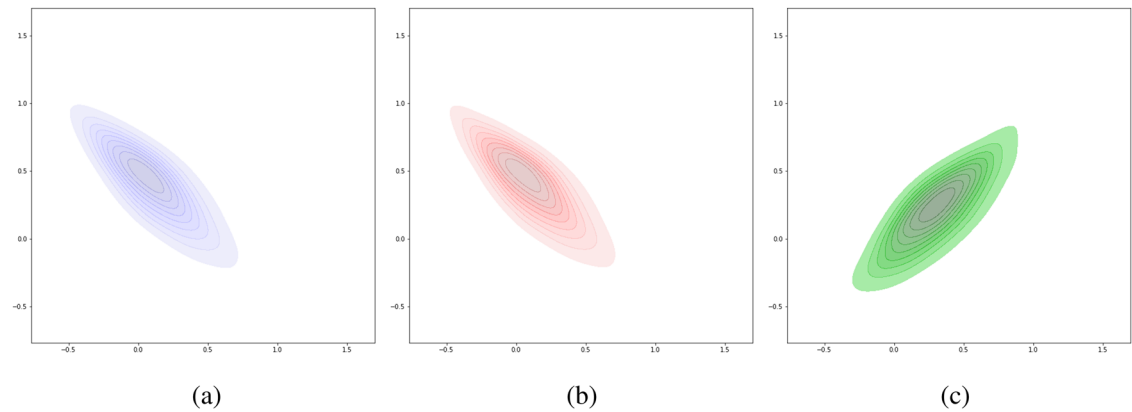

Figure 3: KDE plots of the distribution of counterfactual outcomes for the treatment group. The first covariate increases along the horizontal axis and the second along the vertical. KDE bandwidth = 0.5. (a) True CF, (b) OT estimate, and (c) CiC estimate.

treatment group before the intervention and its unobserved counterfactual. Further details about the implementation can be found in Appendix C.1.

The recovered marginals in Figure 2 and the kernel density estimator (KDE) of the counterfactual joint distribution in Figure 3 demonstrate that our methodology estimates the true counterfactual distribution almost exactly, while the CiC estimator cannot. Notably, the CiC estimator recovers a joint distribution with a mirrored dependence structure.

As previously noted, the tensorized CiC estimates counterfactuals under the strong assumption that the copula of outcomes preintervention and postintervention are identical. Each coordinate of the potential outcomes vectors is assumed to be a monotone increasing function of the latent random vector's corresponding coordinate. When there are interactions between latent coordinates during potential outcome generation, the multivariate CiC will not identify counterfactual distributions. We believe that proper accounting for correlation structures contributes to our seemingly novel substitution result when reanalyzing the classical minimum wage data from Card and Krueger [13] in Section 4.3.

4.2 Numerical experiment in higher dimensions

In the following numerical experiment, we consider the performance of our estimator against the multivariate tensorized CiC estimator as the dimensionality of the observations grows. Optimal transport's strong recovery of higher-dimensional joint distributions, illustrated by Figure 4, distinguishes it.
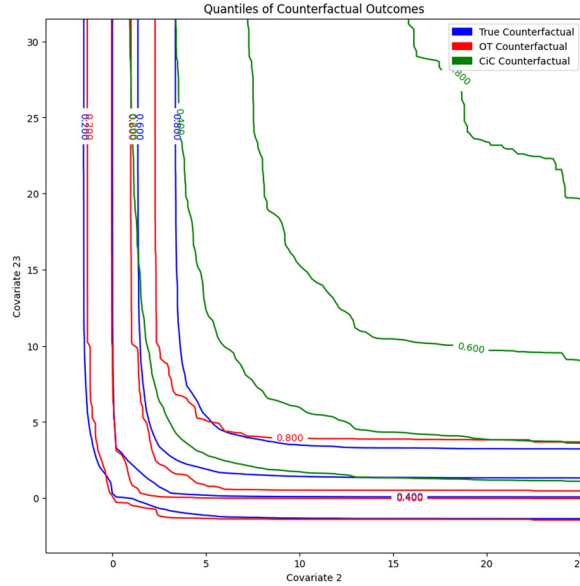
Figure 4: True and estimated eCDF quantiles for counterfactual treatment individuals, $D = 50$.

Recall from Theorem 1 that optimal transport maps can be characterized as the gradients of convex functions. Consider the function $f(x, y) = x^2 y^{-1}$, which is convex when $y > 0$. As the sum of convex functions is convex, summing f evaluated at multiple pairs of coordinates in a random vector constructs a convex function from $\mathbb{R}^d$ to $\mathbb{R}$. Formally, we consider a class convex functions $g(x; \theta) = \sum_{i=1}^p x_{(i_1)}^2 x_{(i_2)}^{-1}$, where $x \in \mathbb{R}^d_{++}$, $x_{(i)}$ represents the ith coordinate of x, and $i_j \in [[d]]$ for $j \in \{1, 2\}$; the parameter θ contains the p pairs of indices to be compared. The gradient of g is straightforward to compute. In this simulation, we specify h_0 to be the identity and h_1 to be ∇g. There are d coordinate pairs randomly selected each run. Both treatment arms have $n = 3,000$ units drawn from beta latent distributions. Further implementation details can be found in Appendix C.2.

Our experiment is designed to quantify how well the two methods estimate a joint counterfactual distribution. For computational reasons, we focus on randomly selected pair-wise coordinate interactions. To this end, we construct the eCDF of the true counterfactual, its optimal transport estimate, and its CiC estimate over a uniform mesh of 10,000 points. We compute the mean absolute difference between the true and estimated eCDFs over each point in the mesh and have 20 runs per trial. As demonstrated in Figure 4, our proposed method better estimates the quantiles of a plane in the outcome space. Table 1 shows that optimal transport has a mean absolute difference consistently smaller than CiC with a smaller standard deviation as well. Our method's monotone increasing loss likely occurs due to the curse of dimensionality and the difficulty of matching vectors in very high dimensions with only 3,000 samples. Regardless, optimal transport still has substantially smaller MAD and standard deviation than CiC when $d = 100$.

Table 1: Mean absolute deviation of the estimated counterfactual CDF to the true one. eCDFs are evaluated over one randomly selected plane per run

Dimension	2	10	20	50	100
OT MAD	0.0006	0.0017	0.0019	0.0033	0.0037
st. dev.	0.0003	0.0011	0.0016	0.0027	0.0055
CiC MAD	0.0083	0.0104	0.0092	0.0088	0.0090
st. dev.	0.0055	0.0078	0.0074	0.0073	0.0125

4.3 Revisiting the Card and Krueger dataset

On April 1, 1992, New Jersey raised its minimum wage from the Federal level of $4.25 per hour to $5.05 per hour. Card and Krueger [13] (CK henceforth) investigated the effect this increase had on employment in fast food restaurants with a case–control experimental design, taking the bordering region of eastern Pennsylvania, where the minimum wage remained at $4.25, as a control group. The original study employs the standard DiD estimator to conclude that the higher minimum wage led to increased employment in New Jersey restaurants. This result spawned much debate within the economics community about the effect of minimum wage policies on employment (e.g., Neumark and Wascher [61], Dube et al. [62], Meer and West [63], Neumark et al. [64], and the references therein).

The treatment effect of interest in CK and many subsequent reevaluations is the change in full-time equivalent employees (FTE), which is defined via FTE = FT + 0.5PT, where FT (PT) is the number of full-time (part-time) employees [61,65]. We use the proposed multivariate extension of the CiC estimator to estimate a bivariate treatment effect in terms of full- and part-time employees, hence fully dissecting the causal effect of raising the minimum wage on these subgroups.

We analyze the number of full- and part-time employees as drawn from an underlying continuous distribution. In the original CK data, "there are 28 records that report fractional full-time employees (the fraction is always one-half), 29 records that report fractional numbers of part-time employees, and one record that reports a fractional number of managers" [66]. These partial observations could represent ambiguity in the classification of a worker as full- or part-time (Alan Krueger via interview reported in by Ehrenberg et al. [66]). As the restaurant employee responding to the phone interview was not necessarily the same at both measurements, Neumark and Wascher [61] argue, "there is no reason to believe that the responses in the first and second waves are based on the same 'definition' of employment." Therefore, these fractional observations can also be considered a degree of belief statement. As discussed earlier, deterministic optimal transport maps are the unique solution in the continuous case while the discrete case also admits probabilistic optimal transport plans as solutions. In practice, we use a linear program implemented in the Python package POT which finds a deterministic plan, matching the structure of our theoretical results.

The original CK study finds a positive average treatment effect of 2.76 FTE jobs added in the treatment group compared to the control group, a result reproduced by Neumark and Wascher [61] using different methodology. Neumark and Wascher [61] also compute an average treatment effect in full- and part-time jobs with separate regressions and find treatment effects of 3.16 gained full-time jobs and 0.60 lost part-time jobs. These classical contributions only focus on aggregate outcomes over all restaurants, irrespective of the size of the respective restaurant. An exception is Ropponen [67] who reanalyzes the data with the CiC estimator, taking into account the heterogeneity in the size of the fast-food restaurants. He finds the average treatment effect bounded in [0.90, 1.70]. Ropponen also notes in New Jersey that large restaurants preintervention (measured in FTE employees) have a negative treatment effect, while small restaurants preintervention have a positive treatment effect. Our analysis recovers this trend for both FT and PT employees in New Jersey. As we also generate counterfactual controls, we find an opposite trend for Pennsylvania restaurants. This opposite trend is intriguing and worth exploring further as it does not seem to be explained by substitution effects of employees moving across the state line to find jobs (Table 2).

The previously discussed subsampling approach to inference was applied to generate symmetric 95 % confidence intervals around each point estimate. The only interval to contain zero is for the DiD treatment effect on part-time employment. Furthermore, we run subsampled hypothesis tests to check whether the point

Table 2: Estimated average treatment effect on full- and part-time employment by method with subsampled 95% confidence intervals

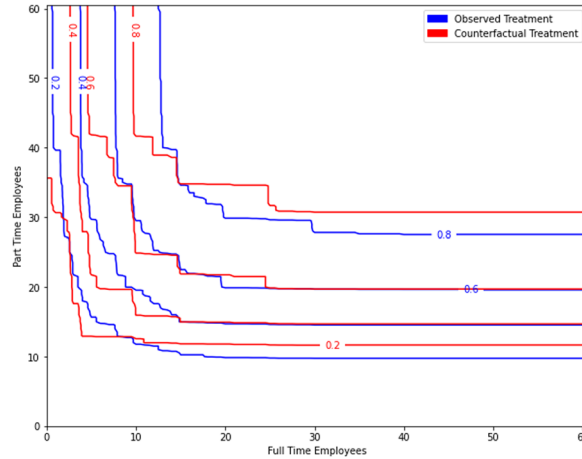|	OT	CiC	DiD
ATE FT	3.07 (2.00, 4.67)	2.61 (0.40,3.45)	3.45 (3.05,5.46)
ATE PT	−1.79 (−3.43, −0.33)	−1.52 (−3.46, −0.51)	−1.00 (−2.07, 1.15)

Figure 5: Postintervention quantile curves for observed treatment group units and their counterfactuals.

estimates are significantly different across estimators. Of the four tests between optimal transport to an alternative, only the difference in treatment effect on part-time employment estimated with CiC was not significantly different from zero at the $p = 0.05$ level. For both the confidence intervals and hypothesis tests, we use a block size of 300 and 10,000 replications.

The average treatment effects estimated using the OT and CiC counterfactual distributions depend only on the marginals, not the full joint structure. However, as illustrated in Figure A3 in Appendix, the strong copula assumptions between preintervention and postintervention periods for CiC can lead to markedly different estimated marginals. That structure implies restaurants with the largest preintervention full/part-time employment will also have the largest counterfactual full/part-time employment. The wider tails of the CiC treatment effect marginals compared to OT suggest that additional joint structure in this setting leads to counterfactuals with less variance from the observed potential outcome.

The proposed multivariate extension of the CiC estimator allows us to jointly estimate the effects on full- and part-time employment while accounting for the heterogeneity in restaurant size. For the results in this section, restaurants are represented in $\mathbb{R}^2$ by our outcomes of interest, the number of full- and part-time employees. In Appendix, we provide an additional experiment with restaurants represented as higher dimensional vectors including additional measurements about wages, prices, and operations. We discard 19 units with missing outcomes, leaving 76 control and 315 treatment restaurants. The transport plans for both groups are solved with a linear program, and we apply nearest neighbor matching between treatment and control preintervention samples to calculate counterfactual outcomes.

Our optimal transport analysis suggests that fast food restaurants responded to an increased minimum wage by substituting full-time employees for part-time ones with a net gain of FTEs. Our results seem to indicate that the negative effect on part-time employees is more pronounced than previously estimated. The quantile plot in Figure 5 suggests this effect applies throughout the distribution of restaurants, not just the mean: fixing the number of full-time employees, counterfactual quantiles tend to have more part-time employees than treatment group quantiles at the same level; likewise fixing the number of part-time employees, treatment group quantiles tend to have more full-time employees. In their original paper, CK suggest that restaurants may respond to higher labor costs with more full-time employees because they tend to be older, more skilled, and more productive, thus a better investment of capital. Furthermore, the univariate methods may underestimate the number of part-time employees lost, a conclusion supported by the higher dimensional analysis in Table A3 in Appendix.

5 Discussion

We have introduced a general method based on optimal transport theory for causal inference in observational treatment and control study designs. It combines two desirable properties: it is designed for multivariate settings while at the same time capturing the heterogeneity in treatment response of individuals. In particular, it complements both the classical DiD estimator, which is applicable in multivariate settings but only captures average effects, and the CiC estimator [6], which allows for general treatment heterogeneity but is only applicable in univariate settings.

Showcasing the utility of the proposed method, we revisit the classical Card and Krueger dataset by decomposing the treatment effects for full-and part-time employees. We find that fast-food restaurants responded to an increased minimum wage by substituting full-time employees for part-time employees, even when accounting for restaurant size. This provides a novel insight by dissecting the relationship between full- and part-time employees while accounting for the heterogeneity in restaurant size. More generally, since our proposed method is able to consider multivariate outcomes, it can help uncover interesting causal effects and nonlinear relations in a wide array of applications.

Author contributions: All authors have accepted responsibility for the entire content of this manuscript and approved its submission.

Conflict of interest: The authors state no conflict of interest.

Data availability statement: The datasets analyzed during the current study are available at David Card's public website, https://davidcard.berkeley.edu/data_sets.html.

Code availability statement: Code to reproduce the results of this paper and software implementations of our methodology in Python, R, and STATA are available at otcic.com.

References

[1] Abadie A. Semiparametric difference-in-differences estimators. Rev Econ Stud. 2005;72(1):1–19.
[2] Lechner M. The estimation of causal effects by difference-in-difference methods. Foundations and Trends® in Econometrics. 2011;4(3):165–224.
[3] Imbens GW, Rubin DB. Causal inference in statistics, social, and biomedical sciences. New York, NY: Cambridge University Press; 2015.
[4] Roth J, Sant'Anna PH, Bilinski A, Poe J. What's trending in difference-in-differences? A synthesis of the recent econometrics literature. 2022. arXiv: http://arXiv.org/abs/arXiv:220101194.
[5] Wing C, Simon K, Bello-Gomez RA. Designing difference in difference studies: best practices for public health policy research. Annual Review Public Health. 2018;39:453–69.
[6] Athey S, Imbens GW. Identification and inference in nonlinear difference-in-differences models. Econometrica. 2006;74(2):431–97.
[7] Heckman J. Varieties of selection bias. Amer Econ Rev. 1990;80(2):313–8.
[8] Roth J, Sant'Anna PH. When is parallel trends sensitive to functional form? 2020. arXiv: http://arXiv.org/abs/201004814.
[9] Ryan AM, Kontopantelis E, Linden A, Burgess Jr JF. Now trending: Coping with non-parallel trends in difference-in-differences analysis. Stat Methods Med Res. 2019;28(12):3697–711.
[10] Goodman-Bacon A, Marcus J. Using difference-in-differences to identify causal effects of COVID-19 policies. Survey Res Methods. 2020;14(2):153–8.

[11] Drysdale KM, Hendricks NP. Adaptation to an irrigation water restriction imposed through local governance. J Environ Econ Management. 2018;91:150–65.

[12] Peyré G, Cuturi M. Computational optimal transport. Foundations and Trends® in Machine Learning. 2019;11(5–6):355–607.

[13] Card D, Krueger AB. Minimum wages and employment: a case study of the fast-food industry in New Jersey and Pennsylvania. Amer Econ Rev. 1994;84(4):772–93.

[14] Ghanem D, Sant'Anna PHC, Wüthrich K. Selection and parallel trends. 2022. arXiv: http://arXiv.org/abs/220309001.

[15] Marx P, Tamer E, Tang X. Parallel trends and dynamic choices. 2022. arXiv: http://arXiv.org/abs/220706564.

[16] Rigollet P, Weed J. Uncoupled isotonic regression via minimum Wasserstein deconvolution. Inform Inference A J IMA. 2019;8(4):691–717.

[17] Balabdaoui F, Doss CR, Durot C. Unlinked monotone regression. 2021. arXiv: http://arXiv.org/abs/arXiv:200700830.

[18] Slawski M, Sen B. Permuted and unlinked monotone regression in $\mathbb{R}^d$: an approach based on mixture modeling and optimal transport. 2022. arXiv: http://arXiv.org/abs/arXiv:220103528.

[19] Callaway B, Li T. Quantile treatment effects in difference in differences models with panel data. Quantitative Econom. 2019;10(4):1579–618.

[20] Bonhomme S, Sauder U. Recovering distributions in difference-in-differences models: A comparison of selective and comprehensive schooling. Review Econom Stat. 2011;93(2):479–94.

[21] Brenier Y. Polar factorization and monotone rearrangement of vector-valued functions. Commun Pure Appl Math. 1991;44(4):375–417.

[22] Caffarelli LA. Monotonicity properties of optimal transportation and the FKG and related inequalities. Comm Math Phys. 2000;214(3):547–63.

[23] Rockafellar RT. Convex analysis. Princeton, NJ: Princeton University Press; 1997.

[24] Matzkin RL. Nonparametric estimation of nonadditive random functions. Econometrica. 2003;71(5):1339–75.

[25] Chernozhukov V, Hansen C. An IV model of quantile treatment effects. Econometrica. 2005;73(1):245–61.

[26] Imbens GW. Nonadditive models with endogenous regressors. Econometric Soc Monographs. 2007;43:17.

[27] Ekeland I, Galichon A, Henry M. Comonotonic measures of multivariate risks. Math Finance. 2010 Oct;22(1):109–32.

[28] Boissard E, Le Gouic T, Loubes JM. Distributionas template estimate with Wasserstein metrics. Bernoulli. 2015;21(2):740–59.

[29] Puccetti G, Scarsini M. Multivariate comonotonicity. J Multivariate Anal. 2010;101(1):291–304.

[30] Ashlagi I, Braverman M, Hassidim A, Monderer D. Monotonicity and implementability. Econometrica. 2010;78(5):1749–72.

[31] Rochet JC. A necessary and sufficient condition for rationalizability in a quasi-linear context. J Math Econom. 1987;16(2):191–200.

[32] Shi X, Shum M, Song W. Estimating semi-parametric panel multinomial choice models using cyclic monotonicity. Econometrica. 2018;86(2):737–61.

[33] Chernozhukov V, Galichon A, Henry M, Pass B. Identification of Hedonic equilibrium and nonseparable simultaneous equations. Journal of Political Economy. 2021;129(3):842–70.

[34] Rüschendorf L. On c-optimal random variables. Stat Probability Lett. 1996;27(3):267–70.

[35] Gunsilius FF. A condition for the identification of multivariate models with binary instruments. J Econometrics. 2022;235:220–38.

[36] Galichon A. A survey of some recent applications of optimal transport methods to econometrics. Econometrics J. 2017;20(2):C1–11.

[37] McCann RJ, Pass B. Optimal transportation between unequal dimensions. Archive Rational Mechanics Anal. 2020;238(3):1475–520.

[38] Auricchio G, Veneroni M. On the structure of optimal transportation plans between discrete measures. Appl Math Optimiz. 2022;85(3):42.

[39] Decancq K, Lugo MA. Weights in multidimensional indices of wellbeing: an overview. Econometric Rev. 2013;32(1):7–34.

[40] Duclos JY, Sahn DE, Younger SD. Robust multidimensional poverty comparisons. The Economic J. 2006;116(514):943–68.

[41] Bourguignon F, Chakravarty SR. The measurement of multidimensional poverty. J Economic Inequality. 2003;1:25–49.

[42] Ekeland I, Galichon A, Henry M. Comonotonic measures of multivariate risks. Math Finance Int J Math Stat Financial Econom. 2012;22(1):109–32.

[43] Embrechts P, McNeil A, Straumann D. Correlation and dependence in risk management: properties and pitfalls. Risk Management Value Risk Beyond. 2002;1:176–223.

[44] Hallin M. Measure transportation and statistical decision theory. Ann Rev Stat Appl. 2022;9:401–24.

[45] García-Gómez C, Perez A, Prieto-Alaiz M. A review of stochastic dominance methods for poverty analysis. J Economic Surveys. 2019;33(5):1437–62.

[46] Embrechts P, Puccetti G. Bounds for functions of multivariate risks. J Multivariate Anal. 2006;97(2):526–47.

[47] Shannon CE. A mathematical theory of communication. Bell Syst Technical J. 1948;27(3):379–423.

[48] Forrow A, Hütter JC, Nitzan M, Rigollet P, Schiebinger G, Weed J. Statistical optimal transport via factored couplings. In: The 22nd International Conference on Artificial Intelligence and Statistics. PMLR; 2019. p. 2454–65.

[49] Gunsilius FF. On the convergence rate of potentials of Brenier maps. Econometric Theory. Cambridge, England: Cambridge University Press; 2021.

[50] Hütter JC, Rigollet P. Minimax estimation of smooth optimal transport maps. Ann Stat. 2021;49(2):1166–94.

[51] de Lara L, González-Sanz A, Loubes JM. A consistent extension of discrete optimal transport maps for machine learning applications. 2021. arXiv: http://arXiv.org/abs/210208644.

[52] Manole T, Balakrishnan S, Niles-Weed J, Wasserman L. Plugin estimation of smooth optimal transport maps. 2021. arXiv: http://arXiv.org/abs/arXiv:210712364.

[53] Manole T, Niles-Weed J. Sharp convergence rates for empirical optimal transport with smooth costs. 2021. arXiv: http://arXiv.org/abs/arXiv:210613181.

[54] Pooladian AA, Niles-Weed J. Entropic estimation of optimal transport maps. 2021. arXiv: http://arXiv.org/abs/arXiv:210912004.

[55] Vacher A, Muzellec B, Rudi A, Bach F, Vialard FX. A dimension-free computational upper-bound for smooth optimal transport estimation. In: Proceedings of Thirty Fourth Conference on Learning Theory. PMLR; 2021. p. 4143–73.

[56] Muzellec B, Vacher A, Bach F, Vialard FX, Rudi A. Near-optimal estimation of smooth transport maps with kernel sums-of-squares. 2021. arXiv: http://arXiv.org/abs/arXiv:211201907.

[57] Deb N, Ghosal P, Sen B. Rates of estimation of optimal transport maps using plug-in estimators via barycentric projections. 2021. arXiv: http://arXiv.org/abs/arXiv:210701718.

[58] Politis DN, Romano JP, Wolf M. On the asymptotic theory of subsampling. Stat Sinica. 2001;11:1105–24.

[59] Romano JP, Shaikh AM. On the uniform asymptotic validity of subsampling and the bootstrap. Ann Stat. 2012;40(6):2798–822.

[60] Shao X, Politis DN. Fixed b subsampling and the block bootstrap: improved confidence sets based on p-value calibration. J R Stat Soc Ser B (Stat Methodol). 2013;75(1):161–84.

[61] Neumark D, Wascher W. Minimum wages and employment: a case study of the fast-food industry in new jersey and pennsylvania: comment. Amer Econ Rev. 2000;90(5):1362–96.

[62] Dube A, Lester TW, Reich M. Minimum wage effects across state borders: Estimates using contiguous counties. Review Econom Stat. 2010;92(4):945–64.

[63] Meer J, West J. Effects of the minimum wage on employment dynamics. J Human Resources. 2016;51(2):500–22.

[64] Neumark D, Salas JI, Wascher W. Revisiting the minimum wage-Employment debate: Throwing out the baby with the bathwater? ILR Review. 2014;67(3 suppl):608–48.

[65] Lu B, Rosenbaum PR. Optimal pair matching with two control groups. J Comput Graph Stat. 2004;13(2):422–34.

[66] Ehrenberg RG, Brown C, Freeman RB, Hamermesh DS, Osterman P, Welch F. Myth and measurement: the new economics of the minimum wage. Industrial Labor Relations Review. 1995;48(4):827.

[67] Ropponen O. Reconciling the evidence of Card and Krueger (1994) and Neumark and Wascher (2000). J Appl Econometrics. 2011;26(6):1051–7.

[68] Segers J. Graphical and uniform consistency of estimated optimal transport plans. 2022. arXiv: http://arXiv.org/abs/arXiv:220802508.

[69] Varadarajan VS. Weak convergence of measures on separable metric spaces. Sankhy Indian J Stat. 1958;19:15–22.

[70] Rockafellar R. Characterization of the subdifferentials of convex functions. Pacific J Math. 1966;17(3):497–510.

[71] Shorack GR, Wellner JA. Empirical processes with applications to statistics. Classics in Applied Mathematics. Philadelphia, PA: Society for Industrial and Applied Mathematics; 2009.

[72] Vapnik VN, Chervonenkis AY. On the uniform convergence of relative frequencies of events to their probabilities. Theory Probability Appl. 1971;16(2):264–80.

[73] Fournier N, Guillin A. On the rate of convergence in Wasserstein distance of the empirical measure. Probability Theory Related Fields. 2015;162(3):707–38.

[74] Canas G, Rosasco L. Learning probability measures with respect to optimal transport metrics. In: Pereira F, Burges CJ, Bottou L, Weinberger KQ, editors. Advances in Neural Information Processing Systems. vol. 25. Boston, MA: Curran Associates, Inc.; 2012. https://proceedings.neurips.cc/paper/2012/file/c54e7837e0cd0ced286cb5995327d1ab-Paper.pdf.

[75] Abadie A, Imbens GW. Large sample properties of matching estimators for average treatment effects. Econometrica. 2006 Jan;74(1):235–67.

[76] Cuturi M. Sinkhorn distances: lightspeed computation of optimal transport. In: Advances in neural information processing systems. vol. 26. Boston, MA: Curran Associates, Inc.; 2013.

[77] Altschuler J, Weed J, Rigollet P. Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. 2017. arXiv: http://arXiv.org/abs/170509634.

[78] Papadakis N. Optimal transport for image processing [Habilitation à diriger des recherches]. Université de Bordeaux; Habilitation thesis; 2015. https://hal.science/tel-01246096.

[79] Seguy V, Damodaran BB, Flamary R, Courty N, Rolet A, Blondel M. Large-scale optimal transport and mapping estimation. 2018. arXiv: http://arXiv.org/abs/171102283.

[80] Makkuva AV, Taghvaei A, Oh S, Lee JD. Optimal transport mapping via input convex neural networks. 2020. arXiv: http://arXiv.org/abs/190810962.

[81] Lu Y, Lu J. A Universal Approximation theorem of deep neural networks for expressing probability distributions. 2020. arXiv: http://arXiv.org/abs/200408867.

[82] Flamary R, Courty N, Gramfort A, Alaya MZ, Boisbunon A, Chambon S, et al. POT: python optimal transport. J Machine Learn Res. 2021;22(78):1–8.

Appendix

A Proofs of stated results

For compactness and clarity, we use a different notation for measures in these proofs than in the main text. For control group outcomes, $\mathbb{P}_{Y_{C,0}}$ is equivalent here to μ_0 and $\mathbb{P}_{Y_{C,1}}$ is equivalent to μ_1. For the treated, $\mathbb{P}_{Y_{T,0}}$ is equivalent to $\mu_0^\star$ and $\mathbb{P}_{Y_{T,1}}$ is equivalent to $\mu_1^\star$. The counterfactual measure $\mathbb{P}_{Y_{T,1}^\dagger}$ has $\mu_1^\dagger$ as its counterpart.

A.1 Proof of Theorem 2

Proof. The cyclically comonotonicity assumption reads as follows:

$$\sum_{i=1}^{m} \langle h_0(u_i), h_1(u_{i+1}) - h_1(u_i) \rangle \le 0$$

for all $m \in \mathbb{N}$. Define $y_i = h_0(u_i)$ and $y_{i+1} = h_0(u_{i+1})$ so that

$$\sum_{i=1}^{m} \langle y_i, h_1 \circ h_0^{-1}(y_{i+1}) - h_1 \circ h_0^{-1}(y_i) \rangle \le 0$$

for all x_0, y_0 in the image of h_0. This means that the function $\mathsf{p} = h_1 \circ h_0^{-1}$ is cyclically monotone in the sense of (4). By [23, Theorem 24.8], and this is equivalent to p being the gradient of a convex function and hence the unique Brenier map from μ_0 to μ_1. Moreover, by Assumption 1 $K_0^* \subset K_0$, so that we can apply the map p to $\mu_0^\star$ to identify $\mu_1^\dagger$. $\qquad\square$

A.2 Proof of Proposition 1

Proof. The main result to enable a consistency statement for $\hat{\mu}_{1,m}^\dagger$ will be the uniform convergence for optimal maps presented in Theorem 1.1 of the study by Segers [68]. This states that with the Euclidean norm

$$\lim_{n \to \infty} \sup_{x_0 \in K_0} \sup_{y \in \hat{\mathsf{p}}_{\#\{x_0\}}} \|y - \mathsf{p}_{\#\{x_0\}}\| \to 0.$$

In other words, the pushforward of singletons under $\hat{\mathsf{p}}$ converges uniformly to their population pushforward under p. The aforementioned statement holds by case (b) of that theorem, and its assumptions are justified as follows. Both the empirical source measure $\hat{\mu}_{0,n}$ and target measure $\hat{\mu}_{1,l}$ converge weakly in probability to their population distributions [69]. The Brenier map p is the unique coupling with cyclically monotone support between μ_0 and μ_1. Finally, both $\hat{\mathsf{p}}$ and p are maximally cyclically monotone functions on the nonempty supports of their respective coupling measures by results in Rockafellar [70].

For finite n, the preimage of $\hat{\mathsf{p}}$ will not contain all possible singletons in $K_0^\star \subset K_0$. Instead, we generate counterfactuals by rounding treatment samples $x_0^\star$ to their nearest neighbor in the control pre-image, $\tilde{x}_0^\star$, and using $\hat{\mathsf{p}}_{\#\{\tilde{x}_0^\star\}}$ as the counterfactual estimator. We wish to show

$$\lim_{n \to \infty} \sup_{x_0^\star \in K_0^\star} \sup_{y \in \hat{\mathsf{p}}_{\#\{\tilde{x}_0^\star\}}} \|y - \mathsf{p}_{\#\{x_0^\star\}}\| \to 0.$$

The exchange of $x_0^\star$ for x_0 in the aforementioned statement follows because the set C of all closed balls on $\mathbb{R}^d$ is a Vapnik-Chervonenkis class with VC dimension $d + 2$ [71,72, p. 833]. Therefore, with $\sigma(\cdot)$ indicating the Borel σ-algebra and $\|\cdot\|_C$ denoting the supremum of the Euclidean distance evaluated over the set of closed balls,

$$\sup_{\mu_0 \in \mathcal{P}(K_0, \sigma(K_0))} \mathbb{E}\|\hat{\mu}_{0,n} - \mu_0\|_C \to 0, \quad n \to \infty.$$

This implies for sufficiently large n the empirical $\hat{\mu}_{0,n}$ will contain a sample from all closed balls. It follows that for any $\varepsilon > 0$, $\tilde{x}_0^{\star}$, the nearest neighbor of $x_0^{\star}$ must be an element of $B_\varepsilon(x_0^{\star})$.

We now consider

$$\limsup_{\min(m,n)\to\infty} \int_{K_0^{\star}} \|\hat{\mathsf{p}}(\tilde{x}) - \mathsf{p}(x)\| \mu_0^{\star}(\mathrm{d}x) \leq \limsup_{\min(m,n)\to\infty} \left[\int_{K_0^{\star}} \|\hat{\mathsf{p}}(\tilde{x}) - \mathsf{p}(\tilde{x})\| \mu_0^{\star}(\mathrm{d}x) + \int_{K_0^{\star}} \|\mathsf{p}(\tilde{x}) - \mathsf{p}(x)\| \mu_0^{\star}(\mathrm{d}x) \right].$$

The first integral must converge to 0 in the limit supremum because of the uniform convergence for optimal maps stated earlier. The second integral involves a Brenier map p from $\mu_0^{\star}$ to $\mu_1^{\dagger}$, which is $\mu_0^{\star}$-Lipschitz almost everywhere. This implies $\int_{K_0^{\star}} \|\mathsf{p}(\tilde{x}) - \mathsf{p}(x)\| \mu_0^{\star}(\mathrm{d}x) \leq C\int_{K_0^{\star}} \|\tilde{x} - x\| \mu_0^{\star}(\mathrm{d}x) \leq C \sup_{K_0^{\star}} \|\tilde{x} - x\|$ for some Lipschitz constant $C < +\infty$. We have already shown that in a sufficiently large sample, all nearest neighbor matches $\tilde{x}$ will be arbitrarily close to x. Therefore, the limit supremum of this integral is bounded by $C\varepsilon$ for any $\varepsilon > 0$. It follows that

$$\limsup_{\min(m,n)\to\infty} \int_{K_0^{\star}} \|\hat{\mathsf{p}}(\tilde{x}) - \mathsf{p}(x)\| \mu_0^{\star}(\mathrm{d}x) = 0.$$

This integral also upper bounds $|\int_{K_0^{\star}} \|\hat{\mathsf{p}}(\tilde{x}) - x\| \mu_0^{\star}(\mathrm{d}x) - \int_{K_0^{\star}} \|\mathsf{p}(x) - x\| \mu_0^{\star}(\mathrm{d}x)|$ by the reverse triangle inequality, so this absolute difference must also converge to zero in the limsup. The sequence of empirical $\hat{\mu}_{0,m}^{\star}$ converge weakly in probability to $\mu_0^{\star}$. Segers' theorem implies uniform convergence, so

$$\int_{K_0^{\star}} \|\hat{\mathsf{p}}(\tilde{x}) - x\| \hat{\mu}_{0,m}^{\star}(\mathrm{d}x) \leq C \sup \|\hat{\mathsf{p}}(\tilde{x}) - x\| \cdot \hat{\mu}_{0,m}^{\star}(K_0^{\star}) \to 0.$$

Hölder's inequality and the triangle inequality gives

$$\left| \int_{K_0^{\star}} \|\hat{\mathsf{p}}(\tilde{x}) - x\| \hat{\mu}_{0,m}^{\star}(\mathrm{d}x) - \int_{K_0^{\star}} \|\hat{\mathsf{p}}(\tilde{x}) - x\| \mu_0^{\star}(\mathrm{d}x) \right|$$

$$\leq \left| \int_{K_0^{\star}} \|\hat{\mathsf{p}}(\tilde{x}) - x\| [\hat{\mu}_{0,m}^{\star} - \mu_0^{\star}](\mathrm{d}x) \right| + \left| \int_{K_0^{\star}} \|\hat{\mathsf{p}}(\tilde{x}) - x\| \mu_0^{\star}(\mathrm{d}x) - \int_{K_0^{\star}} \|\hat{\mathsf{p}}(\tilde{x}) - x\| \mu_0^{\star}(\mathrm{d}x) \right|$$

$$\leq \sup \|\hat{\mathsf{p}}(\tilde{x}) - x\| \cdot [\hat{\mu}_{0,m}^{\star} - \mu_0^{\star}](K_0^{\star}) + \sup \|\hat{\mathsf{p}}(\tilde{x}) - x\| \cdot \mu_0^{\star}(K_0^{\star}) \to 0,$$

so we have convergence in probability of $\int_{K_0^{\star}} \|\hat{\mathsf{p}}(\tilde{x}) - x\| \hat{\mu}_{0,m}^{\star}(\mathrm{d}x)$ to $\int_{K_0^{\star}} \|\hat{\mathsf{p}}(\tilde{x}) - x\| \mu_0^{\star}(\mathrm{d}x)$. By combining these results, we can state that for all $\varepsilon > 0$,

$$\lim_{\min(m,n)} \mathbb{P}\left(\left| \int_{K_0^{\star}} \|\hat{\mathsf{p}}(\tilde{x}) - x\| \hat{\mu}_{0,m}^{\star}(\mathrm{d}x) - \int_{K_0^{\star}} \|\mathsf{p}(x) - x\| \mu_0^{\star}(\mathrm{d}x) \right| > \varepsilon \right) \to 0. \qquad \square$$

A.3 Proof of Proposition 2

Proof. We wish to bound the risk between the true and estimated counterfactual outcome distributions with respect to the 2-Wasserstein metric: $\mathbb{E} W_2(\mathsf{p}_{\#\mu_0^{\star}}, \hat{\mathsf{p}}_{\#\hat{\mu}_{0,m}^{\star}})$. We will show this quantity is $O(n^{\frac{-1}{d}})$. Note that this expectation is jointly over $\mu_0^{\star}$, control measure μ_0 onto whose sampled support we project, and μ_1 through estimating $\hat{\mathsf{p}}$. The assumed mutual independence between all samples will sometimes allow us to marginalize out distributions from an expectation; the distributions that we jointly take expectations over will be noted in that operator's subscript.

By applying the triangle inequality, we have

$$\mathbb{E}_{\mu_0,\mu_1,\mu_0^\star} W_2(\mathsf{p}_{\#\mu_0^\star}, \hat{\mathsf{p}}_{\#\tilde{\mu}_{0,m}^\star}) \leq \mathbb{E}_{\mu_0,\mu_0^\star} W_2(\mathsf{p}_{\#\mu_0^\star}, \mathsf{p}_{\#\tilde{\mu}_{0,m}^\star}) + \mathbb{E}_{\mu_0,\mu_0^\star,\mu_1} W_2(\mathsf{p}_{\#\tilde{\mu}_{0,m}^\star}, \hat{\mathsf{p}}_{\#\tilde{\mu}_{0,m}^\star}). \tag{A1}$$

Consider first $\mathbb{E}_{\mu_0,\mu_0^\star} W_2(\mathsf{p}_{\#\mu_0^\star}, \mathsf{p}_{\#\tilde{\mu}_{0,m}^\star})$. We will bound the expected Wasserstein distance from $\mu_0^\star$ to its projection on the controls' support $\tilde{\mu}_{0,m}^\star$. Then we will show this convergence rate does not change under pushforward of both measures by p.

We have $\mathbb{E}_{\mu_0,\mu_0^\star} W_2(\mu_0^\star, \tilde{\mu}_{0,m}^\star) \leq \mathbb{E}_{\mu_0} W_2(\mu_0^\star, \hat{\mu}_{0,m}^\star) + \mathbb{E}_{\mu_0,\mu_0^\star} W_2(\hat{\mu}_{0,m}^\star, \tilde{\mu}_{0,m}^\star)$ again by the triangle inequality. The first term $\mathbb{E}_{\mu_0^\star} W_2(\mu_0^\star, \hat{\mu}_{0,m}^\star)$ is the expected distance between the true distribution and its i.i.d sample; we have an $O(m^{\frac{-1}{d}})$ upper-bound e.g. from Fournier and Guillin [73]. The second term is the expected Wasserstein distance between sample $\hat{\mu}_{0,m}^\star$ and its nearest neighbor match in the support of $\hat{\mu}_{0,n}$.

The distributions of these empirical measures for a fixed sample size are denoted $\sigma_{0,n}$, $\sigma_{0,m}^\star$, and $\sigma_{1,l}$, respectively. We can express the joint expectation $\mathbb{E}_{\mu_0,\mu_0^\star} W_2(\hat{\mu}_{0,m}^\star, \tilde{\mu}_{0,m}^\star)$ as the marginalized integral over the treatment group's sampling distribution $\int_{\hat{\mu}_{0,m}^\star} \mathbb{E}_{\mu_0} W_2(\hat{\mu}_{0,m}^\star, \tilde{\mu}_{0,m}^\star) \sigma_{0,m}^\star(\hat{\mu}_{0,m}^\star)$. We wish to show that for fixed $\hat{\mu}_{0,m}^\star$, $\mathbb{E}_{\mu_0} W_2(\hat{\mu}_{0,m}^\star, \tilde{\mu}_{0,m}^\star) = O(n^{\frac{-1}{d}})$, which will allow us to conclude $\mathbb{E}_{\mu_0,\mu_0^\star} W_2(\hat{\mu}_{0,m}^\star, \tilde{\mu}_{0,m}^\star) = O(n^{\frac{-1}{d}})$.

Lemma 3.1 from Canas and Rosasco [74] states $W_2^2(\mu, \pi_S\mu) = \mathbb{E}_\mu \|x - S\|^2$, where π_S is the nearest neighbor projection onto set S. In words, squared Euclidean nearest neighbor matching solves an optimal transport problem. Let $S = \operatorname{supp}\hat{\mu}_{0,n}$. This formulation allows us to express $\int_{\hat{\mu}_{0,m}^\star} \mathbb{E}_{\mu_0} W_2^2(\hat{\mu}_{0,m}^\star, \tilde{\mu}_{0,m}^\star) \sigma_{0,m}^\star(\hat{\mu}_{0,m}^\star)$ as $\int_{\hat{\mu}_{0,m}^\star} \mathbb{E}_{\mu_0} \mathbb{E}_{\hat{\mu}_{0,m}^\star} \|x - S\|^2 \sigma_{0,m}^\star(\hat{\mu}_{0,m}^\star)$. Replacing $\mathbb{E}_{\hat{\mu}_{0,m}^\star}$ with a summation yields $\int_{\hat{\mu}_{0,m}^\star} \frac{1}{m} \sum_{i=1}^m \mathbb{E}_{\mu_0} \|x_i - S\|^2 \sigma_{0,m}^\star(\hat{\mu}_{0,m}^\star)$. Lemma 1 given by Abadie and Imbens [75] provides bounds on the matching discrepancy between fixed point z and potential matches drawn i.i.d from a distribution with compact, convex, and bounded support. By denoting the discrepancy vector to the nearest neighbor U_1, they find $\mathbb{E}_{\mu_0} \|U_1\|^3 = O(n^{-\frac{3}{d}})$. Applying the Lyapunov's inequality yields $\mathbb{E}_{\mu_0} \|U_1\|^2 = O(n^{-\frac{2}{d}})$. Thus, have $\mathbb{E}_{\mu_0} W_2^2(\hat{\mu}_{0,m}^\star, \tilde{\mu}_{0,m}^\star) = O(n^{-\frac{2}{d}})$. From Jensen's $(\mathbb{E}_{\mu_0} W_2^2(\hat{\mu}_{0,m}^\star, \tilde{\mu}_{0,m}^\star))^{\frac{1}{2}} \geq \mathbb{E}_{\mu_0} W_2(\hat{\mu}_{0,m}^\star, \tilde{\mu}_{0,m}^\star) = O(n^{\frac{-1}{d}})$. Integrating this bound over the sampling distribution does not change it. Therefore, we can conclude by summing the two triangle inequality rates to yield $\mathbb{E}_{\mu_0,\mu_0^\star} W_2(\hat{\mu}_{0,m}^\star, \tilde{\mu}_{0,m}^\star) = O(\min(m,n)^{\frac{1}{d}})$. Results to shortly follow will show this rate holds for the pushforward $\mathbb{E}_{\mu_0,\mu_0^\star} W_2(\mathsf{p}_{\#\mu_0^\star}, \mathsf{p}_{\#\tilde{\mu}_{0,m}^\star})$ as well.

Now we turn our attention to the rate of convergence for $\mathbb{E}_{\mu_0,\mu_0^\star,\mu_1} W_2(\mathsf{p}_{\#\tilde{\mu}_{0,m}^\star}, \hat{\mathsf{p}}_{\#\tilde{\mu}_{0,m}^\star})$. We once again appeal to the triangle inequality to upper bound this expectation. Note that

$$\mathbb{E}_{\mu_0,\mu_0^\star,\mu_1} W_2(\mathsf{p}_{\#\tilde{\mu}_{0,m}^\star}, \hat{\mathsf{p}}_{\#\tilde{\mu}_{0,m}^\star}) \leq \mathbb{E}_{\mu_0,\mu_0^\star} W_2(\mathsf{p}_{\#\tilde{\mu}_{0,m}^\star}, \mathsf{p}_{\#\mu_0^\star}) + \mathbb{E}_{\mu_0,\mu_0^\star,\mu_1} W_2(\mathsf{p}_{\#\mu_0^\star}, \hat{\mathsf{p}}_{\#\mu_0^\star}) + \mathbb{E} W_2(\hat{\mathsf{p}}_{\#\mu_0^\star}, \hat{\mathsf{p}}_{\#\tilde{\mu}_{0,m}^\star}). \tag{A2}$$

To show the first and third terms are $O(\min(m,n)^{\frac{1}{d}})$, it is sufficient to show the rate of converge between two measures does not change under pushforward via an optimal transport map. Recall that optimal transport maps are twice differentiable almost everywhere, which implies they are also Lipschitz almost everywhere.

Let r denote the rounding optimal transport map from $\mu_0^\star$ to $\tilde{\mu}_{0,m}^\star$. For all x in the support $\mu_0^\star$, we have $\|\mathsf{p} \circ \mathsf{r}(x) - \mathsf{p}(x)\| \leq K\|\mathsf{r}(x) - x\|$ for some Lipschitz constant K. Integrating both sides over $\mu_0^\star$ preserves this inequality and yields $W_2(\mathsf{p}_{\#\tilde{\mu}_{0,m}^\star}, \mathsf{p}_{\#\mu_0^\star}) \leq K W_2(\tilde{\mu}_{0,m}^\star, \mu_0^\star)$. It follows $\mathbb{E}_{\mu_0,\mu_0^\star} W_2(\mathsf{p}_{\#\tilde{\mu}_{0,m}^\star}, \mathsf{p}_{\#\mu_0^\star}) = O(\min(m,n)^{\frac{1}{d}})$. Adapting this argument with $\hat{\mathsf{p}}$ instead of p shows that $\mathbb{E}_{\mu_0,\mu_0^\star,\mu_1} W_2(\hat{\mathsf{p}}_{\#\mu_0^\star}, \hat{\mathsf{p}}_{\#\tilde{\mu}_{0,m}^\star}) = O(\min(m,n)^{\frac{1}{d}})$ as well.

It remains to show $\mathbb{E}_{\mu_0,\mu_0^\star,\mu_1} W_2(\mathsf{p}_{\#\mu_0^\star}, \hat{\mathsf{p}}_{\#\mu_0^\star}) = O(n^{\frac{-1}{d}})$. The Radon–Nikodym derivative f for $\mu_0^\star$ with respect to μ_0 exists and is bounded by assumption. From the definition of a push forward, we see that $W_2(\mathsf{p}_{\#\mu_0^\star}, \hat{\mathsf{p}}_{\#\mu_0^\star})$ can be expressed as $\int_{K_0} \|\mathsf{p}(x) - \hat{\mathsf{p}}(x)\| f(x)\mu_0(\mathrm{d}x)$. It is well known that $\int_{K_0} \|\mathsf{p}(x) - \hat{\mathsf{p}}(x)\|\mu_0(\mathrm{d}x) = O(n^{\frac{-1}{d}})$; see e.g. Manole and Niles-Weed [53]. Let C equal the upper bound of f. Clearly $\int_{K_0} \|\mathsf{p}(x) - \hat{\mathsf{p}}(x)\| f(x)\mu_0(\mathrm{d}x) \leq C\int_{K_0} \|\mathsf{p}(x) - \hat{\mathsf{p}}(x)\|\mu_0(\mathrm{d}x) = O(n^{\frac{-1}{d}})$.

We have shown each term of equation (A2) converges at a rate of $O(\min(m,n)^{\frac{1}{d}})$. Therefore, $\mathbb{E}_{\mu_0,\mu_0^\star,\mu_1}W_2(\mathsf{p}_{\#\tilde{\mu}_{0,m}^\star},\hat{\mathsf{p}}_{\#\tilde{\mu}_{0,m}^\star}) = O(\min(m,n)^{\frac{1}{d}})$. Both terms on the right-hand side of equation (A1) converges at a rate of $O(\min(m,n)^{\frac{1}{d}})$, so $\mathbb{E}_{\mu_0,\mu_0^\star,\mu_1}W_2(\mu_1^\dagger,\hat{\mu}_{1,m}^\dagger) = O(\min(m,n)^{\frac{1}{d}})$ as desired. $\qquad\square$

B Calculating optimal transport couplings from data

This section outlines how to estimate optimal transport couplings from sampled data and provides references to supporting theory. Given n samples from measure μ and m samples from measure ν, both supported on $\mathbb{R}^d$, the objective is to estimate the Brenier map $\bar{T}$ such that $\nu = \bar{T}_{\#}\mu$. The Brenier map's definition is given by (3).

In practice, the measures μ and ν are replaced by their empirical analogues, denoted $\hat{\mu} \coloneqq \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ and $\hat{\nu} \coloneqq \frac{1}{m}\sum_{j=1}^m \delta_{Y_j}$, which are discrete uniform distributions over their samples.[3] The transport cost between observations can be summarized by the matrix $\mathbf{C} \in \mathbb{R}_{\geq 0}^{n\times m}$, where $\mathbf{C}_{i,j} = c(\hat{\mu}[i],\hat{\nu}[j])$, which is defined with respect to a cost function $c(x,y):\mathbb{R}^d\times\mathbb{R}^d\mapsto[0,\infty)$. Theorem 1 holds under the squared Euclidean distance: $c(x,y) = \|x-y\|^2$, which corresponds to the 2-Wasserstein distance between measures. A nonnegative transport matrix $\mathbf{T}$, which moves the atoms of $\hat{\mu}$ to $\hat{\nu}$ is the object to optimize over. We enforce the recovery of marginals in the following discrete analog of equation (3)'s optimization:

$$\min_{\mathbf{T}\in\mathbb{R}_{\geq 0}^{n\times m}}\sum_{i,j}\mathbf{T}_{i,j}\mathbf{C}_{i,j} \quad \text{s.t.}\quad n\mathbf{T}\mathbb{1}_m = \mathbb{1}_n \quad\text{and}\quad m\mathbf{T}^{\mathsf{T}}\mathbb{1}_n = \mathbb{1}_m. \tag{A3}$$

Here, $\mathbb{1}_n$ denotes the length n vector with all entries equal to 1. Noting that (A3) can be reformulated as a linear program, it is common to apply algorithms from that domain, which yield bijective maps when $n = m$ and sparse plans otherwise. Another approach is to add an entropic penalty term to (A3), which relaxes the problem and enables computation through Sinkhorn's algorithm [12,76]. Sinkhorn-based approaches add the penalty term $\lambda\mathbf{H}(\mathbf{T})$ to (A3), where $\mathbf{H}(\mathbf{T})$ is the matrix entropy of $\mathbf{T}$ and positive tuning parameter λ controls the sparsity of $\mathbf{T}$. For large choices of λ, the optimal solution will be driven toward the naive coupling given by the outer-product of marginals; for small λ, sparse solutions more akin to the unregularized problem are found. Without loss of generality assuming $n \geq m$, the runtime of linear programming approaches is generally $O(n^3\log(n))$, while Sinkhorn can achieve $O(n^2\varepsilon^{-3})$ if we wish to bound the error of our discretized Wasserstein distance by ε [77].

Our methodology involves first estimating $\hat{\mathsf{p}}$ from control group observations and then using that estimate to pushforward $\mu_0^\star$. Representing $\hat{\mathsf{p}}$ as a matrix, we can recover its pushforward of $\hat{\mu}_0$ through $n\mathbf{T}\hat{\mu}_1$. As it is unlikely that the supports of the empirical measures $\hat{\mu}_0$ and $\hat{\mu}_0^\star$ have a large intersection, we must extrapolate $\hat{\mathsf{p}}$'s behavior on $\hat{\mu}_0^\star$. A common approach is using nearest neighbors [78]. Counterfactual treatment outcomes are generated by matching each treatment unit preintervention to the nearest control and then pushing forward along the transport plan between observed control distributions. Neural network-based approaches to learn maps have also been proposed [79,80], and representation results help support this direction of active research [81].

To compute optimal transport maps in this article, we use a linear programming method `ot.lp.emd` from the Python package `Python Optimal Transport`[82]. These maps are then extrapolated into optimal transport plans with Euclidean nearest neighbor matching. In our simulations, the control and treatment groups preintervention have almost complete overlap of sampled support, justifying this technique. All numerical experiments in this article are written and run in Python 3. Replication code is available in Appendix of this article.

---

3 For any measurable set A, $\delta_x(A)$ denotes the Dirac measure, which is 1 if $x \in A$ and 0 otherwise.

C Numerical experiments

In this section, we provide implementation details for results computed in Sections 4.1 and 4.2.

C.1 Implementation details for numerical experiment in Section 4.1

We consider $n = 3{,}000$ units in $\mathbb{R}^2$ for each treatment arm. Samples of latent variables are drawn from the distribution ν at $t = 0$ and $t = 1$, mirroring the setup discussed in Section 2.2. Draws from $\nu^\star$ are fixed across time for counterfactual validation purposes. For controls, ν's first coordinate is distributed according to $\mathsf{Beta}(3, 2)$ and second to $\mathsf{Beta}(2, 3)$; for treatments, $\nu^\star$'s first coordinate is distributed according to $\mathsf{Beta}(2, 3)$ and second to $\mathsf{Beta}(3, 2)$.

Recall that we consider the following set of production functions that are cyclically comonotone:

$$h_0(u) = \begin{bmatrix} 1 & \alpha \\ \alpha & 1 \end{bmatrix} u \quad \text{and} \quad h_1(u) = \begin{bmatrix} 1 & -\alpha \\ -\alpha & 1 \end{bmatrix} u.$$

In our simulations, the α parameter of the production functions is fixed at 0.5. The proof of their cyclically comonotone relationship when $|\alpha| \in (0, 1)$ follows.

Proof. We begin by simplifying the inner product terms contained in the definition of cyclical comonotonicity (6). We have

$$\langle h_0(u_i), h_1(u_i) - h_1(u_{i+1}) \rangle = \left\langle \begin{bmatrix} 1 & \alpha \\ \alpha & 1 \end{bmatrix} u_i, \begin{bmatrix} 1 & -\alpha \\ -\alpha & 1 \end{bmatrix} (u_i - u_{i+1}) \right\rangle.$$

After expansion and simplification, this reduces to $(1 - \alpha)^2 \langle u_i, u_i - u_{i+1} \rangle$. Substituting this into the definition of (6), we now wish to show $(1 - \alpha)^2 \sum_{i=1}^{m} \langle u_i, u_i - u_{i+1} \rangle \geq 0$ for all m. Our restriction of $|\alpha| \in (0, 1)$ ensures $(1 - \alpha)^2$ is always positive, and this summation is equivalent to the identity map being cyclically monotone in the sense of (4). The identity map is the gradient of a convex function, $0.5\|u\|_2^2$, completing our proof. $\square$

By using these production functions, we generate n independent samples from each of the distributions μ_0, μ_1 and $\mu_0^\star$, as well as samples from the true counterfactual distribution $\mu_1^\dagger$ for validation purposes.
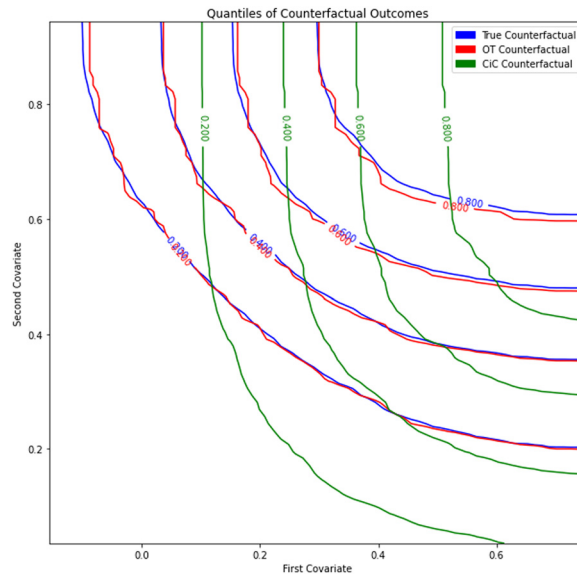

Figure A1: Quantiles of the eCDF for the bivaraite experiment over a uniform mesh with 10,000 points.

Table A1: Recovery of eCDF by method over 20 runs

	MAD OT	MAD CiC
Mean	0.008	0.089
st. dev.	0.002	0.003

To estimate the transport map p, we first compute an optimal plan using observed data and round it to an optimal transport map. This map is only defined on the data from μ_0. To predict counterfactuals treatment on data from $\mu_0^\star$, we employ nearest neighbor interpolation. Since the Beta distribution is supported on $[0, 1]$ for all parameter choices, μ_0 and $\mu_0^\star$ have identical support in this example, and this naive extrapolation technique performs sufficiently well. In particular, it shows that the OT-based estimator remains close to the true counterfactual distribution while naive tensorization of CiC may depart significantly.

We also compute the eCDF for our 3,000 true counterfactual observations and the counterfactuals generated by the two methods over a uniform mesh of 10,000 points. Figure A1 visually demonstrates that OT almost perfectly recovers the eCDF, and Table A1 quantifies this result. The mean absolute difference over each mesh point is an order of magnitude smaller for OT and has a smaller standard deviation.

C.2 Implementation details for numerical experiment in Section 4.2

Recall that we consider a class convex functions indexed by θ, where $g(x; \theta) = \sum_{i=1}^{p} x_{(i_1)}^2 x_{(i_2)}^{-1}$. Here, $x \in \mathbb{R}_{++}^d$, $x_{(i)}$ represents the ith coordinate of x, $i_j \in [[d]]$ for $j \in \{1, 2\}$, and the parameter θ contains the p pairs of indices to be compared.

It is well known the that function $f = x^2/y$ is convex when $y > 0$. Furthermore, the sum of convex functions remains convex. Therefore, the functions $g(x; \theta)$ are convex. In our numerical experiment, we sample θ randomly in each run to contain d coordinate–pair interactions, matching the dimension of the latent vectors.

In our experiment, we define h_0 as the identity and $h_1 = \nabla g$. Therefore, cyclical comonotonicity (6) is equivalent to the cyclical monotonicity (4) of g. As g is convex, ∇g is an optimal transport map and therefore cyclically monotone.

The d coordinate pairs randomly selected from a uniform distribution on $[1, d]$. As these results only hold for strictly positive latent vectors, we sample both latent distributions from $\mathsf{Beta}(2, 3)$. Both treatment arms have $n = 3{,}000$ units.

The implementation of our experiment is otherwise identical to that described in Appendix C.1; we generate data for the four relevant distributions and compute transport plans with nearest neighbor extrapolation. The eCDF and MAD computation is also the same.

D Additional CK experiment

We begin by presenting an additional figure from our bivariate CK analysis in Section 4.3. Figure A2 demonstrates a negative correlation between the change in full-time employees after the intervention and the change in part-time employees. This suggests that fast food restaurants substituted full-time employees for part-time ones.

For our reanalysis of the CK data on the minimum wage and fast food employment, we use the original dataset provided by David Card on his website. It contains survey information, collected via telephone, for 410 fast food restaurants across four fast-food chains. We remove 19 restaurants with missing outcome observations. Of the remaining, 76 are in the control region of northeastern Pennsylvania and 315 are in New Jersey. In
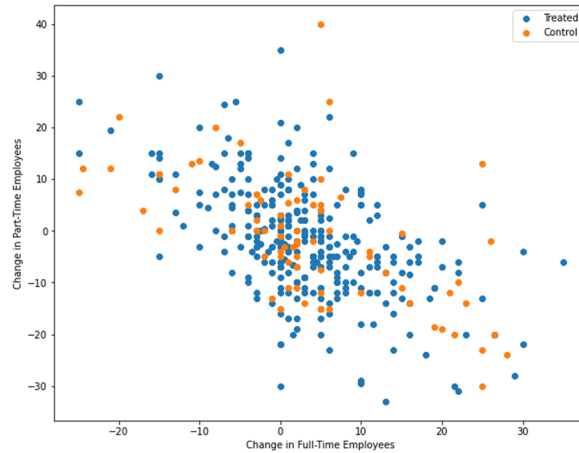
Figure A2: Distribution of unit-level treatment effect estimates, demonstrating a negative correlation between change in full-time and part-time employees.

addition to the two employment outcomes of interest, a number of other observations about a restaurant's wages, prices, and operations were collected. Our optimal transport methodology allows us to estimate treatment effects more richly with potential outcomes including nonemployment time varying restaurant characteristics. We emphasize that this experiment is presented to illustrate the applicability of our method in higher dimensions. Our variable selection is not data driven and the ratio of covariate to units is low.

We select a representative subset of 10 numerical covariates, listed in Table A2, and remove any restaurant with a missing entry for any of these covariates. This leaves a final sample of 52 controls and 200 treatments. The full dataset has 15 numerical covariates but including all 15 leads to a small sample size due to missing data. The covariates we do not include add little additional information, such as the price of fries (we include the price of entrees and sodas), have high rates of missingness, or have a different distribution in the subsample with complete data than the entire survey population. Unlike the outcome-only analysis presented in Section 4.3, the included covariates have different scales. Thus, we standardize each to zero mean and unit standard deviation. The dataset also includes categorical data such as the fast-food chain, which we do not include because the Euclidean distance between restaurants would depend on how these are encoded.

Instead of presenting results conditional on some set of covariates, we present aggregate results over subsets of all subsets of potential outcome vectors including both employment outcomes and 2, 3, or 4 other covariates. The positive results of this experiment are included in Table A3. The sign of our two estimates
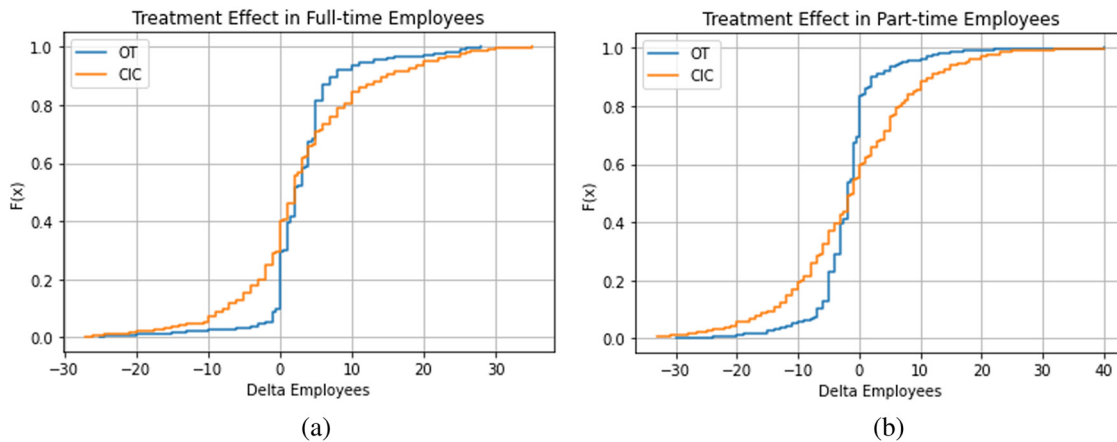

Figure A3: Recovery of counterfactual marginals by OT and CiC. (a) Full-time employment and (b) part-time employment.

Table A2: Outcomes and covariates included in our analysis

Covariate name	Description
EMPFT	Number of full-time employees
EMPPT	Number of part-time employees
PCTAFF	Percent of employees affected by new minimum wage
NMGRS	Number of managers
INCTIME	Months until usual first raise
PENTREE	Price of an entrẮe with tax
PSODA	Price of a soda with tax
NREGS	Number of cash registers in the restaurant
OPEN	Hour of opening
HRSOPEN	Number of hours open per day

Table A3: Summary statistics of higher dimensional subsets for the CK ATE

	TE FT	TE PT
Mean	1.56	−1.97
st. dev.	0.38	0.56
Min	0.69	−3.59
Max	2.63	−0.78

never change in this experiment. Furthermore, addings covariates leads to optimal transport estimators with smaller estimated full-time ATE and larger part-time ATE on aggregate. In future work with more tests for covariate selection and further sensitivity analysis, we hope stronger claims about the estimates' magnitude can be made.