Dealing with a Technological Bias: The Difference-in-Difference Approach*

Dmitry Arkhangelsky †

Current version December 11, 2017
The most recent version is here

Abstract

I construct a nonlinear model for causal inference in the empirical settings where researchers observe individual-level data for few large clusters over at least two time periods. It allows for identification (sometimes partial) of the counterfactual distribution, in particular, identifying average treatment effects and quantile treatment effects. The model is flexible enough to handle multiple outcome variables, multidimensional heterogeneity, and multiple clusters. It applies to the settings where the new policy is introduced in some of the clusters, and a researcher additionally has information about the pretreatment periods. I argue that in such environments we need to deal with two different sources of bias: selection and technological. In my model, I employ standard methods of causal inference to address the selection problem and use pretreatment information to eliminate the technological bias. In case of one-dimensional heterogeneity, identification is achieved under natural monotonicity assumptions. The situation is considerably more complicated in case of multidimensional heterogeneity where I propose three different approaches to identification using results from transportation theory.

Keywords: Treatment effects; Difference-in-difference; Multidimensional heterogeneity; Optimal transportation

*I thank my advisors Guido Imbens and Lanier Benkard for their numerous comments, suggestions and support. I also thank Susan Athey, Stefan Wager, Jens Hainmueller and Filippo Santambrogio for their comments and suggestions. All errors are mine.

†Graduate School of Business, Stanford University, darkhang@stanford.edu.

1 Introduction

I propose a new nonlinear model for causal inference in the frameworks where researchers observe individual-level data for few large clusters over at least two time periods. The proposed model is flexible enough to handle multiple outcome variables, multidimensional heterogeneity, and multiple clusters. It allows for identification (sometimes partial) of the counterfactual distribution, in particular, identifying average treatment effects and quantile treatment effects. The model applies to the settings where the new policy is introduced in some of the clusters (and applies to everybody in the cluster), and a researcher additionally has information about the pretreatment periods.

As a motivating example, consider two schools and assume there is a policy intervention in one of the schools (e.g., change in the curriculum). There are two different sources of the bias that we need to address to make causal statements. The first one is the familiar selection bias: different schools attract students that are different regarding their underlying ability. The second one is the technological bias: different schools use different "production functions", and thus even students with the same skills might have different outcomes if assigned to a different school. These differences in production functions can arise due to various reasons: schools can be different in terms of teachers' qualifications, class sizes, curriculum or general specialization. Athey and Imbens [2006] propose a particular nonparametric strategy for dealing with the selection bias using pretreatment information. In their model subjects in two clusters are different regarding their unobservable characteristic which is assumed to be fixed over time. One way to interpret their identification strategy is to say that the pretreatment outcomes should be used to control for the underlying heterogeneity.[1] This approach is explicitly based on the assumption that there are no structural differences between clusters. In the school example described above it reasonable to expect that the technological bias is present. If this is the case, then the identification strategy is no longer valid.

In this paper, I deal with both technological and selection biases. I assume that the clusters are different regarding the distributions of underlying unobservables. Instead of using pretreatment outcomes to control for this heterogeneity I employ standard techniques, e.g., randomized assignment, covariates or instruments, to deal with it. At the same time, I explicitly allow clusters to be structurally different. Technological bias can be handled if the production functions

[1] This is a conceptual interpretation, one of the central results of the paper is identification in the repeated sampling case, where we observe only the marginal distribution of the pretreatment outcomes.

do not change much over time. I focus on the case when the technology is fixed. With this structure, it is natural to restrict attention to situations in which only two periods are observed, and this is the case that I analyze in the paper. With more periods we can either test that the technology does not change or explicitly model the way it evolves. I view this as a separate question and leave it to future research.

I focus on the situations with few large clusters. In this case, even if the treated clusters are randomly selected we cannot expect the average technological difference between treated and control clusters to be small. As a result, I need to assume that the production functions satisfy specific structural properties. I am trying to find a middle ground by imposing restrictions that are practical and can be motivated by general economic intuition. While not universally applicable, they can serve as an approximation that applied researchers can use either as a first part of the analysis or in the absence of a better model.

This strategy relies heavily on the assumption that the underlying heterogeneity is one-dimensional. In this case, identification can be achieved using monotonicity restrictions. In the school example discussed above it can be motivated in the following way: assume that we take a top student in one of the schools and reassign her to the second school. If schools are similar, then it is natural to expect that the student will still be at the top of the class. I impose a stronger form of this restriction, making production functions in two clusters strictly co-monotone (rank invariant) with respect to some underlying order on the unobservables. Monotonicity restrictions are often used in the identification literature. Two papers that are particularly relevant to my work are Matzkin [2003] and Altonji and Matzkin [2005]. The main difference is that I do not focus on identification of the production function (which can be non-identified in my setup); instead, I use structural properties to connect outcomes in two clusters and explicitly address the technological bias. Identification results in the one-dimensional model can be interpreted in terms of optimal transportation theory. The problem of optimal transportation of measure has a long history in mathematics (see Villani [2008]) and recently attracted considerable attention in economics (see Galichon [2016] for particular applications) and statistics (see Chernozhukov et al. [2017] and Carlier et al. [2016]). This theory is crucial in my work because it paves the way for dealing with multidimensional heterogeneity.

In many cases, it appears unlikely that all the relevant differences between units can be captured by a single unobserved random variable. At the same time, one-dimensional heterogeneity cannot be rejected by the data in the framework with two clusters, two periods and a single

outcome variable. To address this, I assume that more than one outcome variable is observed. This assumption is motivated by the empirical work, where researchers frequently have data on multiple outcome variables. Multiple outcomes allow me to check the validity of the previously described model. In particular, I develop a new, consistent test for one-dimensional heterogeneity. The null hypothesis states that a distinct one-dimensional model generates each of the observed outcomes. Rejection of this hypothesis implies that variables should be analyzed jointly, rather than separately.

Presence of multidimensional heterogeneity considerably complicates the identification strategy. I propose three different approaches. The first method is based on the low-level structural restrictions on production functions. I assume that the relationship between unobservables and outcomes can be identified as an optimal transportation mapping. This idea was previously used in a different context by Chernozhukov et al. [2014]. I show that this approach leads to identification only under very restrictive informational assumptions on the distribution of unobservables. This is not the case in the one-dimensional model, where we can be utterly ignorant about these distributions.

The second approach is based on a generalization of the order restrictions that were used in the one-dimensional case. I prove that such restrictions imply a triangular structure of the production function. This leads to the identification strategy using Knothe-Rosenblatt transportation map (see Chapter 2 of Santambrogio [2015] for definition). This identification technique was previously used in the nonlinear IV context in Imbens and Newey [2009]. In my case, production function itself is not identified, but we can still construct a counterfactual distribution. The key requirement for this approach to be applicable is that researcher needs to select a fixed order on the outcomes. In the applications where there is no information about this order, it does not lead to exact identification.

My last strategy puts explicit high-level restrictions on the relationship between two clusters. These constraints generalize the one-dimensional ones but are arguably less intuitive. The main advantage of this approach is that the counterfactual distribution is identified using a well-defined extremal program: optimal transportation problem with quadratic costs. This result implies that the solution satisfies natural properties that make it reasonable in practice.

I consider several extensions of the basic model. These extensions are developed in the context of one-dimensional heterogeneity. The first extension deals with outcomes with a discrete component in the distribution. I show that in this case counterfactual distribution is partially

identified. The second extension deals with multiple clusters. I demonstrate that in this case identification can be achieved under weaker assumptions using a particular matching algorithm.

Finally, I apply the developed methods to a particular empirical study. I use data from Engelhardt and Gruber [2011]. One of the questions addressed in this paper is the size of the effect that introduction of Medicare Part D had on the out-of-pocket spending. I use my methodology to estimate this effect and get qualitatively similar but more conservative results.

This paper is the first one to explicitly address both technological and selection biases in the diff-in-diff framework both with single and multi-dimensional heterogeneity. My results can be used on the conceptual level, emphasizing the importance of two types of biases and on the practical level, providing a flexible strategy to deal with these issues. I view this as my main contribution to the literature. I also develop new statistical results. For the one-dimensional case, I prove uniform consistency and convergence of the relevant transportation maps, generalizing previously available results. This allows me to construct a powerful test in the case with multiple clusters. This test has a non-pivotal asymptotic distribution, and I show that its distribution can be approximated using a bootstrap-type procedure.

Notation

For any $\Omega \subseteq \mathbb{R}^n$, $\mathcal{B}(\Omega)$ is a Borel σ-algebra on Ω (topology induced by $\mathbb{R}^n$). For any measure μ on $(\Omega, \mathcal{B}(\Omega))$ and function $h : \Omega \to \mathbb{R}$ the expectation with respect to this measure is denoted by $\mathbb{E}_\mu[h(X)]$. I drop the subscript if the measure naturally follows from the context. For any measurable map T $T_\# \mu$ denotes the image measure (pushforward) of μ. $\lambda(\Omega)$ is used for the Lebesgue measure on Ω. For any random vectors X, Z μ_X is used for the distribution of X (image measure), $\sigma(X)$ for the generating σ-algebra and $\mu_{Z|X}$ for the conditional distribution. For a scalar random variable X and any random element Z $F_{X|Z}$ denotes a conditional distribution function of X given Z and $F^{-1}_{X|Z}$ denotes a conditional quantile function.[2] For any random variables (X, Y, Z) I write $X \perp Y|Z$ when X and Y are independent conditionally on Z.

$\mathcal{P}_2(\Omega)$ is a set of measures such that $\mathbb{E}_\mu[\|X\|^2] < \infty$ for any $\mu \in \mathcal{P}_2(\Omega)$. $C_b(\Omega)$ is the vector space of bounded continuous real-valued functions on Ω endowed with the sup-norm; $C^{k,\alpha}(\Omega)$ is a set of k-differentiable functions with α-Hőlder continuous derivatives. For any Ω and $p \in [1, \infty]$ I denote the Wasserstein space with L_p cost function by $W_p(\Omega)$ (see Santambrogio [2015] for definition of this space).

[2] Conditional distributions and measures are well-defined, because I work $\mathbb{R}^n$.

$\|\cdot\|_p$ is used for L_p-type norm ($p = \infty$ corresponds to sup-norm). For vectors $x, y \in \mathbb{R}^n$ $(x, y) = \sum_{j=1}^n x_i y_i$ – the standard dot product on $\mathbb{R}^n$. For any function $\psi : \Omega \to R \cup \{+\infty\}$ Legendre-Fenchel transform of ψ is defined in the following way $\psi^\star(y) = \sup_{x\in\Omega}\{(x, y) - \psi(x)\}$. For any product space $\times_{k=1}^K \Omega_k$ π_k denotes the projection on k-th coordinate. $\{A\}$ is the indicator function for the event A. For any function $f : A \to B$ restriction of f to a set $C \subset A$ is denoted by $f_{|C}$.

For any function $f : \mathcal{X} \to \mathbb{R}$ and a random sample $\{X_i\}_{i=1}^n \in \mathcal{X}^n$ with a distribution $\mathbb{P}$ define $\mathbb{P}_n f := \sum_{i=1}^n f(X_i)$, and $\mathbb{G}_n f := \sqrt{n}\,(\mathbb{P}_n f - \mathbb{P} f)$.[3] For a sequence of random elements $\{X_n\}_{n=1}^\infty$ I use $X_n \xrightarrow{w^\star} X$ to denote weak*-convergence.[4]

2 2×2 model

2.1 Setup

There are two periods $t = 1, 2$ and two clusters $c = 1, 2$. We are interested in the causal effect of a policy variable $w \in \{0, 1\}$. In the period t potential outcomes in cluster c with policy variable equal to w have the following form:

$$Y_t(w, c) = h(w, c, \nu_t) \tag{2.1}$$

where $\nu_t \in V$ is an unobservable random element on (V, σ, μ_{ν_t}) and $h : \{0, 1\}^2 \times V \to \mathbb{R}$ is a real-valued function. The goal is to construct a counterfactual distribution $Y_2(0, 1)$ and estimate a causal effect, e.g., the average treatment effect:

$$\tau := \mathbb{E}_{\mu_{\nu_2}}[Y_2(1, 1) - Y_2(0, 1)] \tag{2.2}$$

or the quantile treatment effect:

$$\tau(q) := Y_2^{-1}(1, 1)(q) - Y_2^{-1}(0, 1)(q) \tag{2.3}$$

The form (2.1) places a restrictions on the potential outcomes. In particular, it assumes that the production function h does not explicitly depend on time, assigning all time variation to the differences in ν_t. This is the key element of the setup that connects two periods. At the same

[3] I use $\mathbb{P}$ instead of $\mathbb{E}$ for the statistical analysis in order to be consistent with the standard notation used in the empirical processes literature.

[4] Whenever measurability issues can arise this should be understood in terms of convergence of outer expectations.

time, h explicitly depends on c, implying that the technology is different between clusters. The following examples suggest particular empirical situations in which this structure makes sense.

Example 2.1.1. (Cohort Analysis) Clusters are two schools; time periods correspond to different cohorts of students. The outcome variable is a success metric (e.g., SAT score). Policy corresponds to a change in the curriculum in the first school that affects the second cohort of students. Different schools might teach the same students differently and thus $h(w, c, \cdot)$ explicitly depends on c. Cohorts are likely to have a different distribution of ν_t (underlying ability). Function h does not explicitly depend on time, meaning that in the absence of treatment schools do not change the way they teach.

Example 2.1.2. (Mobile applications) Clusters are two different mobile platforms, and t are two different time periods. In both applications, we observe user-level data, and the outcome is a success metric (installs, rides, payments). Treatment variable corresponds to a marketing campaign. Applications themselves do not change (technologically) over t, so that function h does not depend on t explicitly. ν_t represents some latent engagement level that has a different distribution in two time periods.

Example 2.1.3. (Medicare example) We observe two time periods and the policy change (expansion of Medicare) happens in the second period. Clusters are defined in terms of eligibility for Medicare: $C_t := \{\text{subject is eligible for Medicare in period } t\}$. Outcome variables are different types of medical expenditures. We assume that people are different in their underlying level of health and preferences for treatment ν_t. The main assumption is that the policy does not change the market environment for non-eligible subjects.

In applications, we can have three types of data. We can either observe the same population over two periods as in Example 2.1.2, two separate populations (cohorts) as in Example 2.1.1 or a mixed case as in Example 2.1.3. In the main text of the paper, I assume that we observe two different populations. I describe the case with a single population observed over time in Appendix A. Identification strategy in the single-population case is conceptually similar, but the underlying assumptions are different and are related to the evolution of unobservables over time.

I assume that the cluster assignment C_t is a measurable function of ν_t, $W_t = \{C_t = 1\}\{t = 2\}$ (diff-in-diff setup) and observable outcomes Y_t are generated in the following way:

$$Y_t = Y_t(W_t, C_t) \tag{2.4}$$

Given μ_{ν_t} this construction defines a distribution $\mu_{(Y_t,C_t,W_t)}$ on $(\mathbb{R}\times\{0,1\}^2, \mathcal{B}(\mathbb{R}\times\{0,1\}^2))$. In the identification part of the paper these distributions are assumed to be known. By definition $\mu_{Y_1(0,c)|C_t=c} = \mu_{Y_1|C_1=c}$, $\mu_{Y_2(0,2)|C_2=2} = \mu_{Y_2|C_2=2}$ and $\mu_{Y_2(1,2)|C_2=1} = \mu_{Y_1|C_2=1}$. I will use this notation interchangeably.

2.2 Discussion of the identification strategy

The primary goal of the paper is to construct a counterfactual distribution of $Y_2(0,1)$ from the observable random variables. There are two sources of differences that should be addressed with this construction. The first difference corresponds to the fact that generally $\mu_{\nu_t|C_t=1} \neq \mu_{\nu_t|C_t=2}$. This is a manifestation of the selection bias: different units select themselves into different clusters. The second difference arises from the fact that $h(w,c,\nu_t)$ explicitly depends on c and corresponds to the technological bias: even in the absence of selection, the same individuals will have different outcomes in different clusters.

These two problems require different approaches. The selection issue is a classical one, and there are many strategies available to cope with it, with the most notable ones based on random experiments, unconfoundedness assumptions, and instrumental variables. The second problem is structural since it requires comparing variables that are inherently distinct. As a result, we need conceptually different assumptions that would allow us to connect outcomes between the clusters. These assumptions are more controversial than those that address the selection problem because there is no perfect benchmark in this case (randomized experiments cannot solve this problem).

My approach to identification in this setup is the following: I will assume that the selection issue can be dealt with by the standard methods (experiments, covariates or instruments) and will focus on the technological problem. This approach should be contrasted with the existent identification diff-in-diff literature (e.g., Athey and Imbens [2006]) that uses the pretreatment periods to address the selection problem.

2.3 Identification: independent case

I start with an assumption that for $t = 1, 2$ ν_t and C_t are independent. This assumption is justified if the subjects were randomly assigned into clusters (due to either controlled or natural experiment). In practice we rarely expect this to hold exactly, but it is a useful starting point.

Assumption 2.3.1. (INDEPENDENT SELECTION) *For $t = 1, 2$ unobservables aren't affected by the selection process:*

$$\mu_{\nu_t|C_t=1} = \mu_{\nu_t|C_t=1} = \mu_{\nu_t} \tag{2.5}$$

This assumption effectively assumes away the selection problem, allowing me to focus on the technological differences between the clusters. I relax this hypothesis in the next subsection, allowing for the selection bias. The main consequence of this assumption is summarized in the following lemma:

Lemma 2.1. (IDENTIFICATION OF POTENITAL OUTCOMES) *Under Assumption 2.3.1 the distributions $\mu_{Y_1(0,1)}, \mu_{Y_1(0,2)}, \mu_{Y_2(0,2)}$, and $\mu_{Y_2(1,1)}$ are identified.*

Proof. By construction $Y_1(0,1) = h(0,1,\nu_1)$ and since by assumption $\mu_{\nu_1|C_1} = \mu_{\nu_1}$ we have that $\mu_{Y_1(0,c)} = \mu_{Y_1(0,C_1)|C_1=1}$. The same logic works for other distributions. □

As a next step, I fix several properties of h and V that pin down the relationship between different clusters. The first assumption restricts the way unobservables and observables are related:

Assumption 2.3.2. (UNIVARIANCE) *For any (w, c) function $h(w, c, \cdot) : V \to \mathbb{R}$ is a bijection.*

While being a standard assumption in the identification literature, the univariance is very restrictive from a theoretical viewpoint. It implies that the heterogeneity is one-dimensional. If only a single outcome is observed then this is non-testable (can not be rejected by any data), but it does not make it more plausible.

The next assumption restricts the measures μ_{ν_1} and μ_{ν_2}:

Assumption 2.3.3. (OVERLAP) *Measure μ_{ν_2} is absolutely continuous with respect to μ_{ν_1}.*

This assumption connects the unobservables in both periods within the cluster. It is essential if we want to achieve full identification. Intuitively the assumption means that in both periods we observe the same units in terms of ν. What is different is the number (measure) of subjects with a particular value of ν. This assumption can be restrictive in specific applications, where we expect ν to increase over time. In this case, it can be relaxed leading to partial identification results.

With Assumptions 2.3.2 and 2.3.3 we can state the following straightforward result:

Lemma 2.2. *Let Assumptions 2.3.2 and 2.3.3 be satisfied. Then for any (w, c) distribution $\mu_{Y_2(w,c)}$ is absolutely continuous with respect to $\mu_{Y_2(w,c)}$.*

Proof. Due to univariance, for any measurable A_t we have the equality $\{Y_t(w, c) \in A_t\} = \{\nu_t \in h^{-1}(w, c)(A_t)\}$. Then the overlap assumption implies the result. □

If Assumptions 2.3.1 and 2.3.2 are satisfied then this lemma implies that overlap is testable. This is natural and important from a practical point of view: if in the second period we observe outcomes that we have never seen before, then we should be particularly cautious.

Finally, I make the structural assumption that connects outcomes in different clusters:

Assumption 2.3.4. (Monotonicity) *There exists a **linear** order $\succsim$ on V such that for any $x, y \in V$, $x \succ y$ implies either $h(0, c, x) > h(0, c, y)$ for $c = 1, 2$ or $h(0, c, x) < h(0, c, y)$ for $c = 1, 2$.*

Again, this is a standard assumption made in the identification literature. Typically, it is stated in a more restrictive form, where V is set to be $[0, 1]$ and the functions are required to be strictly monotonic (increasing). In my case, this is excessive because I do not need to identify function h. The assumption implies that there is an order on observables such that **both** clusters behave similarly with respect to this order. This is a structural assumption that connects two clusters.

Monotonicity is non-testable (since we never observe the same unit in both clusters) but we would not expect it to hold exactly. Its validity depends on the application at hand. E.g., in the school example, we expect it to be satisfied, at least approximately, if both schools have a similar specialty. At the same time, if we are comparing the performance of students from MIT to those from Juilliard, then monotonicity is unreasonable.

Univariance and monotonicity imply the following restriction.

Lemma 2.3. *Let Assumptions 2.3.4, 2.3.2 be satisfied, then the function $\gamma^\star := h(0, 1, \cdot) \circ h^{-1}(0, 2, \cdot) : \mathbb{R} \to \mathbb{R}$ is strictly monotone.*

Proof. Monotonicity and univariance implies that for any $x \neq y$ we have the following inequality $(h(0, 1, x) - h(0, 1, y))(h(0, 2, x) - h(0, 2, y)) > 0$ which implies the statement. □

Assumption 2.3.5. (No atoms) *For $t = 1, 2$ and $c = 1, 2$ distributions $\mu_{Y_t|C_t=c}$ are absolutely continuous with respect to Lebesgue measure on $\mathbb{R}$.*

This assumption restricts the applications of the model to the cases with absolutely continuous outcomes. Discrete outcomes can be included assuming that $h(0, c, \cdot)$ is weakly increasing, leading to partial identification results. I analyze this extension in Section 4.

Combining all the assumptions and lemmas we get the following identification result:

Proposition 2.1. (Identification under independence) *Let Assumptions 2.3.1, 2.3.2, 2.3.3 2.3.4 and 2.3.5 hold. Then the counterfactual distribution of $\mu_{Y_2(0,1)|C_2=1}$ is identified: $\mu_{Y_2(0,1)|C_2=1} = \gamma_{\#}\mu_{Y_2(0,2)|C_2=2}$, where $\gamma = F^{-1}_{Y_1|C_1=1} \circ F_{Y_1|C_1=2}$.*

Proof. Assumptions 2.3.1,2.3.2, 2.3.4 imply that the function $\gamma^\star := h(0, 1, \cdot) \circ h^{-1}(0, 2, \cdot)$ is a transportation map between $\mu_{Y_t(0,2)}$ and $\mu_{Y_t(0,1)}$. Lemma 2.4 implies that the function $\gamma^\star$ is strictly monotone. General results from optimal transportation (see Santambrogio [2015]) then imply that under Assumption 2.3.5 $(\gamma^\star)_{|A} = F^{-1}_{Y_1(0,1)} \circ F_{Y_1(0,2)}$, where A is the support of $\mu_{Y_1|C_1=1}$. Assumption 2.3.3 then implies that $\gamma := (\gamma^\star)_{|A}$ is a valid transportation map from $\mu_{Y_2(0,2)}$ to $\mu_{Y_2(0,1)}$. □

Remark 2.3.1. Observe that the function $\gamma^\star$ is not identified, only its restriction to the support of $\mu_{Y_1(0,2)}$ is. If Assumption 2.3.3 does not hold then we can map only absolutely continuous part of $\mu_{Y_2(0,2)}$ achieving partial identification.

The proposition implies that $Y_t(0, 1) \sim \gamma(Y_t(0, 2))$ and thus using function γ we can construct the counterfactual outcome in the second period and estimate the causal effects, e.g., the average treatment effect:

$$\tau = \mathbb{E}[Y_2|C_2 = 1] - \mathbb{E}[\gamma(Y_2)|C_2 = 0] \tag{2.6}$$

This construction and the restrictions it implies are shown in the following simple example.

Example 2.3.1. (Normal outcomes) Let $\mu_{Y_1(0,c)} = \mathcal{N}(m_c, \sigma_c^2)$. Normalizing $\nu_1 = \lambda([0, 1])$ and assuming stricti monotonicity we have the following expression for function h: $h(0, c, \nu_t) = \sigma_c\Phi^{-1}(\nu_t) + m_c$, where Φ^{-1} is the quantile function of the standard normal distribution.

In this case he function γ has the following form:

$$\gamma(x) = m_1 + \frac{\sigma_1}{\sigma_2}(x - m_2) \tag{2.7}$$

The relationship (2.7) generalizes the linear diff-in-diff type approach. In fact, if $\sigma_1 = \sigma_2$ then we have the standard diff-in-diff identification.

2.3.1 Discussion

Using quantile-quantile methods for identification is not new, the same approach was used in Altonji and Matzkin [2005] for the second estimator that they consider. There this construction is used to identify an analog of function h, while in my case this is just an intermediate step in relating potential outcomes in two different clusters.

The assumptions above were stated in a more abstract way than required by the problem. Indeed, monotonicity assumption 2.3.4 can be substituted by the following, more straightforward one:

Assumption 2.3.6. (Simple monotonicity) *$V = [0,1]$, $\mu_{\nu_1} = \lambda([0,1])$ and for $c = 1,2$ functions $h(0,c,\cdot) : [0,1] \to \mathbb{R}$ are strictly increasing.*

This assumption implies that restriction of function $h(0,c,\cdot)$ is identified as a quantile function in each cluster leading to the same expression for function γ.

I present a more abstract version of the assumptions for two reasons. First of all, I want to emphasize the essence of the identification argument: it does not depend on any topology on V, it does not rely on the distribution of μ_{ν_1}, and it is independent of the identification of the function h. Instead, the central assumptions are univariance and monotonicity, which imply that the technologies can be combined and specify the way to connect them. Secondly, I want to ensure logical continuity between the assumptions in the one-dimensional case and the multivariate case. As I show later, there is a natural extension of these assumptions in the multivariate case.

In the rest of the paper, I relax the three out of four assumptions made above. I start with the independence assumption. I describe how we can achieve identification using the standard methods of causal inference: controlling for observable covariates and instruments. Other approaches might be adapted to this framework; I am focusing on the most conventional ones.

2.4 Identification: selection on observables

In this section, I assume that we have access to a characteristic $X_t \in A$. I explicitly include covariates in the definition of the potential outcome function:

$$Y_t(w,c,x) = h(w,c,\nu_t,x) \tag{2.8}$$

Characteristics have a distribution μ_{X_t}, unobservables have a conditional distribution $\mu_{\nu_t|X_t}$. I assume that cluster assignment C_t is a measurable function of X_t and ν_t, $W_t = \{C_t = 1\}\{t = 2\}$ and the observed outcomes are generated in the following way:

$$Y_t = h(W_t, C_t, \nu_t, X_t) \tag{2.9}$$

This construction defines measures $\mu_{(Y_t,W_t,C_t,X_t)}$ that are used for identification below.

Assumption 2.4.1. (Selection on observables) *Cluster assignment is independent of unobservables given the covariates:*

$$\mu_{\nu_t|X_t,C_t=1} = \mu_{\nu_t|X_t,C_t=2} = \mu_{\nu_t|X_t} \tag{2.10}$$

This is a standard assumption used in the causal inference literature (e.g., Imbens and Rubin [2015]). Its validity depends on the application at hand, but in general a rich set of covariates makes it more plausible. Because the production function h now depends on x univariance, overlap and monotonicity should be generalized in the following way:

Assumption 2.4.2. (Univariance with covariance) *$\nu_t \in V$; for each x and $c = 1, 2$ function $h(0, c, \cdot, x) : V \to \mathbb{R}$ is a bijection.*

Assumption 2.4.3. (Overlap with covariance) *For each x $\mu_{\nu_2|X_2=x}$ is absolutely continuous with respect to $\mu_{\nu_1|X_1=x}$.*

Assumption 2.4.4. (Monotonicity with covariance) *For each x there exists a **linear** order $\succsim_x$ on V such that for any $z_1, z_2 \in V$, $z_1 \succ_x z_2$ implies either $h(0, c, z_1, x) > h(0, c, z_1, x)$ for $c = 1, 2$ or $h(0, c, z_1, x) < h(0, c, z_2, z)$.*

Assumption 2.4.5. (No atoms with covariates) *For $t = 1, 2$, $c = 1, 2$ and x distributions $\mu_{Y_t|C_t=c,X_t=x}$ are absolutely continuous with respect to Lebesgue measure on $\mathbb{R}$.*

We have the same lemma as before (the proof is omitted):

Lemma 2.4. *Let Assumption 2.4.4, 2.4.2 be satisfied, then the function $\gamma_x^\star := h(0, 1, \cdot, x) \circ h^{-1}(0, 2, \cdot, x) : \mathbb{R} \to \mathbb{R}$ is strictly monotone.*

This result leads to the straightforward generalization of Proposition 2.1 (proof is omitted):

Proposition 2.2. (Identification with Covariates) *Let Assumptions 2.4.1,2.4.2, 2.4.3, 2.4.4, and 2.4.5 to be satisfied. Then the conditional (on X_t) counterfactual distribution of $Y_2(0,1)$ is identified: $\mu_{Y_2|C_t=1,X_t=x} = (\gamma_x)_{\#}\mu_{Y_2|C_t=2,X_t=x}$, where $\gamma_x = F^{-1}_{Y_1|C_1=1,X_1=x} \circ F_{Y_1|C_1=2,X_1=x}$.*

Similar to the independent case we have the following: $Y_t(0,1,x) \sim \gamma_x(Y_t(0,2,x))$. The conditional treatment effect is identified in the following way:

$$\tau(x) = \mathbb{E}[Y_2|C_2 = 1, X_2 = x] - \mathbb{E}[\gamma_x(Y_2)|C_2 = 1, X_2 = x] \tag{2.11}$$

Conditional treatment effects can be integrated with respect to different measures on X_t leading to a variety of average effects.

General non-parametric identification can be significantly simplified if we consider a particular conditional distribution of outcomes.

Example 2.4.1. (Conditional Normal Outcomes) It is straightforward to extend the normal example we considered before to the case with covariates. In this case, I assume that $\mu_{Y_1(0,c)|X_1=x} = \mathcal{N}(m_c(x), \sigma_c^2(x))$. The function γ_x has the following form:

$$\gamma_x(y) = m_1(x) + \frac{\sigma_1(x)}{\sigma_2(x)}(y - m_2(x)) \tag{2.12}$$

Similar to the independent case, if $\sigma_1(x) = \sigma_2(x)$ then we have the standard conditional diff-in-diff identification.

Covariates make the univariance assumption more plausible since it needs to hold only for a fixed value of x but might not on average. The same is true for monotonicity because the order can now depend on x. This implies that covariates play two different roles in this setup: they are making independence more plausible, but at the same time they allow for relaxation of the structural assumptions as well. If $\mu_{X_t|C_t=c}$ does not depend on c, meaning that observable characteristics are perfectly balanced in two clusters, then Assumption 2.4.1 implies Assumption 2.3.1. At the same time, even in this case structural assumptions 2.4.2 and 2.4.4 do not imply their unconditional analogs 2.3.2 and 2.3.4. In particular, this means, that even in the perfectly balanced sample we need to use Proposition 2.2 for identification.

2.5 Identification: binary instrument

Selection on observables is not always plausible, especially in cases where we do not have access to a large set of covariates. Identification can be achieved if instead, we have access to an instrument.

Let $\{Z_t\}_{t=1,2}$ be a sequence of **binary** instruments. I make the standard assumptions about the relationship between Z_t and C_t.

Assumption 2.5.1. (Z_t IS AN INSTRUMENT)

(a) Random assignment: $\mu_{\nu_t|Z_t=z} = \mu_{\nu_t}$.

(b) Exclusion: $h(w, c, \nu)$ *doesn't explicitly depend on* z.

(c) Potential cluster assignment function $C_t(z)$ *is increasing in* z.

Values of $C_t(z)$ define three standard groups (for each period): never-takers ($C_t(0) = C_t(1) = 1$), compliers ($C_t(1) > C_t(0) = 1$) and always-takers ($C_t(1) = C_t(0) = 2$). See Imbens and Angrist [1994] for discussion of these groups. I use subscripts (at,cmp,nt) to denote a member of a particular group. Let $\pi_{k,t}$ denote the total measure of group k in period t. Under Assumption 2.5.1 these proportions are identified (Imbens and Angrist [1994]).

A standard result in the instrumental variable literature (Imbens and Rubin [1997]) implies that we can identify the part of the distribution of potential outcomes that corresponds to compliers' groups using the following equalities:

$$\begin{cases} \pi_{cmp,1}\mu_{Y_1(0,1)|cmp} = (\pi_{cmp,1} + \pi_{at,1})\mu_{Y_1(0,1)|C=1,Z_1=1} - \pi_{at,1}\mu_{Y_1(0,1)|C=1,Z_1=0} \\ \pi_{cmp,1}\mu_{Y_1(0,2)|cmp} = (\pi_{cmp,1} + \pi_{nt,1})\mu_{Y_1(0,2)|C=2,Z_1=0} - \pi_{nt,1}\mu_{Y_1(0,2)|C=2,Z_1=1} \\ \pi_{cmp,2}\mu_{Y_2(1,1)|cmp} = (\pi_{cmp,2} + \pi_{at,2})\mu_{Y_2(1,1)|C=1,Z_2=1} - \pi_{at,2}\mu_{Y_2(1,1)|C=1,Z_2=0} \\ \pi_{cmp,2}\mu_{Y_2(0,2)|cmp} = (\pi_{cmp,2} + \pi_{nt,2})\mu_{Y_2(0,2)|C=2,Z_2=0} - \pi_{nt,2}\mu_{Y_2(0,2)|C=2,Z_2=1} \end{cases} \tag{2.13}$$

Thus the distributions $(\mu_{Y_1(0,1)|cmp}, \mu_{Y_1(0,2)|cmp}, \mu_{Y_2(1,1)|cmp}, \mu_{Y_2(0,2)|cmp})$ can be treated as known. This result plays the same role as Corollary 2.1 in the independent case. Compliers' groups can be different in both periods, but this is not important for identification.

Technical assumptions need to be adjusted:

Assumption 2.5.2. (OVERLAP WITH INSTRUMENTS) *Measure* $\mu_{\nu_2|cmp}$ *is absolutely continuous with respect to* $\mu_{\nu_1|cmp}$

Assumption 2.5.3. (NO ATOMS WITH INSTRUMENTS) *For* $c = 1, 2$ *distributions* $\mu_{Y_1(0,c)|cmp}$ *are absolutely continuous with respect to Lebesgue measure on* $\mathbb{R}$.

Overlap is a restrictive assumption, that connects potentially different compliers groups in two periods. Similar to the Assumption 2.3.3 it is testable under monotonicity. I use it as a

high-level restriction that can be adjusted for a particular setup leading to partial identification results.

Together these assumptions imply the following identification result (proof is omitted):

Proposition 2.3. (IDENTIFICATION WITH INSTRUMENTS) *Assume that Assumptions 2.5.1, 2.3.2, 2.3.4, 2.5.2, and 2.5.3 hold; let* $\gamma_{ins} := F^{-1}_{Y_1(0,2)|cmp} \circ F_{Y_1(0,1)|cmp}$. *Then the counterfactual distribution* $\mu_{Y_2(0,1)|cmp}$ *is identified:* $\mu_{Y_2(0,1)|cmp} = (\gamma_{ins})_{\#}\mu_{Y_2(0,2)|cmp}$.

We can construct the following causal effect:

$$\tau_{cmp} = \mathbb{E}[Y_{2|cmp}(1,1)] - \mathbb{E}[\gamma_{ins}(Y_{2|cmp}(0,2))] \tag{2.14}$$

where $Y_{t|cmp}(w,c)$ are random variables with distributions $\mu_{Y_t(w,c)|cmp}$. The average treatment effect isn't identified in this case, only the effect for compliers.

Compared to the case with covariates identification using instruments is more involved: as an intermediate step, we need to compute $\mu_{Y_t(w,c)|cmp}$. In practice, it is difficult to construct the whole distribution function; it is much easier to identify particular moments. If we are willing to assume a parametric model for $\mu_{Y_t(w,c)|cmp}$ then these moments might sufficient for identification. This situation is illustrated with the following example:

Example 2.5.1. (NORMAL OUTCOMES (III)) Assume that $\mu_{Y_1(0,c)|cmp} = \mathcal{N}(m_c, \sigma_c^2)$. Using formulas above we can identify relevant parameters. The means are identified in the following way:

$$\begin{cases} m_1 = \frac{(\pi_{cmp,1}+\pi_{at,1})\mathbb{E}[Y_1(0,1)|C=1,Z_1=1]-\pi_{at,1}\mathbb{E}[Y_1(0,1)|C=1,Z_1=0]}{\pi_{cmp,1}} \\ m_2 = \frac{(\pi_{cmp,1}+\pi_{nt,1})\mathbb{E}[Y_1(0,2)|C=2,Z_1=0]-\pi_{nt,1}\mathbb{E}[Y_1(0,2)|C=2,Z_1=1]}{\pi_{cmp,1}} \end{cases} \tag{2.15}$$

The variances are identified similarly:

$$\begin{cases} \sigma_1^2 = \frac{(\pi_{cmp,1}+\pi_{at,1})\mathbb{E}[(Y_1(0,1)-m_{11})^2|C=1,Z_1=1]-\pi_{at,1}\mathbb{E}[(Y_1(0,1)-m_{11})^2|C=1,Z_1=0]}{\pi_{cmp,1}} \\ \sigma_2^2 = \frac{(\pi_{cmp,1}+\pi_{nt,1})\mathbb{E}[(Y_1(0,1)-m_{12})^2|C=2,Z_1=0]-\pi_{nt,1}\mathbb{E}[(Y_1(0,1)-m_{12})^2|C=1,Z_1=1]}{\pi_{cmp,1}} \end{cases} \tag{2.16}$$

The γ_{ins} function has the same form as before:

$$\gamma_{ins}(x) = m_1 + \frac{\sigma_1}{\sigma_2}(x - m_2) \tag{2.17}$$

with $\sigma_1 = \sigma_2$ this can be viewed as IV diff-in-diff strategy.

The analysis above does not include covariates, but they can be added similarly as before. Identification algorithm becomes even more complicated because we need to construct conditional distributions for compliers. At the same time, if we are willing to make parametric assumptions then the results from (Abadie [2003]) can be applied, making the identification easier.

3 Multidimensional heterogeneity

3.1 Motivation

All the identification results in previous section were achieved under restrictive univariance and monotonicity assumptions (2.3.4,2.3.2). In the framework described so far, these assumptions do not place any testable restrictions on the observable data and cannot be rejected based on empirical evidence. At the same time, they are questionable from a general theoretical perspective: there is no reason to believe that a one-dimensional characteristic can summarize all the inherent heterogeneity of the subjects.

The empirical situation changes if we consider a framework with multiple outcome variables. If these outcomes are connected, then the information contained in their joint distribution might allow us to reject the univariance assumption. I start this section showing that a natural extension of the basic model makes this possible.

The model that can handle the multiple outcomes is useful on its own, not just for testing. In applications we rarely observe a single outcome variable, it is more typical to have a variety of metrics that we are interested in. At the same time, the current practice is to analyze these variables separately, ignoring the information that is contained in their joint distribution.[5] It is a priori unclear how this information should be utilized, and I show that under some assumptions there is a way to do that.

Finally, it is natural to ask what happens to the identification results of the previous section if we still observe only a single outcome, but the restrictive assumptions are relaxed in some way. The extension with the multiple outcomes allows us to approach this question in the following way. Assume that the underlying heterogeneity is two-dimensional, but in the data, we observe only a single outcome. Viewing the second outcome as a latent variable, we can apply the solution concept for the two-dimensional model and use it as a relaxation of the one-dimensional model. I show that this type of argument leads to some qualitative results about the one-dimensional

[5]Situation is different in structural models, where typically all the available information is used.

model.

To construct and analyze the model with multi-dimensional heterogeneity, I use results from optimal transportation theory. There is a variety of great sources on optimal transportation (e.g., Villani [2008] or Santambrogio [2015]). Optimal transportation has a deep connection with several classical economic problems and recently started to attract considerable attention in econometrics (see Galichon [2016] for examples).

3.2 Notation and basic assumptions

I assume that the researcher observes a K-dimensional vector of outcomes

$$Y_t(w,c) := (Y_{t1}(w,c), \dots, Y_{tK}(w,c)) \tag{3.1}$$

Potential outcomes are generated in the following way:

$$Y_{tk}(w,c) = h_k(w,c,\nu_t) \tag{3.2}$$

Defining $h(w,c,\nu_t) := (h_1(w,c,\nu_t), \dots, h_K(w,c,\nu_t))$ we have that $Y_t(w,c) = h(w,c,\nu_t)$. This structure directly generalizes the model from Section 2. Technology (function h) is assumed to be constant in time, with all differences between periods coming from the differences between ν_1 and ν_2. Each function h_k depends on the same unobservables ν_t, implying that all the outcomes should be analyzed together as a vector. Observable data is defined in the same way as before: $W_t = \{C_t = 1\}\{t = 2\}$, C_t is assumed to be the measurable function of ν_t and $Y_t = Y_t(W_t, C_t)$. For $t = 1, 2$ distributions μ_{ν_t} define measures $\mu_{(Y_t,W_t,C_t)}$ on $(\mathbb{R}^K \times \{0,1\}^2, \mathcal{B}(\mathbb{R}^K \times \{0,1\}^2))$ that are assumed to be known in this section.

To focus on the technological bias and abstract away from the selection problem I let Assumption 2.3.1 hold: $\nu_t \perp C_t$ in both periods. As before, it can be relaxed using additional information (covariates and instruments). The first structural assumption that I will adopt throughout the whole section is the following generalization of univariance:

Assumption 3.2.1. (Multivariance) *For $c = 1, 2$ function $h(0, c, \cdot) : V \to \mathbb{R}^K$ is a bijection.*

Informally, this restriction means that we have access to enough outcomes to control the underlying heterogeneity. If we believe that the variables are measured without any idiosyncratic error (which is implicitly assumed throughout the paper), then this assumption is reasonable for K large enough.

Assumption 3.2.2. (Full support) *For $c = 1, 2$ the set $h(0, c, V)$ is open in $\mathbb{R}^K$.*

This assumption implies that the underlying heterogeneity is K-dimensional. This rules out some applications, where be might believe that outcomes lie on a low-dimensional manifold. Identification in this setup is a challenging but conceptually different problem that I leave to further research.

3.3 Testable restrictions in the univariate model

One-dimensional model from Section 2 can be embedded in the framework described above. This requires the following assumption:

Assumption 3.3.1. (Product structure) *$V = \times_{k=1}^K V_k$, $\nu_t = (\nu_{t1}, \dots, \nu_{tK})$ and $h_k(0, c, \nu_t) = \tilde{h}_k(0, c, \nu_{tk})$ for some function $\tilde{h}_k : \{0, 1\}^2 \times V_k \to \mathbb{R}$.*

This restriction means that each outcome variable can only depend on its own unobservable. This does not imply that the outcomes are mutually independent because the random variables $\{\nu_{tl}\}_{l=1}^K$ can be dependent. Together with Assumptions 2.3.2 and 2.3.4 product structure implies the following Proposition.

Proposition 3.1. (Diagonal transportation) *Let Assumption 2.3.1 be satisfied and assume that for each $k = 1, \dots, K$ functions $\tilde{h}_k$ satisfy Assumptions 2.3.2 and 2.3.4 and for each k 2.3.5 holds for $\mu_{Y_{1k}|C_1=c}$ for $c = 1, 2$. Define $\gamma_k := F^{-1}_{Y_{1k}|C_1=1} \circ F_{Y_{1k}|C_1=2}$, then $\gamma := (\gamma_1, \dots, \gamma_k)$ is transportation map: $\mu_{Y_1|C_1=1} = \gamma_{\#}\mu_{Y_1|C_1=2}$.*

Proof. Assumption 2.3.2 holding for $k = 1, \dots, K$ implies that h satisfies Assumption 3.2.1. This implies that $\gamma^\star := h(0, 1, \cdot) \circ h^{-1}(0, 2, \cdot)$ is well-defined and due to independence is a transportation map between $\mu_{Y_1(0,2)}$ and $\mu_{Y_1(0,1)}$. By Assumption 3.3.1 we have that $\gamma^\star$ has the diagonal structure: k-th outcome depends only on k-th coordinate. $\gamma^\star = (\gamma_1^\star, \dots, \gamma_K^\star)$, where each $\gamma_k^\star$ is strictly monotone by Assumption 2.3.4. Proposition 2.1 implies that $\gamma_k = (\gamma_k^\star)_{|\text{supp}\{\mu_{Y_{1k}|C_1=2}\}}$ and the result follows □

The main consequence of Proposition 3.1 is that it implies that a particular function γ (identified from the data) is a transportation map. This is a testable restriction for the following reason:

Corollary 3.3.1. (Testable restriction) *Under the same assumptions as in Proposition 3.1 distributions $\mu_{Y_1|C_1=1}$ and $\mu_{Y_1|C_1=2}$ have the same copula.*

Proof. Above I showed that $(Y_{11}(0,1),\dots,Y_{1K}(0,1)) \sim (\gamma_1(Y_{11}(0,2),\dots,\gamma_K(Y_{1K}(0,2)))$ for strictly increasing functions $\{\gamma_1,\dots,\gamma_K\}$. By defintions it implies that two measures share the same copula. □

Copulas are identified from distributions $\mu_{Y_1|C_1=c}$ $(c = 1, 2)$ and this corollary provides a testable restriction. In particular, if we assume that 2.3.5 holds for each k and 2.3.1 holds (which sometimes can be justified due to explicit randomization) then Corollary 3.3.1 can be viewed as a testable restriction for Assumptions 3.3.1,2.3.4 and 2.3.2. This has a practical implication: in the situations where we observe multiple outcomes, there is a way to understand whether the outcomes should be analyzed jointly or can be analyzed separately using the one-dimensional model. In Section 5 I describe a particular consistent test that can be used to reject this hypothesis.

Corollary 3.3.1 can be used to jointly reject several assumptions. One can ask whether these assumptions can be tested separately. Univariance and monotonicity (for each component) without 3.3.1 are very restrictive, because they imply that for each c and t the distribution $\mu_{Y_t|C_t=c}$ is supported on the one-dimensional subset of $\mathbb{R}^K$ which violates Assumption 3.2.2. Situation is more complicated for 3.3.1: under additional continuity assumptions on functions $h(0,c,\cdot)$ and 3.2.1 with 2.3.1 (which we assume to hold) it can be tested in a similar way. Whether the same can be done without continuity remains an open question.

3.4 Preview of the identification results

In the univariate setting, monotonicity is a low-level assumption, in a sense that it restricts the primitives (function h). It has another advantage: it is expressed entirely in terms of order restrictions, which are arguably more easy to understand and motivate in the applied work. It is natural to follow the same approach in the case with multiple outcomes. Unfortunately, this leads to identifications results that are considerably weaker. In particular, I show that the order restrictions that generalize Assumption 2.3.4 lead to exact identification only under additional informational assumptions. There are infinitely many linear orders on $\mathbb{R}^K$ and to identify the counterfactual distribution we need to know which is the right one.

An alternative approach is to put restrictions on the function h that would lead to its identification. I did not focus on this approach in the one-dimensional case, but it is possible to do this and have the same identification results as before. A standard way to identify a func-

tion is to assume that it solves a known extremal problem. In particular, under multivariance and independence assumptions $h(0, c, \cdot)$ is a transportation map from μ_{ν_t} to $\mu_{Y_t(0,c)}$. Structural restrictions on h can identify it as a particular transportation map, which can then be characterized as a solution to an extremal problem.[6] This approach also leads to weak identification results: the counterfactual distribution is identified if we know a lot about the distribution of unobservables. I show that if there is no such information then in the particular example the mean of the counterfactual distribution is not identified.

Limited applicability of two natural approaches suggests that in order to achieve positive results one might focus on the high-level assumptions on the function $\gamma^\star = h(0, 1, \cdot) \circ h^{-1}(0, 2, \cdot)$. Multivariance and independence imply that this function is a transportation map. I show that natural structural assumptions allow us to identify this function. I emphasize that these assumptions are not the only ones that guarantee the exact identification. In a particular application, one might use different assumptions.

3.5 Identification with low-level assumptions

3.5.1 Order restrictions

Identification in the one-dimensional case was particularly appealing because it was based on the order restrictions. I start with a generalization of this assumption to the multidimensional case. Let $\succsim$ be a standard lexicographic order on $\mathbb{R}^K$: $(x_1, \ldots, x_K) \succ (y_1, \ldots, y_K)$ if either $(x_1 > y_1)$ or $(x_k > y_k$ and $x_l = y_l)$ for $l = 1, \ldots, k-1$ and $1 < k \leq K$. For any permutation $\sigma : \{1, \ldots, K\} \to \{1, \ldots, K\}$ define $\succsim^\sigma$ in the following way:

$$(x_1, \ldots, x_K) \succsim^\sigma (y_1, \ldots, y_K) \Leftrightarrow (x_{\sigma(1)}, \ldots, x_{\sigma(K)}) \succsim (y_{\sigma(1)}, \ldots, y_{\sigma(K)}) \tag{3.3}$$

For any function $T = (T_1, \ldots, T_k) : V \to \mathbb{R}^K$ and permutation σ define T^σ in the following way:

$$T^\sigma(x) := (T_{\sigma(1)}(x), \ldots, T_{\sigma(K)}(x)) \tag{3.4}$$

Using this notation, I make the following assumption:

Assumption 3.5.1. (Order restriction) *There exists a linear order $\succ_\nu$ on V and a permutation σ such that for any $x, y \in V$ with $x \succ_\nu y$ we have either $h(0, c, x) \succ^\sigma h(0, c, y)$ or*

[6] Transportation maps are extremal points in the convex and compact set of joint distributions with given marginals and thus can be supported by some linear functionals.

$h(0,c,y) \succ^{\sigma} h(0,c,x)$ *for* $c = 1, 2$.

This restrictions implies that unobservables are 'ordered' is the same way in both clusters. It is especially powerful when combined with the following assumption:

Assumption 3.5.2. (CONTINUITY) *Function* $\gamma^{\star} := h(0,1,\cdot)^{-1} \circ h^{-1}(0,2,\cdot)$ *is continuous.*

I formulate this assumption for $\gamma^{\star}$ which is well-defined by Assumption 3.2.1, alternatively, one can specify a topology on V and assume that production functions are homeomorphisms.

Proposition 3.2. (TRIANGULAR FORM) *Fix a permutation* σ *and let Assumptions 3.5.1,3.5.2, 3.2.2 and 3.2.1 hold. Then function* $\gamma^{\star}$ *has the following form:*

$$\gamma^{\star}(x) = \begin{pmatrix} \gamma^{\star}_{\sigma(1)}(x_{\sigma(1)}) \\ \gamma^{\star}_{\sigma(2)}(x_{\sigma(1)}, x_{\sigma(2)}) \\ \dots \\ \gamma^{\star}_{\sigma(K)}(x_{\sigma(1)}, x_{\sigma(2)}, \dots, x_{\sigma(K)}) \end{pmatrix} \tag{3.5}$$

Proof. Fix $x, y \in \mathbb{R}^K$ such that $x \neq y$ and $\pi_{\sigma(1)}(x) = \pi_{\sigma(1)}(y) = x_{\sigma(1)}$ Assume that $x \succ^{\sigma} y$ and define the set $A := \{z \in \mathbb{R}^K : x \succsim^{\sigma} z \succsim^{\sigma} y\} \cap h(0,2,V)$. Consider the images of x, y under $\gamma^{\star}$. The goal is to prove that $\gamma^{\star}_{\sigma(1)}(x) = \gamma^{\star}_{\sigma(1)}(y)$, so I assume that this isn't true. Assumption 3.5.1 implies that $\gamma^{\star}(x) \succ^{\sigma} \gamma^{\star}(y)$ and thus $\gamma^{\star}_{\sigma(1)}(x) > \gamma^{\star}_{\sigma(1)}(y)$. Define the set $B := \{z \in \mathbb{R}^K : \gamma^{\star}(x) \succsim^{\sigma} z \succsim^{\sigma} \gamma^{\star}(y)\} \cap h(0,1,V)$ and observe that by definition of $\succsim^{\sigma}$, the fact that $\gamma^{\star}_{\sigma(1)}(x) > \gamma^{\star}_{\sigma(1)}(y)$ and openness of $h(0,1,V)$ we have $\text{int}(B) \neq \emptyset$ (in $\mathbb{R}^K$).

To prove the contradiction observe that by Assumption 3.5.1 we have that $B = \gamma^{\star}(A)$ and by Assumption 3.2.1 $A = \gamma^{-1}(B)$. But then Assumption 3.5.2 implies that A has non-empty interior in $\mathbb{R}^K$ (because $h(0,2,V)$ is open) which is impossible because $\pi_1(A) = x_1$ by construction. This implies that $\gamma^{\star}_{\sigma(1)}(x) = \gamma^{\star}_{\sigma(1)}(0,c,y)$ proving the first line in 3.5.

To prove the rest, I proceed by induction, assuming that the claim is proved up k-th line. Again, fix c and let $x, y \in \mathbb{R}^K$ be such that $x \succ^{\sigma} y$ and $\pi_{\sigma(l)}(x) = \pi_{\sigma(l)}(y) = x_{\sigma(l)}$ for $1 \leq l \leq k+1$. By the induction assumption we have that $\gamma^{\star}_{\sigma(l)}(x) = \gamma^{\star}_{\sigma(l)}(y)$ for $1 \leq l \leq k$. Proving by contradiction assume that $\gamma^{\star}_{\sigma(k+1)}(x) > \gamma^{\star}_{\sigma(k+1)}(y)$. Define the sets A and B as before and observe that the projection of the set B on $K-k$ coordinates has non-empty interior (in $\mathbb{R}^{K-k}$), while the same projection of B does not, proving the contradiction. As a result, the whole claim is proved. □

Remark 3.5.1. It is clear, that for each k function $\gamma^{\star}_k(\cdot)$ is strictly monotone in each of its arguments.

Proposition 3.3. (Identification under order restrictions) *Let Assumptions of Proposition 3.5 and Assumption 2.3.1 hold; assume that $\mu_{Y_1|C_1=c}$ are absolutely continuous with respect to $\lambda(\mathbb{R}^K)$;*[7] *define function $\gamma := \gamma^{\star}_{supp\{\mu_{Y_1|C_1=2}\}}$. Then function γ is equal to Knothe-Rosenblatt transportation map between $\mu_{Y_1|C_1=2}$ and $\mu_{Y_1|C_1=1}$ and thus is identified. If additionally 2.3.3 holds then the counterfactual distribution $\mu_{Y_2(0,1)|C_2=1}$ is identified.*

Proof. Relabel the outcomes according to permutation σ: $Y_t(w,c) := Y_t^{\sigma}(w,c)$. Independence and multivariance implies that $\gamma^{\star}$ is a transportation mapping. Proposition 3.5 implies that $\gamma^{\star}$ has the triangular form with each of functions $\gamma_k^{\star}$ strictly monotone. Together with restrictions on distributions $\mu_{Y_c|C_1=c}$ this implies that $\gamma^{\star}_{|\text{supp}\{\mu_{Y_1|C_1=2}\}}$ is equal to a Knothe-Rosenblatt transportation map between $\mu_{Y_1|C_1=2}$ and $\mu_{Y_1|C_1=1}$ (see Santambrogio [2015] for the precise definition of Knothe-Rosenblatt transportation) proving the first clam. The last claim is obvious. □

Observe that the map γ defined above depends on the labeling of the coordinates: each permutation σ leads to a different function γ. Since there is no reason to assume that σ is known in applications we have the following corollary:

Corollary 3.5.1. *Let Assumptions 3.2.1,2.3.1, 3.5.2 hold. Also assume that there exists an unknown permutation σ such that Assumption 3.5.1 holds. The the function γ is identified up to relabeling the outcomes ($K!$ possible combinations).*

In applications, Assumption 3.5.1 might be too restrictive, and we can substitute it with the following assumption:

Assumption 3.5.3. (Weak order restriction) *There exists a linear order $\succ_{\nu}$ on V and a homeomorphism $g : \mathbb{R}^K \to \mathbb{R}^K$ such that for any $x, y \in V$ with $x \succ_{\nu} y$ we have either $g(h(0,c,x)) \succ g(h(0,c,y))$ or $g(h(0,c,y)) \succ g(h(0,c,x))$ for $c = 1, 2$.*

Using this assumption we can prove in exactly the same way as before that function γ is identified up to a continuous bijection g:

$$\gamma = g^{-1} \circ \tilde{\gamma} \circ g \tag{3.6}$$

where $\tilde{\gamma}$ is Knothe-Rosenblatt transportation between $\mu_{g(Y_1)|C_1=2}$ and $\mu_{g(Y_1)|C_1=1}$. The previous result shows that in general different g will lead to a different γ (relabeling is a particular example of a homeomorphism that satisfies 3.5.3).

[7] This is more restrictive than necessary but suffices in applications.

These results show that unless we have a lot of additional information (know the function g), then the identification using only order restrictions is impossible. It is a consequence of the fact that $\mathbb{R}^K$ has a lot of linear orders that are continuous bijections of the lexicographic order, while $\mathbb{R}$ has only two ($g(x) = x$ and $g(x) = -x$) and $g^{-1} \circ \tilde{\gamma} \circ g$ is the same map in both cases.

Identification using Knothe-Rosenblatt transportation has been used before in the econometric literate (e.g., Matzkin [2003]). One particular application is a non-linear IV case (see Imbens and Newey [2009]) where a particular triangular structure is natural and is a consequence of the exclusion and exogeneity.

3.5.2 Structural technological assumption

I start with an example that illustrates the general problem with the structural assumptions on the function h. I assume that independence assumption 2.3.1 holds.

Example 3.5.1. Consider the following set of distributions: $\mu_{Y_1(0,c)} = \mathcal{N}(m_c, \Sigma_c)$ and $\mu_{\nu_1} = \mathcal{N}(0, \Sigma_0)$. I put a structural assumption on the function h:

Assumption 3.5.4. (Linearity) *For $c = 1, 2$ we $h(0, c, x) = a_c + B_c x$, where B_c is a symmetric positive-definite matrix.*

Under this assumption it is easy to show that a_c and B_c are identified and have the following form (e.g., this follows from the results in Dowson and Landau [1982]):

$$\begin{cases} a_c = m_c \\ B_c = \Sigma_c^{\frac{1}{2}} \left(\Sigma_c^{\frac{1}{2}} \Sigma_0 \Sigma_c^{\frac{1}{2}} \right)^{-\frac{1}{2}} \Sigma_c^{\frac{1}{2}} \end{cases} \tag{3.7}$$

Putting this together we have the following form for the function γ:

$$\gamma(x) = a_1 + B_1 B_2^{-1}(x - a_2) = m_1 + \Sigma_1^{\frac{1}{2}} U \Sigma_2^{-\frac{1}{2}}(x - m_2) \tag{3.8}$$

where $U = \left(\Sigma_1^{\frac{1}{2}} \Sigma_0 \Sigma_1^{\frac{1}{2}} \right)^{-\frac{1}{2}} \Sigma_1^{\frac{1}{2}} \Sigma_2^{-\frac{1}{2}} \left(\Sigma_2^{\frac{1}{2}} \Sigma_0 \Sigma_2^{\frac{1}{2}} \right)^{\frac{1}{2}}$ is an orthogonal matrix that depends on Σ_0. It is easy to see that for any matrix U we can construct a matrix Σ_0 that leads to U.

In applications, it is unreasonable to assume that Σ_0. As a result, we have that γ is identified up to an orthogonal transformation, which in particular implies that the mean of the counterfactual distribution is non-identified.

This example emphasizes the main problem with the identification using structural restrictions on function h. Assumptions that guarantee identification of h most likely lead to results that depend on the unknown distribution μ_{ν_1}.

Assumption 3.5.5. *V and $h(0,c,V)$ are convex and bounded subsets of $\mathbb{R}^K$.*

For any measure μ_ν on $(V, \mathcal{B}(V))$ absolutely continuous with respect to $\lambda(\mathbb{R}^K)$ construct the following function:

$$h_{\mu_\nu}(c,\cdot) := \underset{T:\mu_{Y_1(0,c)}=T_\# \mu_\nu}{\arg\min} \mathbb{E}[\|\nu - T(\nu)\|_2^2] \tag{3.9}$$

This is a consistent definition, because for each absolutely continuous μ_ν the solution is exists and unique (e.g., Villani [2008]). Moreover, different measure μ_ν lead to different functions h_{μ_ν}. This follows from the fact that h is a bijection and $\mu_{Y_1(0,1)}$ is fixed.

Assumption 3.5.6. (Identification of h) *For $c = 1, 2$ function h satisfies the following restriction:*

$$h(0,c,\cdot)_{|supp\{\mu_{\nu_1}\}} = h_{\mu_{\nu_1}}(c,\cdot) \tag{3.10}$$

Similar identification restriction (in different context) was used recently in Chernozhukov et al. [2014]. This assumption leads to the following identification result:

Proposition 3.4. *Let Assumptions 3.2.1, 2.3.1, 3.5.6 hold and assume that μ_{ν_1} is known. Then $\gamma := \gamma^\star_{|supp\{\mu_{Y_1|C_1=2}\}}$ is identified. If additionally Assumption 2.3.3 holds then the counterfactual distribution $\mu_{Y_2(0,1)|C_2=1}$ is identified.*

Proof. Assumption 3.2.1 implies that $\gamma^\star$ is a well-define function. Assumption 2.3.1 implies that $\gamma^\star$ is a transportation map between measures $\mu_{Y_t(0,2)}$ and $\mu_{Y_t(0,1)}$. By independence we have $\gamma^\star_{|\text{supp}\{\mu_{Y_1|C_1=2}\}} = h(0,1,\cdot)_{|\text{supp}\{\mu_{\nu_1}\}} \circ h^{-1}(0,2,\cdot)_{|\text{supp}\{\mu_{\nu_1}\}}$. Assumption 3.5.6 implies that $h(0,c,\cdot)_{|\text{supp}\{\mu_{\nu_1}\}} = h_{\mu_{\nu_1}}(c,\cdot)$ and thus $\gamma = \gamma^\star_{|\text{supp}\{\mu_{Y_1|C_1=2}\}}$ is identified from the data. The second claim follows in a standard way. □

This approach works either if we know μ_{ν_1} or $\gamma^\star_{\mu_{\nu_1}} := h_{\mu_{\nu_1}}(1,\cdot) \circ h^{-1}_{\mu_{\nu_1}}(2,\cdot)$ does not depend on μ_{ν_1}. In applications there is no reason to expect that μ_{ν_1} is known, so I focus on the latter case. The following proposition shows that if μ_{ν_1} belongs to a certain set, then we have identification (proof in Appendix B).

Proposition 3.5. *Define* $\Omega = h(0,1,V) \cup h(0,2,V)$. *Fix arbitrary* μ_1 *and* μ_2 *that belong to the geodesic between* $\mu_{Y_1(0,1)}$ *and* $\mu_{Y_1(0,2)}$ *in* $W_2(\Omega)$. *Then* $\gamma^\star_{\mu_1} = \gamma^\star_{\mu_2}$.

This proposition shows that we can achieve identification using only partial knowledge of μ_{ν_1}, in particular, the fact that it belongs to a geodesic. One might wonder, whether this result can be generalized to other sets in $W_2(\Omega)$. I leave the full characterization for future research, but the following simple example shows that possibilities for identification are severely limited, in a sense that even simple transformations of μ_{ν_1} affect $\gamma^\star_{\mu_{\nu_1}}$.

Example 3.5.2. Let $K = 2$, fix arbitrary $\mu_{Y_1(0,c)} \in \mathcal{P}_2(\Omega)$ and consider two different measures for ν_1: $\mu_1 = \mu_{Y_1(0,2)}$ and $\mu_2 = \mu_{\left(Y_{11}(0,2), \frac{Y_{12}(0,2)}{2}\right)}$. Then it follows that $\gamma^\star_{\mu_1}$ is equal to the optimal transformation map from $\mu_{Y_1(0,2)}$ to $\mu_{Y_1(0,1)}$. At the same time, it is clear that $h^{-1}_{\mu_2}(2,x) = (x_1, \frac{x_2}{2})$. From the general theory we know that $h_{\mu_2}(1,\cdot)$ is a gradient of a convex function. The same is true for $\gamma^\star_{\mu_1}$ (because it is the optimal transportation map). It is easy to see that the map defined a composition of a gradient with a diagonal linear map can't be a gradient (e.g., its Jacobian is not symmetric). This implies that we can have $\gamma^\star_{\mu_1} = h_{\mu_2}(1,\cdot) \circ h^{-1}_{\mu_2}(2,\cdot)$ and thus the counterfactual distribution isn't identified for any set that contains μ_1 and μ_2.

The results presented above imply that identification using structural assumptions on h is very fragile. Similar effects can be achieved using different assumptions, in particular if h is identified as a solution to a transportation problem with a different cost function.

3.6 Identification under high-level assumptions

In this subsection I directly restrict function $\gamma^\star := h(0,1,\cdot) \circ h^{-1}(0,2,\cdot)$. Intuitively this function describes the relationship between the outcomes in two clusters for the same subjects. The first assumption that I make is a version of monotonicity:

Assumption 3.6.1. (Multidimensional monotonicity) *Function* $\gamma^\star := h(0,1,\cdot) \circ h^{-1}(0,2,\cdot)$ *is strictly monotone (as operator in* $\mathbb{R}^K$*). In particular, for any* $x, y \in \mathbb{R}^K$ *the following restriction holds:*

$$(x - y, \gamma^\star(x) - \gamma^\star(y)) > 0 \tag{3.11}$$

This form of operator monotonicity is standard in analysis (e.g., see Boyd and Vandenberghe [2004]). In particular, it implies the following coordinate-wise montonicity:

Corollary 3.6.1. (Coordinate-wise monotonicity) *If $\gamma^\star = (\gamma_1^\star, \dots, \gamma_K^\star)$ satisfies Assumption 3.6.1 then for each k the real-valued function $\gamma_k^\star(y_1, \dots, y^K)$ is strictly monotone with respect to y_k.*

Proof. Consider two vectors $x_1 = (x_{11}, \dots, x_{1k}, \dots, x_{1K})$ and $\tilde{x}_1 = (x_{11}, \dots, \tilde{x}_{1k}, \dots, x_{1K})$ with $x_{1k} > \tilde{x}_{1k}$. Assumption 3.6.1 then implies that $\gamma_k^\star(x_1) > \gamma_k^\star(\tilde{x}_1)$. □

Informally, it implies that changes in outcomes in both clusters are aligned (form an acute angle). This type of restriction might be possible to motivate in specific applications. In the school example with different subject scores as outcome metric, it is natural to expect that changes should be monotone.

The next assumption that I make restricts the relationship between different outcomes:

Assumption 3.6.2. (Symmetry) *Function $\gamma^\star : \mathbb{R}^K \to \mathbb{R}^K$ is differentiable and has a symmetric Jacobian:*

$$\frac{\partial \gamma_k^\star(x)}{\partial x_l} = \frac{\partial \gamma_l^\star(x)}{\partial x_k} \tag{3.12}$$

for all k, l.

This assumption has two parts: a technical part where I assume differentiability of the function $\gamma^\star$ and a conceptual part, where I restrict the relationship between different outcomes. I view this assumption as a more controversial one, because it puts a lot of restrictions on the function $\gamma^\star$ and it is harder to motivate from a practical perspective.

It is important to emphasize that both Assumption 3.6.2 and 3.6.1 are not scale-invariant. In particular, Assumption 3.6.2 is not invariant under (diagonal) linear transformations. This implies that the appropriate scale should be selected before the analysis. In particular, the scale should be chosen in such a way that Assumption 3.6.2 makes sense. One particular option is to normalize the outcomes using quantile functions, but depending on the applications other normalizations might be more attractive.

These two assumptions imply the following lemma:

Lemma 3.1. (Convex potenital) *Under Assumptions 3.6.1, 3.6.2 function $\gamma^\star = \nabla g$, where $g : \mathbb{R}^K \to \mathbb{R}$ is a strictly convex function.*

Proof. Symmetry (and the fact that $\mathbb{R}^K$ is simply connected) implies that $\gamma^\star$ is a path-independent vector field and thus is a gradient for some function $g : \mathbb{R}^K \to \mathbb{R}$. Strict monotonicity then implies that g is strictly convex (see Boyd and Vandenberghe [2004]). □

This lemma is the main part of the the theorem below and thus Assumptions 3.6.1 and 3.6.2 can be substituted directly with assumption that $\gamma^\star = \nabla g$ with g strictly convex. I used an indirect approach, because I believe that symmetry and monotonicity directly emphasize the restriction we are making and thus might be more helpful in applications.[8]

Finally, I need to put some mild technical restrictions on measures that will guarantee that γ can be identified from the data.

Assumption 3.6.3. (Technical conditions) *For $c = 1, 2$ measure $\mu_{Y_1|C_1=c}$ is absolutely continuous with respect to $\lambda(\mathbb{R}^K)$; for $c = 1, 2$ outcomes are square-integrable: $\mathbb{E}[\|Y_1\|^2|C_1 = c] < \infty$; densities $f_{Y_1|C_1=c}$ are supported on the open, bounded and convex regions Λ_c, bounded from below and above and belong to $C^\alpha(\Lambda_c)$ for some α.*

Square-integrability is a standard restriction that guarantees that the optimal transportation problem I state below is well-defined. Restrictions on the densities follow from Caffarelli's regularity theory (see Villani [2008] and references therein). These constraints guarantee that the solution of the problem (and in turn $\gamma^\star$) is smooth (differentiable).

Using all the assumptions and results discussed above, I can state the following identification proposition.

Proposition 3.6. (Identification under high-level assumptions) *Let Assumptions 2.3.1, 3.2.1, 3.6.1, 3.6.2, 3.6.3 hold. Then the function $\gamma := \gamma^\star_{|\Lambda_2}$ is identified as a solution to the following transportation problem:*

$$\gamma := \underset{T:\mu_{Y_1|C_1=1}=T_\#\mu_{Y_1|C_1=2}}{\arg\min} \mathbb{E}\left[\|Y_1 - T(Y_1)\|_2^2 | C_1 = 2\right] \tag{3.13}$$

If additionally Assumption 2.3.3 holds then $\mu_{Y_2(0,1)} = \gamma_\#\mu_{Y_2|C_2=2}$ and thus the counterfactual distribution is identified.

Proof. Independence and multivariance imply that $\gamma^\star := h(0, 1, \cdot) \circ h^{-1}(0, 2, \cdot)$ is a transportation map from $\mu_{Y_1(0,2)}$ to $\mu_{Y_1(0,1)}$. Monotonicity and symmetry allow us to use the result of Lemma

[8] Another approach is to restrict monotonicity to cyclical-montonicity which is a necessary and sufficient condition for $\gamma^\star \in \nabla g$. I do not follow this route because cycical-monotonicity is a global property that is hard to motivate directly.

3.1. The result then follows from the basic optimal transportation theory (see Villani [2008]). Overlap assumption guarantees that $\gamma_{|\Lambda_2}$ is enough to identify the counterfactual in the second period. □

This result allows us to compute any causal effects, e.g., the average treatment effects:

$$\tau_k = \mathbb{E}[Y_{2k}(1,1) - Y_{2k}(0,1)] = \mathbb{E}[Y_{2k}|C_2 = 1] - \mathbb{E}[\gamma_k(Y_2)|C_2 = 2] \tag{3.14}$$

Note that in this case, to compute the treatment effect for the single outcome we need to use the whole vector Y_2.

3.6.1 Properties of the solution

Consider the case when the measures $\mu_{Y_1(0,1)}, \mu_{Y_1(0,2)}$ have the same copulas. It is natural to expect that the solution should be 'diagonal': for each coordinate, we just match the outcomes in the same way as in the one-dimensional model. Motivating any different solution from an agnostic viewpoint is hard. The proposed solution passes this test, which is summarized in the following proposition.

Lemma 3.2. (Consistency) *Let the Assumptions of Proposition 3.6 hold and assume that $\mu_{Y_1(0,1)}, \mu_{Y_1(0,2)}$ share the same copula. Then γ has a diagonal structure:*

$$\gamma = (\tilde{\gamma}_1, \dots, \tilde{\gamma}_K) \tag{3.15}$$

where $\tilde{\gamma}_k = F^{-1}_{Y_{1k}|C_1=1} \circ F_{Y_{1k}|C_1=2}$

Proof. For measures $\mu_{Y_1(0,1)}$ and $\mu_{Y_1(0,2)}$ let $\Gamma(\mu_{Y_1(0,1)}, \mu_{Y_1(0,2)})$ be the set of random vectors (Y_1, Y_2) such that marginal distribution of Y_k is equal to $\mu_{Y_1(0,k)}$. For each $k \in \{1, \dots, K\}$ and $c = 1, 2$ let μ_{ck} be the marginal distribution of k-th coordinate: $\mu_{ck} = (\pi_k)_{\#}\mu_{Y_1(0,c)}$. Under the assumptions of Proposition 3.6 problem (3.13) is equivalent to the following (e.g., Villani [2008]):

$$\min_{(X_1,X_2)\in\Gamma(\mu_{Y_1(0,1)},\mu_{Y_1(0,2)})} \mathbb{E}[\|X_1 - X_2\|^2] \geq \sum_{k=1}^{K} \min_{(X_{1k},X_{2k})\in\Gamma(\mu_{1k},\mu_{2k})} \mathbb{E}[\|X_{1k} - X_{2k}\|^2] \tag{3.16}$$

The solution of the problem on the right side of the inequality is given by $\tilde{\gamma}_k$ (Villani [2008]) and because the distributions share the same copula $\gamma = (\tilde{\gamma}_1, \dots, \tilde{\gamma}_K)$ is a transportation map. It follows from the inequality that it is optimal. □

Now consider a situation where $K = 2$ and $Y_1^{(n)}(0,c) = (Y_{11}(0,c), \frac{Y_{12}(0,c)}{n})$. This corresponds to the case where the variance of the second outcome is small. It is reasonable that in this case, the solution should approach the one-dimensional one. It is, in fact, the case, which is summarized in the following proposition.

Lemma 3.3. (Continuity at zero) *Let the assumptions of Proposition 3.6 hold and assume that* $Y_1^{(n)}(0,c) = (Y_{11}(0,c), \frac{Y_{12}(0,c)}{n})$. *Let* $\gamma^{(n)}$ *be the corresponding solution. Then* $\gamma_1^{(n)}(Y_{11}(0,2), \frac{Y_{12}(0,c)}{n})$ *converges to the one-dimensional solution in* $L^2(\Lambda_2)$ *sense.*

Proof. The minimization program has the following form in this case:

$$\mathbb{E}[\|Y_{11}(0,1) - Y_{11}(0,2)\|^2] + \mathbb{E}\left[\left\|\frac{Y_{12}(0,1)}{n} - \frac{Y_{12}(0,2)}{n}\right\|^2\right] =$$
$$\mathbb{E}[\|Y_{11}(0,1) - Y_{11}(0,2)\|^2] + \frac{1}{n^2}\mathbb{E}[\|Y_{12}(0,1) - Y_{12}(0,2)\|^2] \quad (3.17)$$

The result then follows from the result in Carlier et al. [2010]. □

The two properties represent necessary consistency requirements that we would expect from a reasonable solution concept. It is clear that other approaches might satisfy these restrictions as well.[9]

3.6.2 Sensitivity analysis

Identification results in the multidimensional model have a consequence for the one-dimensional model as well. As I argued before, monotonicity and univariance are very restrictive in the one-dimensional model. A particular way to relax them is to assume that the actual model is the one described in this section, but some of the outcomes are not observed.

In particular, fix $K = 2$, define $Y_t(0,c) = (Y_{t1}(0,c), V_{t2}(0,c))$ and assume that only $Y_{t1}(0,c)$ is observed. This can be viewed as a way to embed the one-dimensional model in a two-dimensional one. Assume for a moment that $V_{t2}(0,c)$ is observed and all the identification assumptions of Proposition 3.6 hold. In this case, that counterfactual distribution is identified using function $\gamma = (\gamma_1, \gamma_2)$. Now, returning back to the fact that only one outcome is observed we are interested in the properties of the function $\gamma_1(Y_{t1}(0,2), V_{t2}(0,2))$, in particular, it is interesting to compare this

[9] Generalization of the first one to the case when the cost function is convex and separable (no cross-coordinate terms) is straightforward.

function with the solution of the one-dimensional model. Below I summarize several qualitative features of this function.

Corollary 3.6.2. *Let Assumptions 2.3.1, 3.2.1,3.6.1,3.6.2,3.6.3 hold. Then the function γ_1 defined above satisfies the following properties:*

(a) γ_1 is monotone in $Y_{t1}(0,2)$;

(b) If $\mu_{V_{t2}(0,2)|Y_{t1}(0,2)}$ doesn't depend on t then the stochastic coupling between $Y_{t1}(0,1)$ and $Y_{t1}(0,2)$ is the same in two periods.

(c) If the copula between $Y_{t1}(0,c), V_{t2}(0,c)$ doesn't depend on c (in particular if $V_{t2}(0,c)$ is independent) then $\gamma_1 = F^{-1}_{Y_{11}|C_1=1} \circ F_{Y_{11}|C_1=2}$ – solution in the one-dimensional model.

(d) Let $V^{(n)}_{t2}(0,2) = \frac{V_{t2}(0,2)]}{n}$ then we have the following:

$$\mathbb{E}[\|\gamma_1(Y_{t1}(0,2), V^{(n)}_{t2}(0,2)) - F^{-1}_{Y_{11}|C_1=1} \circ F_{Y_{11}|C_1=2}(Y_{t1}(0,2))\|^2] \to 0 \tag{3.18}$$

Proof. (a) follows directly from Corollary 3.6.1; (b) follows by definition of γ_1; (c) is the corollary of Proposition 3.2 and (d) is the corollary of Proposition 3.3. □

4 Extensions

4.1 Multiple clusters

I consider an extension of the basic model to the case of multiple clusters (but still restrict analysis to two time periods). Let $\mathcal{C}$ be a finite set of clusters with $c \in \mathcal{C}$ denoting a generic cluster. In terms of data generating process I assume that $W_t = \{t = 2\}\{C_t \in \mathcal{C}_T\}$, where $\mathcal{C}_T \subseteq \mathcal{C}$ is a set of treated clusters; also define $\mathcal{C}_C := \mathcal{C} \setminus \mathcal{C}_T$ – set of control clusters.[10]

There are multiple ways to proceed in this setting. If we maintain all the identification assumptions of Section 2, then there are different ways to construct a counterfactual distribution for each $c \in \mathcal{C}_T$. These counterfactuals should be equal if the model is correctly specified producing a powerful testable restriction. I discuss this approach in Section 5. Alternatively, we can use multiple clusters to weaken the identification assumptions and construct a more robust counterfactual distribution. In this section, I follow the second path. The first assumption that I

[10]Note that $\mathcal{C}_T$ is fixed (non-random).

make describes potential outcomes for all clusters as monotone functions of some fixed functions (types):

Assumption 4.1.1. (Types of outcomes) *Let $\mathcal{L} = \{h_1, \dots, h_L\}$ – finite set of functions, such that for any $h_l \in \mathcal{L}$ function $h_l : V \to \mathbb{R}$, is a bijection and $\|h_l\|_\infty < \infty$. For any $c \in \mathcal{C}$ there exists a strictly increasing function f_c and a function $h_l \in L$ such that $h(0, c, \cdot) = f_c \circ h_l$.*

Unless we somehow restrict how different types are related, this assumption does not have any power: we can always assume that $f_c = \mathrm{Id}$ and let $L = \cup_{c \in \mathcal{C}} h(0, c, \cdot)$. For any $l \in \{1, \dots, |L|\}$ let $\mathcal{C}(l)$ denote the set of clusters such that $h(0, c, \cdot) = f_c \circ h_l$. For any c let $l[c]$ denote the index of its type. The main property of the clusters of the same type is that the following monotonicity restriction holds:

Corollary 4.1.1. *For any l any $c_1, c_2 \in \mathcal{C}(l)$ and $x \neq y \in V$ monotonicity assumption is satisfied:*

$$(h(0, c_1, x) - h(0, c_1, y))\,(h(0, c_2, x) - h(0, c_2, y)) > 0 \tag{4.1}$$

The next assumption restricts the type space, guaranteeing that there is at least one treated and control cluster of each type:

Assumption 4.1.2. (Richness) *There is at least one treated and control cluster of each type: $\mathcal{C}(l) \cap \mathcal{C}_T \neq \emptyset \neq \mathcal{C}(l) \cap \mathcal{C}_C$ for any $l \in \mathcal{L}$.*

The final assumption restricts the distance between clusters of the same type:

Assumption 4.1.3. (Separation of types) *For any $l_1 \neq l_2$ define:*

$$d_{l_1, l_2} := W_\infty(\mu_{h_{l_1}}, \mu_{h_{l_2}}) \tag{4.2}$$

then for any $l \in \mathcal{L}$ and any $c_1 \in \mathcal{C}(l)$ we have:

$$\|h(0, c_1, \cdot) - h_l(\cdot)\|_\infty < \frac{\min_{k \neq l} d_{l,k}}{3} =: \frac{d_l}{4} \tag{4.3}$$

This is a high-level assumption that guarantees that clusters are separated in Wasserstein space. This is summarized in the following lemma:

Lemma 4.1. *Fix arbitrary $l_1 \neq l_2$ and take any $c_1, c_2 \in \mathcal{C}(l_1)$ and $c_3 \in \mathcal{C}(l_2)$. Then Assumption 4.1.3 implies the following:*

$$W_\infty(\mu_{h(0,c_1,\cdot)}, \mu_{h(0,c_2,\cdot)}) < W_\infty(\mu_{h(0,c_1,\cdot)}, \mu_{h(0,c_3,\cdot)}) \tag{4.4}$$

Proof. Because f_c are strictly monotone we have the following:

$$\frac{d_{l_1,l_2}}{2} \geq \frac{d_{l_1}}{2} > |h(0,c_1,\cdot) - h(0,c_2,\cdot)\|_\infty = W_\infty(\mu_{h(0,c_1,\cdot)}, \mu_{h(0,c_2,\cdot)}) \tag{4.5}$$

where the first inequality follows by definition of d_l, the second one follows by triangle inequality (which is valid because W_∞ is a distance) and equality is a consequence of the strict monotonicity of functions f_c. At the same time we have the following chain of inequalities:

$$\begin{aligned}
&W_\infty(\mu_{h(0,c_1,\cdot)}, \mu_{h(0,c_3,\cdot)}) \geq |W_\infty(\mu_{h(0,c_3,\cdot)}, \mu_{h_{l_1}(\cdot)}) - W_\infty(\mu_{h(0,c_1,\cdot)}, \mu_{h_{l_1}(\cdot)})| = \\
&W_\infty(\mu_{h(0,c_3,\cdot)}, \mu_{h_{l_1}(\cdot)}) - W_\infty(\mu_{h(0,c_1,\cdot)}, \mu_{h_{l_1}(\cdot)}) \geq \\
&|W_\infty(\mu_{h_{l_1}(\cdot)}, \mu_{h_{l_2}(\cdot)}) - W_\infty(\mu_{h(0,c_3,\cdot)}, \mu_{h_{l_2}(\cdot)})| - W_\infty(\mu_{h(0,c_1,\cdot)}, \mu_{h_{l_1}(\cdot)}) = \\
&W_\infty(\mu_{h_{l_1}(\cdot)}, \mu_{h_{l_2}(\cdot)}) - W_\infty(\mu_{h(0,c_3,\cdot)}, \mu_{h_{l_2}(\cdot)}) - W_\infty(\mu_{h(0,c_1,\cdot)}, \mu_{h_{l_1}(\cdot)}) \geq d_{l_1,l_2} - \frac{d_{l_1}}{4} - \frac{d_{l_2}}{2} \geq \frac{d_{l_1,l_2}}{4} > 0
\end{aligned} \tag{4.6}$$

To see that these are correct start from the last one and go backwards. Putting the two inequalities together we get:

$$W_\infty(\mu_{h(0,c_1,\cdot)}, \mu_{h(0,c_2,\cdot)}) < \frac{d_{l_1,l_2}}{4} \leq W_\infty(\mu_{h(0,c_1,\cdot)}, \mu_{h(0,c_3,\cdot)}) \tag{4.7}$$

which concludes the proof. □

Together with Assumption 4.1.2 this lemma implies the following corollary:

Corollary 4.1.2. *Let Assumptions 4.1.2 and 4.1.3 hold. For each $c \in \mathcal{C}_T$ define:*

$$\mathcal{C}_C(c) := \arg\min_{k \in \mathcal{C}_C} W_\infty(\mu_{Y_1(0,c)}, \mu_{Y_1(0,k)}) \tag{4.8}$$

the set of nearest control clusters with respect to W_∞ distance. Then $\mathcal{C}_C(c) \in l[c]$, that is control clusters have the same type as the treated one.

Proof. Corollary follows directly from the fact that clusters of the same type are closer than the clusters of the different types (previous lemma) and the fact that there are treated and control clusters of the same type (Assumption 4.1.2). □

These results lead naturally to the following theorem:

Proposition 4.1. (Identification with multiple clusters) *Let Assumptions 2.3.1,4.1.1, 4.1.2,4.1.3 hold. For any $c \in \mathcal{C}_T$ and $k \in \mathcal{C}_C(c)$ define the function $\gamma^\star_{c,k} := h^{-1}(0,c,\cdot) \circ h^{-1}(0,k,\cdot)$. Then $\gamma_{c,k} := (\gamma^\star_{c,k})_{|supp(\mu_{Y_2|C_2=k})}$ is identified. If additionally Assumption 2.3.3 holds then the counterfactual distribution $\mu_{Y_2(0,k)}$ is identified.*

Proof. Independence guarantees that $\mu_{Y_t(0,c)}$ are identified for $t = 1, 2$ and $c \in \mathcal{C}$. Corollary 4.1.1 and Corollary 4.1.2 guarantees that for any $c \in \mathcal{C}_T$ and $k \in \mathcal{C}_C(c)$ function $\gamma^\star_{c,k}$ is strictly monotone and thus its restriction to the support is equal to the optimal transportation map. This implies the identification result. The overlap condition guarantees that the counterfactual distribution in the second period is identified. □

This result prescribes a way to identify a counterfactual distribution for each treated cluster. It is safe to assume that in applications $\mathcal{C}_C(k)$ will be a singleton for all c, implying that this process will produce a single counterfactual distribution. At the same time, this result does not mean that the model does not have overidentifying restrictions. In particular, it might happen that the partition induced by nearest neighbors matching does not satisfy Assumption 4.1.3 (e.g., if the treated clusters matched to different control clusters are close). There always exist a set of types that all assumptions are satisfied: just assume that there is a single type located at the barycenter (in Wasserstein space) of all clusters. Given a particular configuration of distributions, there might be other sets of types that all assumptions are satisfied. It is natural to look for the richest set of types (the finest partition). This partition might be non-unique, but the previous theorem implies that it will always contain nearest neighbors. From a practical perspective, it means that if we are interested in constructing a counterfactual distribution rather than testing the underlying model, we can always use the matching algorithm.

4.2 Semicontinuous outcomes

I assume that the distribution $\mu_{Y_t(0,c)}$ has a discrete component. I develop this extension in the one-dimensional case because it can be done with a simple change in assumptions on primitives. The resulting approach can in principle be adapted to the multi-dimensional case as an ad hoc solution. I do not restrict the absolutely continuous component of $\mu_{Y_t(0,c)}$ in any way. As a result, the proposed solution can be applied both to the cases where outcomes are purely discrete (e.g., binary) or mostly absolutely continuous, but have atoms (e.g., due to censoring at zero).

If the distribution of $\mu_{Y_1(0,c)}$ has a discrete component, then the univariance and monotonicity imply that discrete part of $\mu_{Y_1(0,1)}$ and $\mu_{Y_1(0,2)}$ should have the same structure. This is unnecessarily restrictive and most likely does not hold in empirical applications. It is natural to drop the univariance assumption and weaken the monotonicity assumption. This leads to the following restriction:

Assumption 4.2.1. (WEAK MONOTONICITY) *There exists a linear order $\succsim$ on V such that for any $x, y \in V$, $x \succ y$ implies either $h(0,c,x) \geq h(0,c,y)$ for $c = 1,2$ or $h(0,c,x) \leq h(0,c,y)$ for $c = 1,2$.*

This assumption has the same interpretation as the monotonicity assumption before, but is less restrictive, allowing the outcomes to stay constant. To continue, I need the following definition:

Definition 1. For any two measures μ_1 and μ_2 on $\mathbb{R}$ define $\Gamma(\mu_1, \mu_2) := \left(F^{-1}_{\mu_1}(U), F^{-1}_{\mu_2}(U)\right)$, where $\mu_U = \lambda([0,1])$ and $\gamma_{mon}(\mu_1, \mu_2) := \mu_{\Gamma(\mu_1,\mu_2)}$; $\gamma_{mon}(\mu_1, \mu_2)$ is called a *co-monotone* transport plan between measures μ_1 and μ_2.

The crucial role of monotone plans is a consequence of the following lemma:

Lemma 4.2. *Assume that 4.2.1 is satisfied. Let $Z_t := (h(0,1,\nu_t), h(0,2,\nu_t))$; then $\mu_{Z_t} = \gamma_{mon}(\mu_{h(0,1,\nu_t)}, \mu_{h(0,2,\nu_t)})$. Additionally, if Assumption 2.3.3 holds then the following restriction on the supports is satisfied:*

$$supp\left(\gamma_{mon}(\mu_{h(0,1,\nu_2)}, \mu_{h(0,2,\nu_2)})\right) \subseteq supp\left(\gamma_{mon}(\mu_{h(0,1,\nu_1)}, \mu_{h(0,2,\nu_1)})\right) \tag{4.9}$$

Proof. The first part of the lemma follows directly from Lemma 2.8 in (Santambrogio [2015]). Restriction on supports is a direct consequence of the overlap assumption that implies that μ_{Z_2} is absolutely continuous with respect to μ_{Z_1} and the fact that $\gamma_{mon}(\mu_{h(0,1,\nu_2)}, \mu_{h(0,2,\nu_2)}) = \mu_{Z_2}$. □

Remark 4.2.1. Overlap assumption has another straightforward but important consequence. Define $\mu_t := \mu_{h(0,2,\nu_t)}$ and let μ_t^d be the discrete part of measure μ_t (see Kolmogorov and Fomin [1968]). Absolute continuity of μ_2 with respect to μ_1 implies that μ_2^d is absolutely continuous with respect to μ_1^2 and as a result $\text{supp}(\mu_2^d) \subseteq \text{supp}(\mu_1^d)$.

Before I state the main result I need to define several additional sets. In particular, for any $x \in \mathbb{R}$, define $A(x) := \{y : (y,x) \in \text{supp}\left(\gamma_{mon}\left(\mu_{Y_1|C_1=1}, \mu_{Y_1|C_1=2}\right)\right)$ and $A^d(x) := A(x) \cap$

$\text{supp}\left(\mu^d_{Y_1|C_1=1}\right)$, where $\mu^d_{Y_1|C_1=1}$ is a discrete component of $\mu_{Y_1|C_1=1}$. For any $x \in \mathbb{R}$ let $\mathcal{M}(x)$ be the set of probability measures such that for any $\mu \in \mathcal{M}(x)$ we have $\text{supp}(\mu) = A(x)$ and $\text{supp}(\mu^d) = A^d(x)$. Let $\mathcal{K}$ be the set of probability kernels (conditional probability distributions) such that for any $K : \mathbb{R} \times \mathcal{B}(\mathbb{R}) \to [0,1]$ in this set we have that $K(x,\cdot) \in \mathcal{M}(x)$.[11]

Proposition 4.2. *Let Assumptions 2.3.1,4.2.1 and 2.3.3 hold. Then the counterfactual distribution $\mu_{Y_2(0,1)}$ is partially identified; in particular, the following inclusion holds:*

$$\mu_{Y_2(0,1)|C_1=1} \in \{\mu \in \mathcal{P}(\mathbb{R}) : \mu(A) = \int K(x,A) d\mu_{Y_2(0,2)|C_2=2} \text{ for some } K \in \mathcal{K}\} \tag{4.10}$$

Proof. Independence implies that $\mu_{Y_1(0,c)} = \mu_{Y_1|C_1=c}$. It follows from Lemma 4.2 and independence that $\text{supp}\left(\gamma_{mon}(\mu_{h(0,1,\nu_1)}, \mu_{h(0,2,\nu_1)})\right)$ is identified. Overlap assumption implies the restriction on supports 4.9. Decomposing measure $\mu_{(Y_2(0,1),Y_2(0,2))}$ into a marginal component $\mu_{Y_2(0,2)}$ and kernel K we have that $K \in \mathcal{K}$. This implies the result. □

Statement of the proposition might look unnecessarily complicated because it covers a lot of different situation. Informally the result is simple: in the first period we identify the support of the joint distribution, and in the second period we observe only one marginal (but know the support of the joint distribution). Kernels then just specify how the marginal measure is split between for each x. Additional properties can help with constructing a particular kernel that satisfies the restriction. First, if $F_{Y_2|C_2=2}$ is continuous at x then by construction of γ_{mon} it follows that $A(x)$ is a singleton and thus $K(x,\cdot)$ is a Dirac measure for any $K \in \mathcal{K}$. Second, let K_1 be the kernel of $\gamma_{mon}(\left(\mu_{Y_1|C_1=1}, \mu_{Y_1|C_1=2}\right))$, then $K_1 \in \mathcal{K}$.[12] As a result, we can always use the kernel from the first period. If the measures are entirely discrete then kernels are simple: each kernel can be represented as a collection of function from $\text{supp}(\mu_{Y_1(0,2)})$ into the finite-dimensional simplex. In this case, it is easy to optimize over this set, and thus partial identification analysis is possible. If we are not that interested in the partial identification but instead want to construct an answer, then the kernel from the first period seems to be the most natural choice.

5 Estimation and inference

I present formal statistical results for the several versions of the model considered in the previous parts. I focus on the case with no covariates or instruments. Discrete covariates that take a finite

[11] I include necessary measurability restrictions requiring that $K(\cdot,\cdot)$ is a valid probability kernel.
[12] In particular, it proves that the set $\mathcal{K}$ is non-empty.

number of values can be introduced straightforwardly.

5.1 One-dimensional model

Basic estimation and inference in the (continuous) one-dimensional model can be done using the well-known results from Athey and Imbens [2006] or Matzkin [2003]. I present a generalization: I prove functional central limit theorems for the transportation maps and quantile function of the counterfactual distribution. This provides additional opportunities for testing in the case with multiple clusters.

For $t = 1, 2$ define $\mathcal{D}_t := (\mathbb{R} \times \mathcal{C} \times \{0, 1\})^{n_t}$. In each period t we observe a random element $\{Y_{it}, C_{it}, W_{it}\}_{i \in n_t} \in \mathcal{D}_t$ from $\left(\mathcal{D}_t, \mathcal{B}(\mathcal{D}_t), \otimes_1^{n_t} \mu_{(Y_t, C_t, W_t)}\right)$.

5.1.1 Estimation, continuous case

I start with the following assumption:

Assumption 5.1.1. *For $c \in \mathcal{C}$ and $t = 1, 2$ distributions $F_{Y_t|C_t=c}$ are strictly increasing, continuously differentiable functions on $[a, b]$ with strictly positive derivative on (a, b).*

This assumption guarantees that the transportation function is strictly increasing and maps $[a, b]$ into $[a, b]$, in particular it is bounded. Fix arbitrary $c \in \mathcal{C}_T$ and $k \in \mathcal{C}_C$. In the continuous case the main object of interest is the following transportation function:

$$\gamma_{c,k} := F^{-1}_{Y_1|C_1=c} \circ F_{Y_1|C_1=k} \tag{5.1}$$

Following Athey and Imbens [2006] I propose to use a plug-in estimator:

$$\hat{\gamma}_{c,k} := \hat{F}^{-1}_{Y_1|C_1=1} \circ \hat{F}_{Y_1|C_1=2} \tag{5.2}$$

where $\hat{F}_{Y_1|C_1=c}$ and $\hat{F}^{-1}_{Y_1|C_1=k}$ are some estimators of $F_{Y_1|C_1=c}$ and $F^{-1}_{Y_1|C_1=k}$. Perhaps, the most natural choice is to use empirical distribution and empirical quantile function, but other estimators can be used, e.g., those arising from smoothing the empirical distribution. Depending on different estimators $\hat{f}_{c,k}$ might have different statistical properties, I discuss this later in the section on inference.

Given the function $\hat{\gamma}_{c,k}$ we can estimate the average effect in a straightforward way:

$$\hat{\tau}_{c,k} = \frac{1}{n_{2c}} \sum_{i=1}^{n_{2c}} Y_{2ci} - \frac{1}{n_{2k}} \sum_{i=1}^{n_{2k}} \hat{\gamma}_{c,k}(Y_{2ki}) \tag{5.3}$$

Another opportunity, is to estimate the difference in quantiles. Since $\gamma_{c,k}$ is strictly monotone, quantile function is given by $\gamma_{c,k} \circ F^{-1}_{Y_2|C_2=k}$. In this case, the plug-in estimator of the difference in quantile functions is given by:

$$\hat{\tau}_{c,k}(q) = \hat{F}^{-1}_{Y_2|C_2=c}(q) - \hat{\gamma}_{c,k} \circ \hat{F}^{-1}_{Y_2|C_2=k}(q) \tag{5.4}$$

As a preparation for the inference step, I introduce two additional operators:

$$\begin{cases} C_1 : l([a,b]) \to l([0,1]) \\ C_1(F)(p) := \inf\{x \in [a,b] : F(x) > p\} \\ C_2 : l([a,b]) \times l([c,d]) \to l([a,b]) \\ C_2(F,G) := F \circ G \end{cases} \tag{5.5}$$

Be definition $C_1(F) = F^{-1}$ for any distribution function F. Using this notation, I define the following operators:

$$\begin{cases} C_3 : l([a,b]) \times l([a,b]) \to l([a,b]) \\ C_3(F,G) := C_2(C_1(F),G) \\ C_4 : l([a,b]) \times l([a,b]) \times l([a,b]) \to l([a,b]) \\ C_4(F,G,L) := C_2(C_2(C_1(F),G),C_1(L)) \end{cases} \tag{5.6}$$

With this notation function $\gamma_{c,k}$ and the quantile function of the counterfactual distribution have the following form:

$$\begin{cases} \gamma_{c,k} = C_3(F_{Y_1|C_1=c}, F_{Y_1|C_1=k}) \\ \gamma_{c,k} \circ F^{-1}_{Y_2|C_2=k} = C_4(F_{Y_1|C_1=c}, F_{Y_1|C_1=k}, F_{Y_2|C_2=l}) \end{cases} \tag{5.7}$$

This form is useful, because statistical properties of the estimators will follow directly from continuity and appropriate differentiability of operators C_3 and C_4.

5.1.2 Consistency, continuous case

Consistency is a direct consequence of the following lemma:

Lemma 5.1. *Let $F_1 \in D[a,b]$ be a strictly increasing (but probably discontinuous) function such that $f_1 := F_1' > c > 0$ at all points where it exists. Then C_1 is continuous at F (with respect to $\|\cdot\|_\infty$). For any $F_2 \in C[a,b]$ and $F_3 \in D[a,b]$ C_2 is continuous at (F_2, F_3) (with respect to* sup*-norm).*

The proof is in Appendix C. Note that I allow for F_1 to be discontinuous.

Corollary 5.1.1. *For any $c \in \mathcal{C}_T$ and $k \in \mathcal{C}_C$, C_3 and C_4 are continuous at $(F_{Y_1|C_1=c}, F_{Y_1|C_1=k})$ and $(F_{Y_1|C_1=c}, F_{Y_1|C_1=k}, F_{Y_2|C_2=k})$, respectively.*

Proof. Result follows directly from the lemma and definitions of C_3 and C_4 in terms of operators C_1, C_2, the fact that $F_{Y_1|C_1=c}$ is continuous and strictly increasing and thus $C_1(F_{Y_1|C_1=c})$ is continuous and finally because $\gamma_{c,k}$ is continuous (as a composition of two continuous functions). □

These two results lead to the following proposition (proof is in Appendix C):

Proposition 5.1. (Consistency) *For $c = \mathcal{C}$ and $t = 1, 2$ let the estimator $\hat{F}_{Y_t|C_t=c}$ be such that the following holds:*

$$\|\hat{F}_{Y_t|C_t=c} - F_{Y_t|C_t=c}\|_\infty = o_p(1) \tag{5.8}$$

Then $|\hat{\tau}_{c,k} - \tau_{c,k}| = o_p(1)$ and $\|\hat{\tau}_{c,k}(q) - \tau_{c,k}(q)\|_\infty = o_p(1)$.

Remark 5.1.1. Uniform consistency requirement in the proposition above holds for the empirical distribution function (by virtue of Glivenko-Cantelli theorem) but also for a variety of other estimators. This is the only restriction that we need to achieve consistency.

This result proves a strong (uniform) consistency of the quantiles, generalizing the consistency result from Athey and Imbens [2006].

5.1.3 Asymptotic normality, continuous case

Asymptotic normality will follow from the following facts:

Fact 5.1. *For any $F \in \mathcal{D}$ the map C_1 is Hadamard-differentiable tangentially to $C[a, b]$. The derivative (linear map) is given by the following function:*

$$D(C_1)|G(h) := -\frac{h}{g} \circ G^{-1} \tag{5.9}$$

where $g := G'$. If $g > c > 0$ then for any $h \in C[a, b]$ function $DC_{1|G}(h)$ is uniformly continuous.

Fact 5.2. *For any F, G function C_2 is Hadamard-differentiable and the derivative is given by the following function:*

$$D(C_2)_{|(F,G)}(h_1, h_2)(x) := h_1 \circ G(x) + F'_{G(x)}(h_2(x)) \tag{5.10}$$

For the proof of both facts check Van Der Vaart and Wellner [1996]. These two facts imply the differentiability of operators C_3 and C_4:

Lemma 5.2. *For any $c \in \mathcal{C}_T$ and $k \in \mathcal{C}_C$, operator C_3 is Hadamard-differentiable at $(F_{Y_1|C_1=c}, F_{Y_1|C_1=k})$ tangentially to $C[a,b] \times D[a,b]$; operator C_4 is Hadamard-differentiable at $(F_{Y_1|C_1=c}, F_{Y_1|C_1=k}, F_{Y_2|C_2=k})$ tangentially to $C[a,b] \times D[a,b] \times C[a,b]$. Explicit form of the derivatives is given in Appendix C.*

Proof. The result follows directly from the differentiability of C_1 and C_2 and chain rule for Hadamard differentiation. □

Hadamard differentiability and Donsker theorem for empirical distribution leads to the following proposition:

Proposition 5.2. (Asymptotic normality) *Let Assumption 5.1.1 hold. Let $\hat{F}_{Y_t|C_t=c}$ be the empirical distribution. Then the following is true:*

$$\begin{cases} \sqrt{n_{1c}+n_{1k}}\left(\hat{\gamma}_{c,k}(\cdot)-\gamma_{c,k}(\cdot)\right) \xrightarrow{w^\star} \mathbb{G}_{1,c,k} \\ \sqrt{n}\left(\hat{\tau}_{c,k}-\tau_{c,k}\right) \xrightarrow{w^\star} \mathcal{N}(0,V) \\ \sqrt{n}\left(\hat{\tau}_{c,k}(\cdot)-\tau_{c,k}(\cdot)\right) \xrightarrow{w^\star} \mathbb{G}_{2,c,k} \end{cases} \tag{5.11}$$

where $\mathbb{G}_{1,c,k}$ and $\mathbb{G}_{1,c,k}$ are centered Gaussian processes on $[a,b]$ and $[0,1]$ with covariance functions $\Phi_{\mathbb{G}_{1,c,k}}$ and $\Phi_{\mathbb{G}_{2,c,k}}$, respectively. Exact expressions for asymptotic variance and covariance functions are given in the appendix. Asymptotic distributions can be approximated by a standard nonparametric bootstrap algorithm.

The proof is in Appendix C. In light of the previous lemma, it follows by the functional delta method.

Remark 5.1.2. The result remains valid if instead of $\hat{F}_{Y_t|C_t=c}$ any other $\tilde{F}_{Y_t|C_t=c}$ is used as long as $\sqrt{n}\|\hat{F}_{Y_t|C_t=c} - \tilde{F}_{Y_t|C_t=c}\|_\infty = o_p(1)$.

5.1.4 W_∞ matching

To apply the matching algorithm, I need to estimate W_∞ distance. I use the following estimator:

$$\hat{W}_\infty(c,k) = \|\hat{F}^{-1}_{Y_1|C_1=c} - \hat{F}^{-1}_{Y_1|C_1=k}\|_\infty \tag{5.12}$$

For each $c \in \mathcal{C}_T$ define $\hat{k}(c) \in \mathcal{C}_c$ in the following way:

$$\hat{k}(c) \in \arg\min_k \hat{W}_\infty(c,k) \tag{5.13}$$

where ties can be broken arbitrarily. This results in the following estimator for the treatment effect:

$$\hat{\tau}_c = \frac{1}{n_{2c}} \sum_{i=1}^{n_{2c}} Y_{2ci} - \frac{1}{n_{2\hat{k}(c)}} \sum_{i=1}^{n_{2\hat{k}(c)}} \hat{\gamma}_{c,\hat{k}(c)} \left(Y_{2\hat{k}(c)i} \right) \tag{5.14}$$

The following lemma implies that this estimator has exactly the same asymptotic properties as $\hat{\tau}_{c,k(c)}$:

Lemma 5.3. *For any $c \in \mathcal{C}_T$ and $k \in \mathcal{C}_C$ we have:*

$$|\hat{W}_\infty(c,k) - W_\infty(c,k)| = o_p(1) \tag{5.15}$$

Proof. By Lemma 5.1 C_1 is continuous at $F_{Y_1|C_1=c}$ and $F_{Y_1|C_1=k}$ and supremum is a continuous function (with respect to uniform norm). Thus the result follows by continuous mapping theorem (for metric spaces). □

Using this lemmas and Slutsky lemma one can see that $\hat{k}(c)$ can be substituted with $k(c)$ without affecting any asymptotic properties. Similar construction can be used for difference in quantiles.

5.1.5 Testing

I focus on testing in a situation with a single treated cluster and two control clusters. It can be extended straightforwardly to the case with multiple treated and control clusters. Let $\mathcal{C}_T = \{1\}$ and $\mathcal{C}_C = \{2, 3\}$. In this case, we have two estimators for counterfactual distribution: we can use either $\hat{f}_{1,2}$ or $\hat{f}_{1,3}$. In this case, it is natural to base the test on the distribution of the pseudo-observations:

$$T := \sqrt{n_{22} + n_{23}} \left\| \frac{1}{n_{22}} \sum_{i=1}^{n_{22}} \{\hat{\gamma}_{1,2} < t\} - \frac{1}{n_{23}} \sum_{i=1}^{n_{22}} \{\hat{\gamma}_{1,3} < t\} \right\|_\infty \tag{5.16}$$

The following two lemmas imply that this statistic converges in distribution to a supremum of a centered Gaussian process.

Lemma 5.4. *For $\hat{f}_{1,c}$ defined above the following is true:*

$$\sqrt{n_{12} + n_{13}} \left(\hat{\gamma}_{1,2}(\gamma_{1,2}^{-1}(t)) - \hat{\gamma}_{1,3}(\gamma_{1,3}^{-1}(t)) \right) \xrightarrow{w^\star} \mathbb{G}_3 \tag{5.17}$$

where $\mathbb{G}_3$ *is a centered Gaussian process on* $[a, b]$*, with the following covariance function:*

$$\Phi_{\mathbb{G}_3}(x, y) = \frac{1}{\alpha}\Phi_2(x, y) + \frac{1}{(1-\alpha)}\Phi_3(x, y) \tag{5.18}$$

where $\alpha = \lim \frac{n_{12}}{n_{12}+n_{13}}$ *and* $\Phi_c(x, y)$ *is defined in the following way:*

$$\Phi_c(x, y) := \frac{F_{Y_1|C_1=c}(\gamma_{1,c}^{-1}(\max\{x, y\})) - F_{Y_1|C_1=c}(\gamma_{1,c}^{-1}(x))F_{Y_1|C_1=c}(\gamma_{1,c}^{-1}(y))}{f_{Y_1|C_1=1}(x)f_{Y_1|C_1=1}(y)} \tag{5.19}$$

Also, let $\mathbb{G}$ *be the weak limit of* $\sqrt{n_{11}}\left(\hat{F}_{Y_1|C_1=1} - F_{Y_1|C_1=1}\right)$*. Then* $\mathbb{G}$ *and* $\mathbb{G}_3$ *are independent.*

Lemma 5.5. *Difference of the empirical distributions converges to a centered Gaussian process:*

$$\sqrt{n_{22}+n_{23}}\left(\frac{1}{n_{22}}\sum_{i=1}^{n_{22}}\{\hat{\gamma}_{1,2} < \cdot\} - \frac{1}{n_{23}}\sum_{i=1}^{n_{22}}\{\hat{\gamma}_{1,3} < \cdot\}\right) \xrightarrow{w^\star} \mathbb{G}_{31} + \mathbb{G}_4 \tag{5.20}$$

where $\mathbb{G}_4$ *and* $\mathbb{G}_3$ *are independent centered Gaussian processes with covariance functions* $\Phi_{\mathbb{G}_4}$ *and* $\Phi_{\mathbb{G}_{31}}$*, respectively. Covariance function for* $\mathbb{G}_4$ *has the following form:*

$$\Phi_{\mathbb{G}_4}(x, y) = F_{Y_2(0,1)|C_2=1}(\max\{x, y\})) - F_{Y_2(0,1)|C_2=1}(x)F_{Y_2(0,1)|C_2=1}(y) \tag{5.21}$$

where $F_{Y_2(0,1)|C_1=1}$ *is the counterfactual distribution. Covariance function of* $\mathbb{G}_{3,1}$ *satisfies the following restriction:* $\Phi_{\mathbb{G}_{3,1}}(x, y) = \lambda f_{Y_2(0,1)|C_2=1}(x)f_{Y_2(0,1)|C_2=1}(y)\Phi_{\mathbb{G}_{31}}(x, y)$*, where* $\lambda = \lim \frac{n_{22}+n_{23}}{n_{12}+n_{13}}$*.*

Proofs of both lemmas are in Appendix C. The first lemma is a straightforward consequence of Proposition 5.2. The second lemma follows from the results in van der Vaart and Wellner [2007]. The second lemma together with continuous mapping theorem implies that T converges to a supremum of $\mathbb{G}_{31} + \mathbb{G}_4$. The limit distribution depends on the unknown parameters and thus should be approximated. In Appendix D, I propose a bootstrap algorithm that can be used for such approximation.

5.1.6 Semi-continuous case

I focus on showing that we can consistently estimate the support of the monotone transportation map. Let $S_1, S_2 \subseteq [a, b]^2$ we can define (Hausdorff distance) $d_H(S_1, S_2)$ in a standard way. For

any $F_1, F_2, G_1, G_2 \in D_1[a,b]$ define:

$$\begin{cases} S(F_1, G_1) := \{(x,y) \in [a,b] : (F^{-1}(p), G^{-1}(p), p \in [0,1]\} \\ d(F_1, G_1, F_2, G_2) := d_H(S(F_1, G_1), S(F_2, G_2)) \end{cases} \tag{5.22}$$

Consistency is handled by the following proposition:

Proposition 5.3. *Assume that distribution functions are uniformly consistent:*

$$\|\hat{F}_{Y_1|C_1=c} - F_{Y_1|C_1=c}\|_\infty = o_p(1) \tag{5.23}$$

and for $c = 1, 2$ $F_{Y_1|C_1=c}$ is strictly increasing on $[a,b]$. Then the following is true:

$$d(F_{Y_1|C_1=1}, F_{Y_1|C_1=2}, \hat{F}_{Y_1|C_1=1}, \hat{F}_{Y_1|C_1=2}) = o_p(1) \tag{5.24}$$

Proof. Consider any $G_1, G_2 \in D_1[a,b]$ such that:

$$\begin{cases} \|F_{Y_1|C_1=1} - G_1\|_\infty < \varepsilon \\ \|F_{Y_1|C_1=2} - G_2\|_\infty < \varepsilon \end{cases} \tag{5.25}$$

Since for $c = 1, 2$ $F_{Y_1|C_1=c}$ is strictly increasing and has strictly positive density Lemma 5.1 we get the following:

$$\begin{cases} \|C_1(F_{Y_1|C_1=1}) - C_1(G_1)\|_\infty < K\varepsilon \\ \|C_1(F_{Y_1|C_1=2}) - C_1(G_2)\|_\infty < K\varepsilon \end{cases} \tag{5.26}$$

And by definition it follows that $d(F_{Y_1|C_1=1}, F_{Y_1|C_1=2}, G_1, G_2) < K\varepsilon$. This implies that

$$d(F_{Y_1|C_1=1}, F_{Y_1|C_1=2}, \cdot, \cdot) : D_1[a,b]^2 \to \mathbb{R}_+ \tag{5.27}$$

is continuous at $(F_{Y_1|C_1=1}, F_{Y_1|C_1=2})$ and the result follows by the continuous mapping theorem (and uniform consistency). □

5.2 Multi-dimensional case

5.2.1 Estimation and inference

Assumption 5.2.1. (Caffarelli's Regularity) *For $c = 1, 2$ $\mu_{Y_1|C_1=c}$ are supported on $[0,1]^K$ and admit densities (with respect to Lebegues measure on $\mathbb{R}^K$) which lie in $C^k[0,1]^K$ and are bounded away from zero and infinity.*

Remark 5.2.1. In this assumption $[0,1]^K$ can be substituted for arbitrary closed convex compact in $\mathbb{R}^K$ with non-empty interior and smooth boundary.

Define $\Psi := \{\psi : [0,1]^K \to : \psi(0) = 0, \psi \text{ is convex}\}$. For each $\psi \in \Psi$ let $\psi^\star$ be the restriction of convex conjugate of ψ to $[0,1]^K$. Estimation strategy is based on the solution of the dual transportation problem:

$$\tilde{\psi} := \arg\inf_{\psi\in\Psi} \left(\int \psi(x) d\mu_{Y_1|C_1=1} + \int \psi^\star(u) d\mu_{Y_1|C_1=2} \right) \tag{5.28}$$

The main reason why we are interested in $\tilde{\psi}$ is because of the following fact:

Fact 5.3. *If Assumption 5.2.1 holds then (a) $\tilde{\psi}$ defined above belongs to $C^{k+2}[0,1]^K$ and (b) $Q(x) := \nabla\tilde{\psi}(x)$ solves the optimal quadratic transportation problem between $\mu_{Y_1|C_1=2}$ and $\mu_{Y_1|C_1=1}$*

This fact is useful for two reasons. First, it implies that solution of the transportation problem can be recovered from the solution of the dual problem. Second, it connects smoothness assumptions on measures with the regularity of Q. This fact might be used in statistical analysis. In this paper I do not exploit this point, leaving it for future research. This result suggests the following way of estimating γ: first, solve the dual problem and then compute the gradient. In particular, let $\hat{\mu}_{Y_1|C_1=c}$ be some estimators of $\mu_{Y_1|C_1=c}$ and consider the following estimators:

$$\begin{cases} \hat{\psi} \in \arg\inf_{\psi\in\Psi} \left(\int \psi(x) d\hat{\mu}_{Y_1|C_1=1} + \int \psi^\star(u) d\hat{\mu}_{Y_1|C_1=2} \right) \\ \hat{\gamma}(x) := \arg\sup_{x\in[0,1]^K} \{x^T y - \hat{\psi}(y)\} \end{cases} \tag{5.29}$$

Solution for this program can be found by a variety of algorithms, check Chernozhukov et al. [2017] for details.

If estimators $\hat{\mu}$ are good enough (uniformly consistent) then we have the following result from Chernozhukov et al. [2017]:

Fact 5.4. *Assume that for $c = 1, 2$ $d_{BL}(\hat{\mu}_{Y_1|C_1=c}, \mu_{Y_1|C_1=c}) = o_p(1)$, let Assumption 5.2.1 hold. Then $\|\hat{\gamma} - \gamma\|_\infty = o_p(1)$.*

This uniform consistency result implies that for any continuous function $g : \mathbb{R}^K \to \mathbb{R}$ we have the following:

$$\mathbb{E}[|g(\hat{\gamma}(Y_2)) - g(\gamma(Y_2))||C_2 = 2] \leq \|g \circ \hat{\gamma} - g \circ \gamma\|_\infty = o_p(1) \tag{5.30}$$

where the last equality follows because g is uniformly continuous on $[0,1]^K$ (compact set) and thus $\|g \circ \hat{\gamma} - g \circ \gamma\| = O_p(\|\hat{\gamma} - \gamma\|_\infty)$. This implies that any moments of the counterfactual distribution can be estimated consistently.

5.2.2 Testing

Results from the Section 3 imply that the test for univariance and monotonicity should be based on the difference between copulas of two distributions. I describe it in case of two-dimensional outcomes, the extension to the multidimensional case is straightforward.

Define $C_c(q,p) := F_{Y_1|C_1=c}(F^{-1}_{Y_{11}|C_1=c}(q), F^{-1}_{Y_{12}|C_1=c}(p))$. Let $\hat{F}_{Y_1|C_1=c}$ and $\hat{F}^{-1}_{Y_{1k}|C_1=c}$ be the empirical distribution function and empirical quantile functions, respectively. Using this we can define empirical copula:

$$\hat{C}_c(q,p) := \hat{F}_{Y_1|C_1=c}(\hat{F}^{-1}_{Y_{11}|C_1=c}(q), \hat{F}^{-1}_{Y_{12}|C_1=c}(p)) \tag{5.31}$$

It is well-known that under smoothness assumptions on C_c $\hat{C}_c(q,p)$ converges (once appropriately scaled) to a limit process. In particular, the following fact is true (see Van Der Vaart and Wellner [1996]).

Fact 5.5. *Assume that function C_c is continuously differentiable on $[0,1]^2$, then $\sqrt{n_{1c}}(\hat{C}_1 - C_1) \to \mathbb{C}$, where $\mathbb{C}$ is a centered Gaussian process (indexed by $[0,1]^2$).*

Using this fact we can base the test on the following statistic:

$$T := \int (\hat{C}_1(x) - \hat{C}_2(x))^2 dx \tag{5.32}$$

It follows that the scaled version of T has the asymptotic distribution of $\int_{[0,1]^2} \mathbb{C}^2 dx$. Since covariance function of $\mathbb{C}$ depends on the unknowns this distribution should be approximated. In Rémillard and Scaillet [2009] authors suggest a particular scheme based on multiplier central limit theorem.

In principle, any other test statistic can be used, for example, Kolmogorov-Smirnov type statistic, where L^2 distance is substituted with sup-norm. If the statistic is a continuous function of the copulas, then its asymptotic distribution follows from the continuous mapping theorem. Then the algorithm in Rémillard and Scaillet [2009] can be used to approximate its distribution.

6 Empirical example

In this section, I apply my methodology to the data from Engelhardt and Gruber [2011]. In this paper, authors analyze the introduction of Medicare Part D in 2006. The Medicare Modernization Act of 2003 introduced a new benefit to a Medicare system providing coverage for prescription drugs. The central question that the paper addresses is whether this new (at the time) system increased welfare by reducing the financial risk for elderly or simply redistributed money within the insurance system.

The data consists of 2002-2005 and 2007 waves of the MEPS, which is a nationally representative set of respondents drawn from the National Health Interview Survey (NHIS). The MEPS is a two-year overlapping panel focused on health insurance coverage, health care utilization, and expenditure, and is used to construct data for the National Health Accounts.

In the data, we observe two periods and two clusters. The first period corresponds to 2002-2005 before Part D was introduced and the second period is the year 2007. For both periods we observe a population of Medicare-eligible people (age 65 to 70) and a population of near-elderly (age 60-64) who aren't eligible for Medicare. The primary variable of interest is the out-of-pocket prescription-drug spending. We also have information on demographic covariates (race, sex, etc.).

In this case, it is natural to view the unobservable ν_t as an underlying health characteristic. Since we are focused on the out-of-pocket spendings and the market environment (insurance) is different for Medicare-eligible individuals it is natural to assume that $h(0, c, \nu_t)$ depends on c. Monotonicity requirement is reasonable in this case: we expect healthier individuals to spend less on the prescription drugs irrespective of their eligibility status. The cutoff rule defines the cluster assignment, and it seems likely that the underlying health characteristics do not change sharply at the age of 65. This situation makes independence assumption plausible, at least for individuals who are close the cutoff.

In the paper, this data is analyzed using different methods. The closest one to what I'm proposing is the difference in quantiles regression (see Table 8 in Engelhardt and Gruber [2011]). The central structural assumption underlying this method is that the difference between quantiles in two populations should stay constant over time in the absence of policy intervention. This a restrictive assumption that implies that the quantile function of the counterfactual data is the sum of a quantile function for the control cluster and a difference in quantiles in the first period.

Since the difference in quantiles doesn't need to be monotonic, there is no guarantee that this sum is a bona fide quantile function. In general, it is not clear what assumptions on primitives would imply this identification strategy.

I use the version of the model with one-dimensional unobservables, with and without covariates. Following the paper, I use the out-of-pocket spendings as the outcome variable. We observe several outcomes in the data and thus can potentially apply the multi-dimensional model. At the same time, these outcomes are all different variants of spendings and seem to capture the same underlying heterogeneity. Results for the one-dimensional model are presented in the Table 1.

Table 1: Quantile treatment effects

Quantile	10%	20%	30%	40%	50%	60%	70%	80%	90%
Unconditional	-15.28	-28.58	-45.15	-97.34	-135.51	-186.82	-282.49	-345.90	-660.98
	(3.88)	(6.15)	(26.37)	(50.42)	(81.78)	(107.62)	(108.82)	(210.76)	(317.24)
Conditional	-15.71	3.66	-27.35	-61.99	-120.64	-162.80	-246.99	-305.66	-506.19
	(3.30)	(21.55)	(15.95)	(20.65)	(30.34)	(36.71)	(49.69)	(65.60)	(151.41)

[a] Bootstrap standard errors in parenthesis.

[b] Conditional estimates are based on race: estimation in the first period is done keeping race constant, then these results are aggregated in the second period.

Qualitatively results are similar to Engelhardt and Gruber [2011] but are smaller in magnitude, especially for top quantiles. Conditional effects are estimated more precisely, roughly with the same standard error as in the paper.

7 Conclusion

In this paper, I constructed a new nonlinear model for diff-in-diff empirical settings. The model can be used to address two different sources of bias that are likely present in these frameworks: selection and technological bias. Selection bias can be dealt with using standard methods, while technological bias requires structural assumptions on the underlying production functions. Natural assumptions are available in the case when the underlying heterogeneity is one-dimensional. With multi-dimensional heterogeneity, the situation is considerably more complicated.

In the paper, I present three different approaches to identification with multi-dimensional heterogeneity using multiple outcomes. These strategies are different in motivation; each has its advantages and limitations. I believe that further research is needed in this direction. A

particularly attractive approach is to formulate new restrictions on the transportation maps that might lead either to partial or exact identification. This plan is especially feasible in the discrete framework. In this case, we will face the same problem as in the extension in Section 4: even if the unique transportation plan is selected the counterfactual distribution is only partially identified. The discrete framework is also tractable from the statistical point of view because the optimal transportation map is defined as a solution of the finite-dimensional linear program and for such problems statistical guarantees can be obtained.

References

A. Abadie. Semiparametric instrumental variable estimation of treatment response models. Journal of econometrics, 113(2):231–263, 2003.

J. G. Altonji and R. L. Matzkin. Cross section and panel data estimators for nonseparable models with endogenous regressors. Econometrica, 73(4):1053–1102, 2005.

S. Athey and G. W. Imbens. Identification and inference in nonlinear difference-in-differences models. Econometrica, 74(2):431–497, 2006.

S. Boyd and L. Vandenberghe. Convex optimization. Cambridge university press, 2004.

G. Carlier, A. Galichon, and F. Santambrogio. From knothe's transport to brenier's map and a continuation method for optimal transport. SIAM Journal on Mathematical Analysis, 41(6): 2554–2576, 2010.

G. Carlier, V. Chernozhukov, A. Galichon, et al. Vector quantile regression: an optimal transport approach. The Annals of Statistics, 44(3):1165–1192, 2016.

V. Chernozhukov, A. Galichon, M. Henry, and B. Pass. Single market nonparametric identification of multi-attribute hedonic equilibrium models. Working paper, 2014.

V. Chernozhukov, A. Galichon, M. Hallin, M. Henry, et al. Monge - kantorovich depth, quantiles, ranks and signs. The Annals of Statistics, 45(1):223–256, 2017.

D. Dowson and B. Landau. The fréchet distance between multivariate normal distributions. Journal of multivariate analysis, 12(3):450–455, 1982.

G. V. Engelhardt and J. Gruber. Medicare part d and the financial protection of the elderly. American Economic Journal: Economic Policy, 3(4):77–102, 2011.

A. Galichon. Optimal Transport Methods in Economics. Princeton University Press, 2016.

G. W. Imbens and J. D. Angrist. Identification and estimation of local average treatment effects. Econometrica, 62(2):467–475, 1994.

G. W. Imbens and W. K. Newey. Identification and estimation of triangular simultaneous equations models without additivity. Econometrica, 77(5):1481–1512, 2009.

G. W. Imbens and D. B. Rubin. Estimating outcome distributions for compliers in instrumental variables models. The Review of Economic Studies, 64(4):555–574, 1997.

G. W. Imbens and D. B. Rubin. Causal Inference in Statistics, Social, and Biomedical Sciences. Cambridge University Press, 2015.

A. Kolmogorov and S. Fomin. Elements of function theory and functional analysis. Nauka, Moscow, 1968.

R. L. Matzkin. Nonparametric estimation of nonadditive random functions. Econometrica, 71 (5):1339–1375, 2003.

B. Rémillard and O. Scaillet. Testing for equality between two copulas. Journal of Multivariate Analysis, 100(3):377–386, 2009.

F. Santambrogio. Optimal transport for applied mathematicians. Birkäuser, NY, 2015.

B. W. Silverman et al. Weak and strong uniform consistency of the kernel estimate of a density and its derivatives. The Annals of Statistics, 6(1):177–184, 1978.

A. W. Van der Vaart. Asymptotic statistics, volume 3. Cambridge university press, 1998.

A. W. Van Der Vaart and J. A. Wellner. Weak convergence. In Weak Convergence and Empirical Processes, pages 16–28. Springer, 1996.

A. W. van der Vaart and J. A. Wellner. Empirical processes indexed by estimated functions. Lecture Notes-Monograph Series, pages 234–252, 2007.

C. Villani. Optimal transport: old and new, volume 338. Springer Science & Business Media, 2008.

A Single population

A.1 Restriction on the joint distribution

So far, I have ignored the joint distribution of the data in two periods. This distribution makes sense, only when we are analyzing the same population in different time periods. It doesn't exist if the time periods correspond to different cohorts of people.

I start this subsection assuming that we don't have access to any covariates and also let the Assumption 2.3.1 hold in the first period: $\nu_1 \perp C_1$. I focus on the assumptions that will guarantee that the independence holds in the second period as well: $\nu_2 \perp C_2$. These assumptions turn out to be more restrictive in the repeated cross-section case, emphasizing the value of the panel data.

I start specifying the counterfactual transition probabilities:

$$\begin{cases} \mathbb{E}[\{\nu_2 \in A\}|\nu_1] = \int_{\nu_2 \in A} \tilde{K}(\nu_2|c_2, c_1, \nu_1) d\nu_1 \\ \mathbb{E}[\{C_2 = 1\}|\nu_1, \nu_2] = \tilde{p}(\nu_2, c_1, \nu_1) \end{cases} \tag{1.1}$$

Function $K(\cdot|\cdot) : V^2 \times \{0,1\}^2 \to \mathbb{R}_+$ is the counterfactual probability kernel that describes the evolution of unobservables depending on the counterfactual cluster assignment. Function $p(\cdot) : V^2 \times \{0,1\} \to [0,1]$ is the counterfactual probability that describes the evolution of the cluster assignment. Transitions in the data are generated if we substitute c_1 with C_1.

As a first step, I put an exclusion restriction on the counterfactual transitions:

Assumption A.1.1. (Sequential exogeneity) *Counterfactual transitions depend only on the information available in the first period:*

$$\begin{cases} \tilde{K}(\nu_2|c_2, c_1, \nu_1) = K(\nu_2|c_1, \nu_1) \\ \tilde{p}(\nu_2, c_1, \nu_1) = p(c_1, \nu_1) \end{cases} \tag{1.2}$$

This is a standard assumption made in the panel data literature. It can be separated into two parts: the first is saying the the selection into clusters occurs before ν_2 is known and the second is saying that ν_2 doesn't depend on the cluster assignment. This assumption is reasonable if we believe that ν completely characterizes the subjects. This may be restrictive if we put some additional structure on ν (e.g., univariance).

The following independence restriction is a straightforward corollary of the Assumption A.1.1.

Corollary A.1.1. (Conditional independence) *C_2 and ν_2 are conditionally independent:*

$$\nu_2 \perp C_2 | \nu_1, C_1 \tag{1.3}$$

This simple result emphasizes the connection between the independence in the second period and information in the first period. It also allows me to systemize the application depending on the properties of $p(C_1, \nu_1)$.

A.1.1 Overlap

In the first case that I consider, I assume that the following overlap condition holds:

Assumption A.1.2. (Cluster overlap) *For any l, x $0 < \mathbb{P}(C_2 = 1 | C_1 = l, \nu_1 = x) < 1$.*

One might achieve the full identification even if this assumption holds only for some l, x. I assume the more restrictive version to make identification as simple as possible, it can be adapted given a particular application.

Define the propensity score: $p_k(l, x) = \mathbb{P}(C_2 = k | C_1 = l, \nu_1 = x)$. Then we have the following classical lemma (see Imbens and Rubin [2015] for a thorough discussion on the role of the propensity score methods in causal inference):

Lemma A.1. (Reweighting) *Let Assumption A.1.1 hold; fix a measurable function f, set A and k and define the following random variable:*

$$Z(f, k, A) = \frac{\{f(\nu_2) \in A, C_2 = k\}}{p_k(C_1, \nu_1)} \tag{1.4}$$

Then $\mathbb{E}[Z(f, k, A)] = \mathbb{E}[\{f(\nu_2) \in A\}]$.

Proof. The proof is standard:

$$\mathbb{E}[Z(f, k, A)] = \mathbb{E}\left[\frac{\{f(\nu_2) \in A, C_2 = k\}}{p_k(C_1, \nu_1)}\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{\{f(\nu_2) \in A, C_2 = k\}}{p_k(C_1, \nu_1)}\right] | C_1, \nu_1\right] =$$
$$\mathbb{E}\left[\mathbb{E}\left[\{f(\nu_2) \in A\} | C_1, \nu_1\right] \mathbb{E}\left[\frac{\{C_2 = k\}}{p_k(C_1, \nu_1)} | C_1, \nu_1\right]\right] = \mathbb{E}[\{f(\nu_2) \in A\}] \tag{1.5}$$

where the third equality from the sequential exogeneity. □

This lemma implies that for any function $f(\nu_2)$ we can compute its distribution from the conditional distribution using appropriate weights. Applying this result to potential outcomes we get the following corollary.

Corollary A.1.2. *Let Assumptions 2.3.2, A.1.2 hold, then $\mu_{Y_2(0,2)}$ and $\mu_{Y_2(1,1)}$ are identified.*

Proof. I show the identification of $\mu_{Y_2(0,2)}$, the proof for the second measure is the same. Assumption A.1.2 justifies the use of the Lemma A.1:

$$\mathbb{E}[\{Y_2(0,2) \in A\}] = \mathbb{E}[\{h(0,2,\nu_2) \in A\}] = \mathbb{E}\left[\frac{\{h(0,2,\nu_2) \in A, C_2 = 2\}}{p_2(C_1,\nu_1)}\right] = \mathbb{E}\left[\frac{\{h(0,C_2,\nu_2) \in A, C_2 = 2\}}{p_2(C_1,\nu_1)}\right] = \mathbb{E}\left[\frac{\{Y_2 \in A, C_2 = 2\}}{p_2(C_1,\nu_1)}\right] \quad (1.6)$$

where the second equality follows from the lemma. Assumption 2.3.2 implies that we can define $\nu_1 = h^{-1}(0, C_1, Y_1)$ leading to $p_2(C_1,\nu_1) = p_2(C_1, h^{-1}(0, C_1, Y_1))$ and thus $\frac{\{Y_2 \in A, C_2=2\}}{p_2(C_1,\nu_1)}$ can be constructed. □

This result has the same role as the Corollary 2.1, implying that in the second period under the sequential exogeneity and overlap assumptions we don't need to assume independence, but instead can guarantee it. Of course, this strategy is unavailable in the case where we only observe a repeated cross-section, because we can't identify the propensity score.

A.1.2 No overlap

I consider the situation with no overlap, in particular, constant assignment $C_1 = C_2$:

Assumption A.1.3. (Constant assignment) *Cluster assignment stays constant: $C_1 = C_2$.*

With this assumption we need to restrict the counterfactual transition function $K_1(\nu_2|c_1,\nu_1)$. I assume the following exclusion restriction:

Assumption A.1.4. (No learning) *Counterfactual transition function $K_1(\nu_2|c_1,\nu_1)$ doesn't depend on c_1: $K_1(\nu_2|c_1,\nu_1) = \hat{K}_1(\nu_2|\nu_1)$ for some function $\hat{K}_1$.*

This assumption doesn't allow for causal effect of cluster assignment on future unobservables. This assumption essentially prohibits learning. Its validity depends on the application, definitions on clusters and meaning of ν. The main consequence of this assumptions is that independence in the first period implies independence in the second period:

Corollary A.1.3. *Assume that $\nu_1 \perp C_1$, let Assumptions A.1.4 and A.1.1 hold. Then $\nu_2 \perp C_2$.*

There is no direct relationship between Assumptions A.1.4 and A.1.3. At the same time, if we have at least some degree of overlap (Assumption A.1.2 holds partially) then there is no reason to use Assumption A.1.4 for identification, because it can be achieved without it.

A.1.3 Repeated cross-section

If the data that we observe have no information about the joint distribution of outcomes (repeated cross-section) then even if the overlap assumption holds we can't use the propensity score to reweigh the observations in the second period. If we know that $C_1 = C_2$, then we need Assumption A.1.4 to hold. If we know that $C_1 \neq C_2$, then we can either let Assumption A.1.4 to hold, or use the following:

Assumption A.1.5. (Independent assignment) *Cluster assignment in the second period is independent of the history:* $C_2 \perp (C_1, \nu_1)$

Under the last assumption the propensity score is constant (doesn't depend on (C_1, ν_1)) and thus we can formally apply Lemma A.1.

One can summarize the identification results in the following way: if the cluster assignment is constant then there is no difference between repeated cross-section and panel data in terms of identification. It can be achieved under a restrictive condition on the transition kernel that forbids learning.

If the cluster change over time then there is a crucial difference between different data structures. With the panel data independence can be achieved using propensity score methods using the sequential exogeneity and overlap. With the repeated cross-section we need to make a restrictive assumption on the propensity score to achieve independence.

This logic has a straightforward generalization for the case with time-constant covariates. In this case, the covariates can be included into Assumptions A.1.1,A.1.3,A.1.4, and A.1.5, leading to the conditional version of these restrictions. In turn all the identification results will hold conditional on this covariate.

B Additional identification results

Proof of Proposition 3.5: Define $\lambda_c = \mu_{Y_1(0,c)}$. Let $T^\star$ be the solution of the following program:

$$T^\star = \arg \min_{T:\mu_{Y_1(0,1)}=T_\# \mu_{Y_1(0,1)}} \mathbb{E}[\|Y_1(0,2) - T(Y_1(0,2))\|^2] \tag{2.1}$$

By the assumptions the solution exists and is unique (Villani [2008]). For each $x \in \text{supp}\{\mu_{Y_1(0,2)}\}$ define the curve (displacement interpolation) $\pi_t(x) = (1-t)x + tT^\star(x)$. Then it is known that (constant speed) geodesic between λ_1 and λ_2 is given by $\mu_t := (\pi_t)_\# \lambda_2$ for $t \in [0,1]$ (see Santambrogio [2015]).

Next, take any $t \in [0,1]$ and let $\mu_{\nu_1} = \mu_t$. Standard results about geodesics in $W_2(\Omega)$ (see Santambrogio [2015]) imply that $h_{\mu_t}^{-1}(2) = \pi_t(x)$ and $h_{\mu_t}(1) = T^\star(\pi_t(x))$. Combining this we get that $\gamma_{\mu_t} = T^\star$. Since t was arbitrary this implies the result.

B.1 Covariates

Covariates can be included in the multidimensional model in exactly the same way as before. The potential outcome function is adjusted in the same way:

$$Y_t(0,c) = h(0,c,\nu_t,X_t) \tag{2.2}$$

Multivariance assumption is adapted in the following way:

Assumption B.1.1. (Multivariance with covariance) *For $c = 1,2$ and $x \in X$ function $h(0,c,\cdot,x) : V \to \mathbb{R}^K$ is a bijection.*

In the case with covariates the high-level restrictions should be placed on $\gamma_x^\star = h(0,1,\cdot,x) \circ h^{-1}(0,2,\cdot,x)$. Assumptions are adapted in a straightforward way:

Assumption B.1.2. (Multidimensional monotonicity with covariates) *For each $x \in X$ function $\gamma_x^\star := h(0,1,\cdot,x) \circ h^{-1}(0,2,\cdot,c)$ is monotone (as operator in $\mathbb{R}^K$). In particular, for any $z_1, z_2 \in \mathbb{R}^K$ the following restriction holds:*

$$(z_1 - z_2, \gamma_x^\star(z_1) - \gamma_x^\star(z_2)) > 0 \tag{2.3}$$

Assumption B.1.3. (Symmetry with covariates) *For each $x \in X$ function $\gamma_x : \mathbb{R}^K \to \mathbb{R}^K$ is differentiable and has a symmetric Jacobian:*

$$\frac{\partial \gamma^\star_{k,x}(z)}{\partial z_l} = \frac{\partial \gamma^\star_{l,x}(z)}{\partial z_k} \tag{2.4}$$

for all k, l.

Assumption B.1.4. (Technical conditions with covariates) *For $c = 1, 2$ and $x \in X$ measure $\mu_{Y_1(0,c)|X=x}$ is absolutely continuous with respect to $\lambda(\mathbb{R}^K)$; for $c = 1, 2$ outcomes are square-integrable: $\mathbb{E}[\|Y_1(0,c)\|^2|X = x] < \infty$; densities $f_{Y_1(0,c)|x}$ are supported on the open, bounded and convex regions $\Lambda_c(x)$, bounded from below and above and belong to $C^\alpha(\Lambda_c(x))$ for some α.*

As a result of this assumptions we have a conditional analog of the identification result:

Proposition B.1. (Identification with covariates) *Let Assumptions 2.4.1, B.1.1,B.1.2,B.1.3,B.1.4 hold. Then the function $\gamma_x := (\gamma^\star_x)_{|\Lambda_2}$ is identified as a solution to the following transportation problem:*

$$\gamma := \underset{T:\mu_{Y_1(0,1)|X=x} = T_\# \mu_{Y_1(0,2)|X=x}}{\arg\min} \mathbb{E}[\|Y_1(0,2) - T(Y_1(0,2))\|^2|X = x] \tag{2.5}$$

If additionally Assumption 2.4.3 holds then $\mu_{Y_2(0,1)|C_2=1,X} = (\gamma_x)_\# \mu_{Y_2|C_2=1,X}$ and thus the counterfactual distribution is identified.

From the practical viewpoint, conditional identification is not very useful if we have continuous covariates because the general nonparametric approach would require a tremendous amount of data. As a result, it is essential to have a setting that can be used in practice. I present a particular example below.

Example B.1.1. (Multivariate normal outcomes) Let $\mu_{Y_1(0,c)|X=x} = \mathcal{N}(m_c(x), \Sigma_c(x))$ – outcomes in the first period have a conditional multivariate normal distribution. In this case, a known result (for example, see Dowson and Landau [1982]) implies that γ_x has the following form:

$$\gamma_x(y) = m_1(x) + \Sigma_1^{\frac{1}{2}}(x) \left(\Sigma_1^{\frac{1}{2}}(x)\Sigma_2(x)\Sigma_1^{\frac{1}{2}}(x)\right)^{-\frac{1}{2}} \Sigma_2^{\frac{1}{2}}(x)(y - m_2(x)) \tag{2.6}$$

This is a direct generalization of the standard diff-in-diff algorithm to multiple outcomes.

It is clear that the example above can be generalized to a semiparametric location-scale family, where we can use other generating distribution besides the normal one.

C Statistical proofs

Proof of Lemma 5.1: I start with the first claim. Let F_1 be arbitrary strictly increasing function on $[a,b]$, fix $\varepsilon > 0$ and let $g \in D[a,b]$ be such that $\|F_1 - g\|_\infty < \varepsilon$ and g is a distribution function on $[a,b]$. By construction for any $x, y \in [a,b]$ and $x > y$ we have the following:

$$F_1(x) \geq F_1(y) + \int_y^x f_1(\tau)d\tau \geq F_1(y) + c(x-y) \tag{3.1}$$

Consider any $y \leq b - \frac{4\varepsilon}{c}$, $x \geq a + \frac{5\varepsilon}{c}$ such that $x > y$. Then we have the followin

$$\begin{cases} g(x) \geq F_1(x) - \varepsilon > F_1(y) + \frac{c}{2}(x-y) - \varepsilon \\ g(y) \leq F_1(y) + \varepsilon < F_1(x) - \frac{c}{2}(x-y) + \varepsilon \end{cases} \Rightarrow \begin{cases} g\left(y + \frac{4\varepsilon}{c}\right) > F_1(y) + \varepsilon \\ g\left(x - \frac{4\varepsilon}{c}\right) < F_1(x) - \varepsilon \end{cases} \tag{3.2}$$

Next, take any $p \in [0,1]$. Let $x := C_1(F)(p)$, by definition it follows that $F(x) \geq p$. If $x \geq b - \frac{4\varepsilon}{c}$, then it follows that $C_1(g)(p) \leq b$. Otherwise, by the first inequality above it follows that $g\left(x + \frac{4\varepsilon}{c}\right) > p$. As a result, $C_1(g)(p) < x + \frac{2\varepsilon}{c}$ or $C_1(g)(p) \leq x + \frac{4\varepsilon}{c}$. Next, if $F(x) = p$ and $x \geq a + \frac{5\varepsilon}{c}$ then the second inequality implies that $C_1(g)(p) > x - \frac{2\varepsilon}{c}$. If $F(x) > p$ then it follows that $F(x-) \leq p$, and taking left limits in the second inequality we get the following:

$$g\left(\left(x - \frac{4\varepsilon}{c}\right)-\right) \leq F_1(x-) - \varepsilon < p \tag{3.3}$$

This implies that $g(x - \frac{5\varepsilon}{c}) < p$ and thus it follows that $C_1(g)(p) > x - \frac{5\varepsilon}{c}$. This implies that $\|C_1(g) - C_1(F_1)\|_\infty \leq K\varepsilon$ proving the first claim.

To prove the second claim observe the following:

$$\|g_2(g_3(x)) - F_2(F_1(x))\|_\infty \leq \|g_2(g_3(x)) - F_2(g_3(x))\|_\infty + \|F_2(g_3(x)) - F_2(F_1(x))\|_\infty \tag{3.4}$$

The first summand is less that $\|F_2 - g_2\|_\infty < \varepsilon$, the second summand can be made arbitrarily small because F_2 is continuous by assumption and thus is uniformly continuous because $[a,b]$ is compact.

Proof of Proposition 5.1: The estimator of $\hat{\tau}$ has the following form:

$$\hat{\tau} - \tau = \frac{1}{N_{21}} \sum_{i=1}^{N_{21}} Y_{21i} - \frac{1}{N_{21}} \sum_{i=1}^{N_{21}} \hat{\gamma}(Y_{22i}) - \mathbb{E}[Y_2|C_2 = 1] + \mathbb{E}[\gamma(Y_2)|C_2 = 2] +$$
$$\frac{1}{N_{21}} \sum_{i=1}^{N_{21}} \gamma(Y_{22i}) - \frac{1}{N_{21}} \sum_{i=1}^{N_{21}} \gamma(Y_{22i}) \Rightarrow$$
$$|\hat{\tau} - \tau| \leq \left| \frac{1}{N_{21}} \sum_{i=1}^{N_{21}} Y_{21i} - \mathbb{E}[Y_2|C_2 = 1] \right| + \left| \frac{1}{N_{21}} \sum_{i=1}^{N_{21}} \hat{\gamma}(Y_{22i}) - \frac{1}{N_{21}} \sum_{i=1}^{N_{21}} \gamma(Y_{22i}) \right| +$$
$$\left| \mathbb{E}[\gamma(Y_2)|C_2 = 2] - \frac{1}{N_{21}} \sum_{i=1}^{N_{21}} \gamma(Y_{22i}) \right| \leq \left| \frac{1}{N_{21}} \sum_{i=1}^{N_{21}} Y_{21i} - \mathbb{E}[Y_2|C_2 = 1] \right| +$$
$$\|\hat{f} - f\|_\infty + \left| \mathbb{E}[\gamma(Y_2)|C_2 = 2] - \frac{1}{N_{21}} \sum_{i=1}^{N_{21}} \gamma(Y_{22i}) \right| = o_p(1) + o_p(1) + o_p(1) \quad (3.5)$$

The last equality follows by the law of large numbers and continuous mapping theorem (for general metric spaces) together with Corollary 5.1.1.

The quantile result follows similarly:

$$\|\hat{\tau}(q) - \tau(q)\|_\infty \leq \|\hat{F}^{-1}_{Y_2|C_2=1}(q) - F^{-1}_{Y_2|C_2=1}(q)\|_\infty +$$
$$\|\gamma \circ F^{-1}_{Y_2|C_2=2}(q) - \hat{\gamma} \circ \hat{F}^{-1}_{Y_2|C_2=2}(q)\|_\infty = o_p(1) + o_p(1) \quad (3.6)$$

where results follow from Corollary 5.1.1 and continuous mapping theorem.

Derivatives for Lemma 5.2: Derivative for C_3 has the following form:

$$D(C_3)_{|(F_{Y_1|C_1=1},F_{Y_1|C_1=2})}(h_1,h_2)(x) = -\frac{h_1}{f_{Y_1|C_1=1}} \circ F^{-1}_{Y_1|C_1=1} \circ F_{Y_1|C_1=2}(x) + \frac{h_2(x)}{f_{Y_1|C_1=1}(F^{-1}_{Y_1|C_1=1}(F_{Y_1|C_1=2}(x)))} \tag{3.7}$$

Derivative for C_4 has the following form:

$$D(C_4)_{|(F_{Y_1|C_1=1},F_{Y_1|C_1=2},F_{Y_2|C_2=2})}(h_1,h_2,h_3)(q) = -\frac{h_1}{f_{Y_1|C_1=1}} \circ F^{-1}_{Y_1|C_1=1} \circ F_{Y_1|C_1=2} \circ F^{-1}_{Y_2|C_2=2}(q) + \left(\frac{h_2}{f_{Y_1|C_1=1} \circ F^{-1}_{Y_1|C_1=1} \circ F_{Y_1|C_1=2}}\right) \circ F^{-1}_{Y_2|C_2=2}(q) - \frac{f_{Y_1|C_1=2}(F^{-1}_{Y_2|C_2=2}(q))}{f_{Y_1|C_1=1}(F^{-1}_{Y_1|C_1=1}(F_{Y_1|C_1=2}(F^{-1}_{Y_2|C_2=2}(q))))} \left(\frac{h_3}{f_{Y_2|C_2=2}} \circ F^{-1}_{Y_2|C_2=2}(q)\right) \tag{3.8}$$

Proof of Proposition 5.2: For brevity I assume that we have only two clusters and omit subscript (c,k). Applying Lemma 5.2 and using the fact that Donsker theorem holds for $\hat{F}_{Y_t|C_t=c}$, we get the following by the functional delta method:

$$\sqrt{n_{11}+n_{12}}\,(\hat{\gamma}-\gamma) = -\sqrt{\frac{n_{11}+n_{12}}{n_{11}}}\left(\sqrt{n_{11}}\frac{\hat{F}_{Y_1|C_1=1}-F_{Y_1|C_1=1}}{f_{Y_1|C_1=1}}\circ F^{-1}_{Y_1|C_1=1}\circ F_{Y_1|C_1=2}\right)+$$
$$\sqrt{\frac{n_{11}+n_{12}}{n_{12}}}\left(\frac{1}{f_{Y_1|C_1=1}}\circ F^{-1}_{Y_1|C_1=1}\circ F_{Y_1|C_1=2}\right)\sqrt{n_{12}}\left(\hat{F}_{Y_1|C_1=2}-F_{Y_1|C_1=2}\right)+o_p(1) \quad (3.9)$$

This implies the first result, because $\sqrt{n_{12}}\left(\hat{F}_{Y_1|C_1=2}-F_{Y_1|C_1=2}\right)$ and $\sqrt{n_{11}}\left(\hat{F}_{Y_1|C_1=1}-F_{Y_1|C_1=1}\right)$ converge to two independent centered Gaussian processes on $[a,b]$.

In particular, the covariance function has the following representation:

$$\Phi_{\mathbb{G}_1} = \frac{1}{\lambda}\frac{F_{Y_1|C_1=1}(\gamma(\max\{x,y\}))-F_{Y_1|C_1=1}(\gamma(x))F_{Y_1|C_1=1}(\gamma(y))}{f_{Y_1|C_1=1}(\gamma(x))f_{Y_1|C_1=1}(\gamma(y))}+$$
$$\frac{1}{1-\lambda}\frac{F_{Y_1|C_1=2}(\max\{x,y\})-F_{Y_1|C_1=2}(x)F_{Y_1|C_1=2}(y)}{f_{Y_1|C_1=1}(\gamma(x))f_{Y_1|C_1=1}(\gamma(y))} \quad (3.10)$$

where $\lambda = \lim \frac{n_{11}}{n_{11}+n_{12}}$.

To prove the second claim consider the following representation:

$$\sqrt{n_{22}}\frac{1}{n_{22}}\sum_{i=1}^{n_{22}}\left(\hat{\gamma}(Y_{22i})-\mathbb{E}[\gamma(Y_2)|C_2=2]\right) = \mathbb{G}_{n_{22}}\gamma+\left(\mathbb{G}_{n_{22}}\gamma-\mathbb{G}_{n_{22}}\hat{\gamma}\right)+\sqrt{n_{22}}\left(\mathbb{P}(\hat{\gamma}-\gamma)\right) \quad (3.11)$$

The first term is asymptotically normal by the central limit theorem. The second term is $o_p(1)$ (see Chapter 19 of Van der Vaart [1998]), because the class of monotone functions is $\mathbb{P}$-Donsker and $\hat{\gamma}$ converges in sup-norm. Expansion for the third term follows from the representation defined above:

$$\mathbb{P}(\hat{\gamma}-\gamma) = -\mathbb{P}\left(\frac{\hat{F}_{Y_1|C_1=1}-F_{Y_1|C_1=1}}{f_{Y_1|C_1=1}}\circ F^{-1}_{Y_1|C_1=1}\circ F_{Y_1|C_1=2}\right)+$$
$$\mathbb{P}\left(\left(\frac{1}{f_{Y_1|C_1=1}}\circ F^{-1}_{Y_1|C_1=1}\circ F_{Y_1|C_1=2}\right)\left(\hat{F}_{Y_1|C_1=2}-F_{Y_1|C_1=2}\right)\right)+o_p\left(\frac{1}{\sqrt{n_{22}}}\right) \quad (3.12)$$

Define the following functions:

$$\begin{cases} K_{11}(z) := \int \frac{\{z<F^{-1}_{Y_1|C_1=1}\circ F_{Y_1|C_1=2}(x)\}-F_{Y_1|C_1=2}(x)}{f_{Y_1|C_1=1}(F^{-1}_{Y_1|C_1=1}\circ F_{Y_1|C_1=2}(x))}dF_{Y_2|C_2=2}(x) \\ K_{12}(z) := \int \frac{\{z<x\}-F_{Y_1|C_1=2}(x)}{f_{Y_1|C_1=1}(F^{-1}_{Y_1|C_1=1}\circ F_{Y_1|C_1=2}(x))}dF_{Y_2|C_2=2}(x) \end{cases} \quad (3.13)$$

It follows from the representation above:

$$\mathbb{P}(\hat{\gamma}) = \mathbb{P}(\gamma) - \frac{1}{n_{11}} \sum_{i=1}^{n_{11}} K_{11}(Y_{11i}) + \frac{1}{n_{12}} \sum_{i=1}^{n_{12}} K_{12}(Y_{12i}) + o_p\left(\frac{1}{\sqrt{n_{22}}}\right) \tag{3.14}$$

As a result we have the following:

$$\sqrt{n}\frac{1}{n_{22}} \sum_{i=1}^{n_{22}} \left(\hat{\gamma}(Y_{22i}) - \mathbb{E}[\gamma(Y_2)|C_2 = 2]\right) = \sqrt{\frac{n}{n_{22}}}\mathbb{G}_{n_{22}}\gamma + \sqrt{\frac{n}{n_{11}}}\mathbb{G}_{n_{11}}K_{11} + \sqrt{\frac{n}{n_{12}}}\mathbb{G}_{n_{12}}K_{12} + o_p(1) \tag{3.15}$$

It implies:

$$\begin{aligned}\sqrt{n}\left(\hat{\tau} - \tau\right) = \sqrt{\frac{n}{n_{21}}}\sqrt{\frac{1}{n_{21}}}\left(\sum_{i=1}^{n_{21}} \left(Y_{21i} - \mathbb{E}[Y_2|C_2 = 1]\right)\right) \\ - \sqrt{\frac{n}{n_{22}}}\mathbb{G}_{n_{22}}f - \sqrt{\frac{n}{n_{11}}}\mathbb{G}_{n_{11}}K_{11} - \sqrt{\frac{n}{n_{12}}}\mathbb{G}_{n_{12}}K_{12} + o_p(1)\end{aligned} \tag{3.16}$$

This implies the asymptotic normality. Asymptotic variance has the following expression:

$$\begin{aligned}V = \frac{1}{\alpha_{21}}\mathbb{V}[Y_2|C_2 = 1] + \frac{1}{\alpha_{11}}\mathbb{V}[K_{11}(Y_1)|C_1 = 1] + \\ \frac{1}{\alpha_{12}}\mathbb{V}[K_{12}(Y_1)|C_1 = 1] + \frac{1}{\alpha_{22}}\mathbb{V}[f(Y_2)|C_2 = 2]\end{aligned} \tag{3.17}$$

where $\alpha_{ct} = \lim \frac{n_{ct}}{n}$.

The last case follows directly by delta method and Lemma 5.2:

$$\begin{aligned}&\sqrt{n_{22}}\left(\hat{\tau}(q) - \tau(q)\right) = -\sqrt{\frac{n_{22}}{n_{11}}}\sqrt{n_{11}}\left(\frac{\hat{F}_{Y_1|C_1=1} - F_{Y_1|C_1=1}}{f_{Y_1|C_1=1}} \circ F^{-1}_{Y_1|C_1=1} \circ F_{Y_1|C_1=2} \circ F^{-1}_{Y_2|C_2=2}(q)\right) + \\ &\sqrt{\frac{n_{22}}{n_{12}}}\sqrt{n_{12}}\left(\frac{\hat{F}_{Y_1|C_1=2} - F_{Y_1|C_1=2}}{f_{Y_1|C_1=1} \circ F^{-1}_{Y_1|C_1=1} \circ F_{Y_1|C_1=2}}\right) \circ F^{-1}_{Y_2|C_2=2}(q) - \\ &\sqrt{n_{22}}\frac{f_{Y_1|C_1=2}(F^{-1}_{Y_2|C_2=2}(q))}{f_{Y_1|C_1=1}(F^{-1}_{Y_1|C_1=1}(F_{Y_1|C_1=2}(F^{-1}_{Y_2|C_2=2}(q))))}\left(\frac{\hat{F}_{Y_2|C_2=2} - F_{Y_2|C_2=2}}{f_{Y_2|C_2=2}} \circ F^{-1}_{Y_2|C_2=2}(q)\right) + o_p(1)\end{aligned} \tag{3.18}$$

It follows that estimator converges to a Gaussian process on $(0, 1)$.

Proof of Lemma 5.4: Following the proof of Proposition 5.2 we have the following:

$$\begin{aligned}
&\sqrt{n_{12}+n_{13}}\left(\hat{\gamma}_{1,2}(\gamma_{1,2}^{-1}(\cdot))-\hat{\gamma}_{1,3}(\gamma_{1,3}^{-1}(\cdot))\right)= \\
&\sqrt{n_{12}+n_{13}}\left(\hat{\gamma}_{1,2}(\gamma_{1,2}^{-1}(\cdot))-\gamma_{1,2}(\gamma_{1,2}^{-1}(\cdot))-(\hat{\gamma}_{1,3}(\gamma_{1,3}^{-1}(\cdot))-\gamma_{1,2}(\gamma_{1,2}^{-1}(\cdot)))\right)= \\
&\frac{\sqrt{n_{12}+n_{13}}}{\sqrt{n_{12}}}\frac{\sqrt{n_{12}}\left(\hat{F}_{Y_1|C_1=2}(\gamma_{1,2}^{-1}(\cdot))-F_{Y_1|C_1=2}(\gamma_{1,2}^{-1}(\cdot))\right)}{f_{Y_1|C_1=1}(\cdot)}- \\
&\qquad\frac{\sqrt{n_{12}+n_{13}}}{\sqrt{n_{13}}}\frac{\sqrt{n_{13}}\left(\hat{F}_{Y_1|C_1=3}(\gamma_{1,3}^{-1}(\cdot))-F_{Y_1|C_1=3}(\gamma_{1,3}^{-1}(\cdot))\right)}{f_{Y_1|C_1=1}(\cdot)}+o_p(1) \qquad (3.19)
\end{aligned}$$

From this representation we have the independence result. Also it implies the convergence to Gaussian process (by Donsker's theorem) with the following covariance function:

$$\begin{aligned}
\Phi_{\mathbb{G}_3}=&\frac{1}{\alpha}\frac{F_{Y_1|C_1=2}(\gamma_{1,2}^{-1}(\max\{x,y\}))-F_{Y_1|C_1=2}((\gamma_{1,2}^{-1}(x))F_{Y_1|C_1=2}((\gamma_{1,2}^{-1}(y))}{f_{Y_1|C_1=1}(x)f_{Y_1|C_1=1}(y)}+ \\
&\frac{1}{1-\alpha}\frac{F_{Y_1|C_1=3}(\gamma_{1,3}^{-1}(\max\{x,y\}))-F_{Y_1|C_1=3}((\gamma_{1,3}^{-1}(x))F_{Y_1|C_1=3}((\gamma_{1,3}^{-1}(y))}{f_{Y_1|C_1=1}(x)f_{Y_1|C_1=1}(y)} \qquad (3.20)
\end{aligned}$$

Proof of Lemma 5.5: We start with the following representation:

$$\begin{aligned}
&\sqrt{n_{22}+n_{23}}\left(\frac{1}{n_{22}}\sum_{i=1}^{n_{22}}\{\hat{\gamma}_{1,2}(Y_{22i})<\cdot\}-\frac{1}{n_{23}}\sum_{i=1}^{n_{22}}\{\hat{\gamma}_{1,3}(Y_{23i})<\cdot\}\right)= \\
&\sqrt{n_{22}+n_{23}}\left(\frac{1}{n_{22}}\sum_{i=1}^{n_{22}}\{\hat{\gamma}_{1,2}(Y_{22i})<\cdot\}-\mathbb{P}\{\gamma_{1,2}(Y_{22i})<\cdot\}\right)- \\
&\sqrt{n_{22}+n_{23}}\left(\frac{1}{n_{23}}\sum_{i=1}^{n_{22}}\{\hat{\gamma}_{1,3}(Y_{23i})<\cdot\}-\mathbb{P}\{\gamma_{1,3}(Y_{23i})<\cdot\}\right)= \\
&\frac{\sqrt{n_{22}+n_{23}}}{\sqrt{n_{22}}}\mathbb{G}_{n_{22}}\{\gamma_{1,2}(Y_{22i})<\cdot\}-\frac{\sqrt{n_{22}+n_{23}}}{\sqrt{n_{23}}}\mathbb{G}_{n_{23}}\{\gamma_{1,3}(Y_{23i})<\cdot\}+ \\
&\frac{\sqrt{n_{22}+n_{23}}}{\sqrt{n_{22}}}\mathbb{G}_{n_{22}}\left(\{\hat{\gamma}_{1,2}(Y_{22i})<\cdot\}-\{\gamma_{1,2}(Y_{22i})<\cdot\}\right)+ \\
&\frac{\sqrt{n_{22}+n_{23}}}{\sqrt{n_{23}}}\mathbb{G}_{n_{23}}\left(\{\hat{\gamma}_{1,3}(Y_{23i})<\cdot\}-\{\gamma_{1,3}(Y_{23i})<\cdot\}\right)+ \\
&\sqrt{n_{22}+n_{23}}\mathbb{P}\left(\{\hat{\gamma}_{1,2}(Y_{22i})<\cdot\}-\{\gamma_{1,2}(Y_{22i})<\cdot\}\right)- \\
&\qquad\sqrt{n_{22}+n_{23}}\mathbb{P}\left(\{\hat{\gamma}_{1,3}(Y_{23i})<\cdot\}-\{\gamma_{1,3}(Y_{23i})<\cdot\}\right) \qquad (3.21)
\end{aligned}$$

The first part converges to a standard $\mu_{Y_2(0,1)|C_2=1}$-Brownian bridge by Donsker's theorem. The second two terms are $o_p(1)$ by Theorem 3.2 and Lemma 3.2 in van der Vaart and Wellner [2007].

The last two terms are equivalent to the following:

$$\sqrt{n_{22}+n_{23}}\mathbb{P}\left(\{\hat{\gamma}_{1,c}(Y_{2ci})<\cdot\}-\{\gamma_{1,c}(Y_{2ci})<\cdot\}\right)= \\ \sqrt{n_{22}+n_{23}}f_{Y_2(0,1)|C_2=1}\left(\hat{\gamma}_{1,c}\circ\gamma_{1,c}^{-1}-\gamma_{1,c}\circ\gamma_{1,c}^{-1}\right)+o_p(1). \quad (3.22)$$

This follows from Lemma 4.2 in van der Vaart and Wellner [2007]. Combining this with the result of the previous lemma we get that the last two terms together converge to $\mathbb{G}_{3,1}$ with the following covariance function:

$$\Phi_{\mathbb{G}_{3,1}}(x,y)=\lambda f_{Y_2(0,1)|C_2=1}(x)f_{Y_2(0,1)|C_2=1}(y)\Phi_{\mathbb{G}_3}(x,y) \quad (3.23)$$

where $\lambda=\lim\frac{n_{22}+n_{23}}{n_{12}+n_{13}}$.

D Bootstrap algorithm for testing (preliminary)

In order to compute the quantiles for the test statistic I suggest the following procedure that consists of three steps. On the first step we estimate consistently the density $f_{Y_1|C_1=1}$ and use $\hat{\gamma}_{1,c}(Y_{2ci})$ for $c = 2, 3$ to estimate $f_{Y_2(0,1)|C_1=1}$. Let $\hat{f}_{Y_1|C_1=1}$ and $\hat{f}_{Y_2(0,1)|C_1=1}$ be the resulting estimators. Any estimators that are uniformly consistent in probability can be used. On the second step we construct four bootstrap samples for $Y_{1,c}$ and $Y_{2,c}$, where $c = 2, 3$. Let $\mathbb{G}_n^\star$ be the empirical process with respect to bootstrap samples and let $Y_{t,c}^\star$ be the bootstrapped $Y_{t,c}$. On the the third step we construct the following process:

$$Z(t) := \left(\frac{\sqrt{n_{22}+n_{23}}}{\sqrt{n_{22}}} \mathbb{G}_{n_{22}}^\star \{ Y_{22i}^\star < \hat{\gamma}_{1,2}^{-1}(t) \} + \frac{\sqrt{n_{22}+n_{23}}}{\sqrt{n_{12}}} \frac{\hat{f}_{Y_2(0,1)|C_2=1}(t)}{\hat{f}_{Y_1|C_1=1}(t)} \mathbb{G}_{n_{12}}^\star \{ Y_{12i}^\star < \hat{\gamma}_{1,2}^{-1}(t)) \} \right) - \left(\frac{\sqrt{n_{22}+n_{23}}}{\sqrt{n_{23}}} \mathbb{G}_{n_{23}}^\star \{ Y_{23i}^\star < \hat{\gamma}_{1,3}^{-1}(t) \} + \frac{\sqrt{n_{22}+n_{23}}}{\sqrt{n_{13}}} \frac{\hat{f}_{Y_2(0,1)|C_2=1}(t)}{\hat{f}_{Y_1|C_1=1}(t)} \mathbb{G}_{n_{13}}^\star \{ Y_{13i}^\star < \hat{\gamma}_{1,3}^{-1}(t)) \} \right) \tag{4.1}$$

Given bootstrap samples we can approximate the law of $Z(t)$ with arbitrary precision and compute the distribution of $\|Z\|_\infty$. Quantiles of theses distributions then can be used for testing.

To prove the formal results for this algorithm I start with the following high-level assumption.

Assumption D.0.1. *The following conditions are satisfied for estimators:*

$$\begin{cases} \|\hat{f}_{Y_1|C_1=1} - f_{Y_1|C_1=1}\|_\infty = o_p(1) \\ \|\hat{f}_{Y_2(0,1)|C_2=1} - f_{Y_2(0,1)|C_2=1}\|_\infty = o_p(1) \text{ under } H_0 \\ \|\hat{f}_{Y_2(0,1)|C_2=1} - f\|_\infty = o_p(1) \text{ for some bounded } f \text{ under } H_1 \\ \|\hat{\gamma}_{1,c}^{-1} - \gamma_{1,c}^{-1}\|_\infty = o_p(1) \end{cases} \tag{4.2}$$

Assumptions on density estimator for $\hat{f}_{Y_1|C_1=1}$ are mild, e.g., strong uniform consistency is proved for kernel estimators in Silverman et al. [1978]. The second restriction is trickier, because we use pseudo observations $\hat{\gamma}_{1,c}(Y_{2,c})$ to estimate the density. However, in light of uniform consistency of $\hat{\gamma}_{1,c}$ this fact does not matter for uniform consistency of $\hat{\gamma}_{1,c}(Y_{2,c})$, at least if we use kernel estimators: by linearity the error from using pseudo observations is bounded by $\|\hat{\gamma}_{1,c} - \gamma_{1,c}\|_\infty = o_p(1)$. The third restriction, concerning behavior under H_1 follows if $\hat{\gamma}_{1,c}$ converge under H_1 as well. Inspection of propositions concerning behavior of $\hat{\gamma}$ implies that the proofs do not depend on the fact that the model is correctly specified.

This assumption is needed to have the following equality:

$$\begin{aligned}
&\left(\frac{\sqrt{n_{22}+n_{23}}}{\sqrt{n_{22}}}\mathbb{G}^{\star}_{n_{22}}\{Y^{\star}_{22i} < \hat{\gamma}^{-1}_{1,2}(\cdot)\} + \frac{\sqrt{n_{22}+n_{23}}}{\sqrt{n_{12}}}\frac{\hat{f}_{Y_2(0,1)|C_2=1}}{\hat{f}_{Y_1|C_1=1}}\mathbb{G}^{\star}_{n_{12}}\{Y^{\star}_{12i} < \hat{\gamma}^{-1}_{1,2}(\cdot))\}\right) - \\
&\left(\frac{\sqrt{n_{22}+n_{23}}}{\sqrt{n_{23}}}\mathbb{G}^{\star}_{n_{23}}\{Y^{\star}_{23i} < \hat{\gamma}^{-1}_{1,3}(\cdot)\} + \frac{\sqrt{n_{22}+n_{23}}}{\sqrt{n_{13}}}\frac{\hat{f}_{Y_2(0,1)|C_2=1}}{\hat{f}_{Y_1|C_1=1}}\mathbb{G}^{\star}_{n_{13}}\{Y^{\star}_{13i} < \hat{\gamma}^{-1}_{1,3}(\cdot))\}\right) = \\
&\left(\frac{\sqrt{n_{22}+n_{23}}}{\sqrt{n_{22}}}\mathbb{G}^{\star}_{n_{22}}\{Y^{\star}_{22i} < \gamma^{-1}_{1,2}(\cdot)\} + \frac{\sqrt{n_{22}+n_{23}}}{\sqrt{n_{12}}}\frac{f_{Y_2(0,1)|C_2=1}}{f_{Y_1|C_1=1}}\mathbb{G}^{\star}_{n_{12}}\{Y^{\star}_{12i} < \gamma^{-1}_{1,2}(\cdot))\}\right) - \\
&\left(\frac{\sqrt{n_{22}+n_{23}}}{\sqrt{n_{23}}}\mathbb{G}^{\star}_{n_{23}}\{Y^{\star}_{23i} < \gamma^{-1}_{1,3}(\cdot)\} + \frac{\sqrt{n_{22}+n_{23}}}{\sqrt{n_{13}}}\frac{f_{Y_2(0,1)|C_2=1}}{f_{Y_1|C_1=1}}\mathbb{G}^{\star}_{n_{13}}\{Y^{\star}_{13i} < \gamma^{-1}_{1,3}(\cdot))\}\right) + o_p(1)
\end{aligned} \tag{4.3}$$

The RHS now does not depend on estimated quantities and thus converges to the same limit as the following expression:

$$\begin{aligned}
&\left(\frac{\sqrt{n_{22}+n_{23}}}{\sqrt{n_{23}}}\mathbb{G}_{n_{23}}\{Y_{23i} < \gamma^{-1}_{1,3}(\cdot)\} + \frac{\sqrt{n_{22}+n_{23}}}{\sqrt{n_{13}}}\frac{f_{Y_2(0,1)|C_2=1}}{f_{Y_1|C_1=1}}\mathbb{G}_{n_{13}}\{Y_{13i} < \gamma^{-1}_{1,3}(\cdot))\}\right) - \\
&\left(\frac{\sqrt{n_{22}+n_{23}}}{\sqrt{n_{23}}}\mathbb{G}_{n_{23}}\{Y_{23i} < \gamma^{-1}_{1,3}(\cdot)\} + \frac{\sqrt{n_{22}+n_{23}}}{\sqrt{n_{13}}}\frac{f_{Y_2(0,1)|C_2=1}}{f_{Y_1|C_1=1}}\mathbb{G}_{n_{13}}\{Y_{13i} < \gamma^{-1}_{1,3}(\cdot))\}\right)
\end{aligned} \tag{4.4}$$

And since this is the linearization of the process in Lemma 5.5 and $\|\cdot\|_\infty$ is a continuous functional we have the result by functional continuous mapping theorem.