

# Distributional treatment effect with latent rank invariance\*

Myungkou Shin<sup>†</sup>

October 24, 2025

## Abstract

Treatment effect heterogeneity is of great concern when evaluating policy impacts: e.g., “what is the proportion of people who are better off under the treatment?” However, existing analysis has been mostly limited to an average treatment effect or a quantile treatment effect, due to the fundamental limitation that we cannot simultaneously observe both treated potential outcome and untreated potential outcome for a given unit. To connect the distribution of treated potential outcome to that of untreated potential outcome, I assume that the two potential outcomes are conditionally independent, given a latent variable which is itself informed by two independent proxy variables. With a specific example of strictly increasing conditional expectation, I motivate the identifying assumption as ‘latent rank invariance.’ In implementation, I assume a finite support on the latent variable and propose an estimation strategy based on a nonnegative matrix factorization and plug-in GMM.

**Keywords:** distributional treatment effect, proximal inference, finite mixture, nonnegative matrix factorization,  $U$ -statistic, Neyman orthogonality.

**JEL classification codes:** C13

---

\*I thank Stéphane Bonhomme, Bernard Salanie, Myunghwan Seo and Martin Weidner for their valuable comments. I acknowledge the support from the European Research Council through the grant ERC-2018-CoG-819086-PANEDA. Any and all errors are my own.

<sup>†</sup>School of Social Sciences, University of Surrey. email: m.shin@surrey.ac.uk

# 1 Introduction

The fundamental limitation that we cannot simultaneously observe the two potential outcomes—treated potential outcome and untreated potential outcome—for a given unit makes the task of identifying the distribution of treatment effect particularly complicated. Thus, instead of estimating the entire distribution of treatment effect, researchers often estimate some summary measures of the treatment effect distribution, such as the average treatment effect (ATE) or the quantile treatment effect (QTE). These summary measures provide insights into the treatment effect distribution and thus help researchers with policy recommendations. However, there still remain a lot of questions that can be answered only with the *distribution* of the treatment effect: e.g., is the treatment Pareto improving?; how heterogeneous is the treatment effect at the unit level?; how many people would select into treatment when the cost of opting in is  $c$ ?

Consider a potential outcome setup with a binary treatment:

$$Y = D \cdot Y(1) + (1 - D) \cdot Y(0).$$

$Y(1)$  is the treated potential outcome,  $Y(0)$  is the untreated potential outcome, and  $D \in \{0, 1\}$  is the binary treatment variable. The questions above correspond to testing  $H_0 : F_{Y(1)-Y(0)}(0) = 0$  and estimating  $\text{Var}(Y(1) - Y(0)), 1 - F_{Y(1)-Y(0)}(c)$ . Note that these quantities,  $F_{Y(1)-Y(0)}(0), 1 - F_{Y(1)-Y(0)}(c)$  and  $\text{Var}(Y(1) - Y(0))$ , all come from the distribution of individual-level treatment effect  $Y(1) - Y(0)$ . To answer questions that relate to the distributional concerns in policy recommendation more broadly, I focus on the following two parameters of interest:

$$F_{Y(1), Y(0)}(y_1, y_0) = \Pr \{Y(1) \leq y_1, Y(0) \leq y_0\} \quad \text{for some } (y_1, y_0),$$

$$F_{Y(1)-Y(0)}(\delta) = \Pr \{Y(1) - Y(0) \leq \delta\} \quad \text{for some } \delta.$$

The first parameter is the joint distribution of the two potential outcomes and the second parameter is the marginal distribution of the treatment effect. For the rest of the paper, I

refer to these quantities as the distributional treatment effect (DTE) parameters.<sup>1</sup>

When we believe that there is no dependence between the two potential outcomes, meaning that a realized value of the treated potential outcome has no information on the individual-level heterogeneity and thus has no predictive power for the untreated potential outcome and vice versa, identification of the joint distribution of the two potential outcomes becomes trivial with a randomized treatment. Once we identify the marginal distributions of the two potential outcomes, the joint distribution becomes their product. However, this assumption is extremely restrictive. Thus, I instead impose *conditional* independence, as in Carneiro et al. [2003], and assume that a latent variable captures the individual-level heterogeneity in terms of the dependence between the two potential outcomes.

Conditional independence of potential outcomes in this paper is most plausible when the mechanism of treatment effect has regime-changing nature. As an illustration, consider the following additive model: the two potential outcomes are constructed with a individual-level latent variable  $U \in \mathcal{U} \subset \mathbb{R}$  and two regime-specific random shocks  $\varepsilon^1$  and  $\varepsilon^0$ :

$$Y(1) = \mu^1(U) + \varepsilon^1, \quad (1)$$

$$Y(0) = \mu^0(U) + \varepsilon^0. \quad (2)$$

In this framework, the treatment  $D$  operates in a way that it changes the production function for the outcome  $Y$  altogether; input  $U$  goes through a different function,  $\mu^1$  instead of  $\mu^0$ , and there are two separate random noises drawn for each production function,  $\varepsilon^0$  and  $\varepsilon^1$ . When the noises are truly random, satisfying  $\varepsilon(1) \perp\!\!\!\perp \varepsilon(0) \mid U$ , we can characterize the joint distribution of the two potential outcome as follows:

$$\Pr \{Y(1) \leq y_1, Y(0) \leq y_0\} = \mathbf{E} [\Pr \{Y(1) \leq y_1 | U\} \cdot \Pr \{Y(0) \leq y_0 | U\}].$$

Thus, the task of identifying the joint distribution of the two potential outcomes becomes that of identifying the conditional distribution of  $Y(1)$  given  $U$ , the conditional distribution

---

<sup>1</sup>Some previous works in the literature use the terminology ‘distributional effect’ to discuss parameters that are a functional of the marginal distributions of the potential outcomes; e.g., Firpo and Pinto [2016]. To avoid confusion, I will reserve the expression ‘distributional’ to only when the object involves the joint distribution of the two potential outcomes.

of  $Y(0)$  given  $U$ , and the marginal distribution of  $U$ .

To identify the conditional distribution of  $Y(d)$  given  $U$  and the marginal distribution of  $U$ , I assume that there are two additional proxy variables  $X, Z$  that are conditionally independent of each other and the potential outcomes, given  $U$ . This identification strategy is drawn from the nonclassical measurement error literature and the proximal inference literature: see Hu and Schennach [2008], Miao et al. [2018], Deaner [2023], Kedagni [2023], Nagasawa [2022] and more. In the simple example (1)-(2), the proxy variables  $X, Z$  will shift  $\mu(U)$  independently of  $(\varepsilon(1), \varepsilon(0))$ , allowing us to decompose the variation of  $Y(d)$  into the variation of  $U$  and the variation of  $\varepsilon(d)$ . Additionally, to find labels on  $U$ , I assume that the conditional distribution of  $Y(d)$  given  $U$  orders  $U$ : latent rank invariance.

Given the identification, I propose a two-step estimation strategy to estimate the DTE parameters. In the first step, I solve a nonnegative matrix factorization problem, assuming a finite support on  $U$ . Under the finite support assumption, the conditional independence assumption can be interpreted as finite mixture whose properties are well-studied in the literature. In the second step, I show that the identification of the DTE parameters reduces down to moment conditions and estimate the DTE parameters with plug-in GMM estimators, with the outcome of the nonnegative matrix factorization algorithm as nuisance parameters. Asymptotic normality of the two-step DTE estimator is established.

The framework of this paper allows us to answer important questions in terms of treatment effect heterogeneity, while being applicable to a wide range of empirical contexts where the treatment is randomly assigned, the treatment effect mechanism is regime-changing in its nature, and there exist at least two proxy variables for the individual-level heterogeneity. As an empirical illustration, I revisit Jones et al. [2019] and estimate the effect of a workplace wellness program eligibility on employees' medical spending. Using the DTE framework, I explore the treatment effect beyond the original paper's scope, estimating the entire distribution of the treatment effect. The DTE estimate demonstrates clear information gain compared to partial bounds and shows that the treatment effect has thicker left tail, while having zero treatment effect on the mean.

This paper makes a contribution to the distributional treatment effect literature: see Bedoya et al. [2018] for an overview. In this paper, the joint distribution of the potential

outcomes and thus the marginal distribution of treatment effect are point identified, in contrast to the partial identification results in the literature: Fan and Park [2010], Fan et al. [2014], Firpo and Ridder [2019], Frandsen and Lefgren [2021], Kaji and Cao [2023] and more. There exist several notable point identification results: Heckman et al. [1997], Carneiro et al. [2003]. The closest is Carneiro et al. [2003]; this paper follows their conditional independence approach and contributes by proposing a flexible estimation strategy which does not rely on parametric distributions as does Carneiro et al. [2003]. Other works that discuss estimation of DTE are Wu and Perloff [2006], Noh [2023]; unlike this paper, they both build on the point identification result without latent conditioning variable and develop a deconvolution-based estimator.

This paper also contributes to the nonclassical measurement error/proximal inference literature and the finite mixture literature: Hu and Schennach [2008], Henry et al. [2014], Miao et al. [2018], Deaner [2023], Kedagni [2023], Nagasawa [2022] and more. In terms of identification, this paper draws from Hu and Schennach [2008] and connects the two identification results for the treated subpopulation and the untreated subpopulation, building on Carneiro et al. [2003], Cunha et al. [2010]. In terms of the asymptotic theory on the estimator, this paper is in a similar setup as Hu [2008], assuming a finite support. Unlike existing estimators based on the principal component analysis, the estimation strategy based on the nonnegative matrix factorization as proposed in this paper has guarantee that the estimated conditional distributions are indeed nonnegative and sum-to-one. The  $\sqrt{n}$ -consistency is proven in this paper so that future work may build upon the nonnegative matrix factorization estimator.

The rest of the paper is organized as follows. Section 2 discusses the identification result for the joint distribution of the two potential outcomes and derives a testable implication of the framework. Section 3 explains the estimation method for the two DTE parameters and develops asymptotic theory for the estimators. Section 4 contains Monte Carlo simulation results and Section 5 applies the estimation procedure to an empirical dataset from Jones et al. [2019].

## 2 Identification

An econometrician observes a dataset  $\{Y_i, D_i, X_i, Z_i\}_{i=1}^n$  where  $Y_i, X_i, Z_i \in \mathbb{R}$  and  $D_i \in \{0, 1\}$ .  $Y_i$  is an outcome variable,  $D_i$  is a binary treatment variable and  $X_i, Z_i$  are two proxy variables for individual-level heterogeneity. The outcome  $Y_i$  is constructed with two potential outcomes.

$$Y_i = D_i \cdot Y_i(1) + (1 - D_i) \cdot Y_i(0). \quad (3)$$

In addition to  $(Y_i(1), Y_i(0), D_i, X_i, Z_i)$ , there is a latent variable  $U_i \in \mathcal{U} \subset \mathbb{R}$  that models the individual-level heterogeneity.  $U_i$  plays a key role in putting restrictions on the joint distribution of  $Y_i(1)$  and  $Y_i(0)$  and overcoming the fundamental limitation that we observe only one potential outcome for a given unit. Since  $U_i$  is latent, I assume that the two proxy variables  $X_i$  and  $Z_i$  that are informative for  $U_i$ . Examples of possible proxy variables include repeated measures of  $U_i$  when  $U_i$  has an economic interpretation and past and future outcomes in panel data. The dataset comes from random sampling:  $(Y_i(1), Y_i(0), D_i, X_i, Z_i, U_i) \stackrel{iid}{\sim} \mathcal{F}$ .

Firstly, I assume conditional random assignment on the treatment  $D_i$  and exclusion restriction on the proxy variable  $Z_i$ .

**Assumption 1.** (*assignment/exclusion restriction*)  $(Y_i(1), Y_i(0), X_i) \perp\!\!\!\perp (D_i, Z_i) \mid U_i$ .

Assumption 1 assumes that the treatment is as good as random with regard to the potential outcomes and  $X_i$  after conditioning on the latent variable  $U_i$ . In this sense, Assumption 1 is a restriction on treatment endogeneity. In addition, Assumption 1 assumes that the proxy variable  $Z_i$  does not have any additional information on the potential outcomes after conditioning on the latent variable  $U_i$ , satisfying exclusion restriction. Note that Assumption 1 does not impose any restriction on the dependence between  $Z_i$  and  $D_i$ . The proxy variable  $Z_i$  may still depend on treatment. This is a standard assumption found in the proximal inference literature, making  $X_i$  the outcome-aligned proxy and  $Z_i$  the treatment-aligned proxy.

For the rest of the paper, I focus on a randomly assigned treatment, limiting my attention to randomized controlled trials. The following condition is a sufficient condition for

Assumption 1.

*Remark 1. A sufficient condition for Assumption 1 is*

$$(Y_i(1), Y_i(0), X_i, U_i) \perp\!\!\!\perp D_i \text{ and } (Y_i(1), Y_i(0), X_i) \perp\!\!\!\perp Z_i \mid (D_i, U_i).$$

For example, a control covariate measured at baseline can be used as the proxy variable  $X_i$ . If another control variable is collected at follow-up and only depends on the pretreatment unobserved heterogeneity  $U_i$  and treatment  $D_i$ , it can be used as the proxy variable  $Z_i$ .

When  $U_i$  is observed, Assumption 1 identifies numerous treatment effect parameters such as average treatment effect (ATE), quantile treatment effect (QTE) and more. However, even when  $U_i$  is observed, we still cannot identify the distribution of treatment effect since Assumption 1 does not tell us anything about the dependence between  $Y_i(1)$  and  $Y_i(0)$ .

To impose restrictions on the joint distribution of  $Y_i(1)$  and  $Y_i(0)$  and have more identifying power, I assume that the latent variable  $U_i$  captures all of the dependence between the two potential outcomes and the proxy variable  $X_i$ .

**Assumption 2.** *(conditional independence)  $Y_i(1), Y_i(0)$  and  $X_i$  are all mutually independent given  $U_i$ .*

Two parts of Assumption 2 serve different purposes. Firstly, the conditional independence between  $(Y_i(1), Y_i(0))$  and  $X_i$  assumes that the proxy variable  $X_i$  does not give us additional information for the outcome variable given  $U_i$ . This is a standard assumption for point identification in the nonclassical measurement error literature. Additionally, Assumption 2 assumes that the two potential outcomes are independent of each other given  $U_i$ , as in Carneiro et al. [2003]. This is the key assumption that identifies the joint distribution of the two potential outcomes.

When  $U_i$  is observed, Assumptions 1-2 identify the joint distribution of the two potential outcomes and various distributional treatment effect parameters. Since  $U_i$  is not observed, identifying the conditional densities of  $Y_i(1), Y_i(0)$  given  $U_i$  and the marginal density of  $U_i$  will be the main challenge for identification.

Assumptions 1-2 play a key role in identification. Especially, the conditional indepen-

dence between  $Y_i(1)$  and  $Y_i(0)$  provides crucial identifying power in identifying the joint distribution of potential outcomes. To provide intuition on contexts where these assumptions are plausible and how they can be interpreted in practice, I present two examples based on widely adopted econometric models. The first example is repeated measures on innate ability: Carneiro et al. [2003], Cunha and Heckman [2008], Cunha et al. [2010], Attanasio et al. [2020] and more.

**Example 1.** (*repeated measures*) *The econometrician observes an outcome of interest  $Y_i$  which depends on an innate ability  $U_i$ , a binary treatment  $D_i$ , and two proxy variables  $X_i$  and  $Z_i$  which measure the innate ability  $U_i$ .  $Y_i = D_i \cdot Y_i(1) + (1 - D_i) \cdot Y_i(0)$  and*

$$\begin{aligned} Y_i(d) &= \mu^d + \alpha^d U_i + \varepsilon_i^d \quad \text{for } d = 0, 1, \\ X_i &= \mu^X + \alpha^X U_i + \varepsilon_i^X, \\ Z_i &= \mu^Z + \alpha^Z U_i + \varepsilon_i^Z \end{aligned}$$

where  $\varepsilon_i^0, \varepsilon_i^1, \varepsilon_i^X$  and  $\varepsilon_i^Z$  are mutually independent given  $U_i$ . When treatment  $D_i$  is assigned randomly, Assumptions 1-2 are satisfied.

This example is from Attanasio et al. [2020]. Attanasio et al. [2020] studies the effect of early childhood intervention on cognitive and socio-emotional skills of children aged 12 to 24 months old. The randomized intervention included home visits that provide parenting advice and assistance to parents. From Attanasio et al. [2020]’s dataset, the cognitive ability score measured at follow-up can serve as the outcome  $Y_i$  and the maternally reported measures at baseline—such as the number of words the child can say and the number of complex phrases—can serve as the proxy variables  $X_i$  and  $Z_i$ . Then, the child’s latent cognitive ability at baseline would be the latent variable  $U_i$ .<sup>2</sup> Compared to Attanasio et al. [2020], the only additional assumption that I impose here is that  $\varepsilon_i^0$  and  $\varepsilon_i^1$  are conditionally independent of each other given  $U_i$ .

The second example is the hidden Markov model: Kasahara and Shimotsu [2009], Arcidiacono and Miller [2011], Hu and Shum [2012], Hu and Sasaki [2018] and more.<sup>3</sup>

---

<sup>2</sup>Since Attanasio et al. [2020]’s analysis discusses multiple measures of cognitive and socio-emotional abilities, other variables can similarly be used as  $(Y_i, X_i, Z_i)$ , with a possibly different interpretation on  $U_i$ .

<sup>3</sup>The idea of assuming a hidden Markov model for a panel dataset and using past and future outcomes



**Example 2.** (*hidden Markov model*) The econometrician observes  $\{\{Y_{it}\}_{t=1}^3, D_i\}_{i=1}^n$  where

$$Y_{it}(d) = g_d(V_{it}, \varepsilon_{it}^d)$$

for  $t = 1, 2, 3$  and  $d = 0, 1$  and

$$Y_{it} = \begin{cases} Y_{i1}(0) & \text{if } t = 1 \\ D_i \cdot Y_{it}(1) + (1 - D_i) \cdot Y_{it}(0) & \text{if } t = 2, 3 \end{cases}.$$

$\{V_{it}\}_{t=1}^3$  is first-order Markovian given  $D_i$  and  $(\{V_{it}\}_{t=1}^3, D_i)$ ,  $\varepsilon_{i1}^0, \varepsilon_{i2}^0, \varepsilon_{i2}^1, \varepsilon_{i3}^0$  and  $\varepsilon_{i3}^1$  are mutually independent. When treatment  $D_i$  is randomly assigned at time  $t = 2$ , i.e.  $\{V_{it}\}_{t=1}^3 \perp\!\!\!\perp D_i$ , Assumptions 1-2 are satisfied with  $Y_i = Y_{i2}$ ,  $X_i = Y_{i1}$ ,  $Z_i = Y_{i3}$  and  $U_i = V_{i2}$ .

This example extends the hidden Markov model to a potential outcome setup by modeling the two potential outcomes separately and assuming that the latent state process  $\{V_{it}\}_{t=1}^3$  is first-order Markovian and shared across the two outcome generating processes. To provide a context, let us connect this example to Jones et al. [2019] from Section 5. In Jones et al. [2019], the authors randomly assigned workplace wellness program eligibility to university employees and estimated its effect on medical spendings. The program included in-person classes on physical fitness, healthy workplace habits, etc and online health risk assessments. The medical spending information was collected before, during, and after the treatment. Thus, the pretreatment and the follow-up medical spendings can serve as the proxy variables  $X_i$  and  $Z_i$  and the treatment-period medical spending as the outcome variable  $Y_i$ . In this context,  $\{V_{it}\}_{t=1}^3$  can be thought of as a process of underlying health status.

The common component across the two examples is that the treatment  $D_i$  affects the outcome  $Y_i$  in a regime-changing manner; there are two separate production functions— $\mu^1 + \alpha^1 U_i + \varepsilon_i^1$  and  $\mu^0 + \alpha^0 U_i + \varepsilon_i^0$  in the first example, and  $g_0(V_{i2}, \varepsilon_{i2}^0)$  and  $g_1(V_{i2}, \varepsilon_{i2}^1)$  in the second—and the treatment  $D_i$  decides which production function is applied to generate  $Y_i$ . This point is briefly addressed in Attanasio et al. [2020] where the authors discuss two

---

as proxy variables draws from the proximal inference literature: e.g., [Deaner, 2023]. The key element of the setup is that we observe pretreatment outcome  $Y_{i1}$  to connect the treated subpopulation and the untreated subpopulation, given a random treatment.

possible mechanisms of treatment effect: a change in production function itself and a change in the amount of inputs. However, they abstract away from a comprehensive counterfactual analysis by taking no stance on how the error terms in the two production functions would be connected if the mechanism of the treatment effect is the former. In this paper, I assume that the two error terms are indeed pure random noises, so that a random noise in one regime is independent of a random noise in another regime, conditional on  $U_i$ .

Thus, the framework of this paper is best suited for empirical contexts where a randomized treatment affects an outcome by placing treated units in an alternative regime of outcome generating process. If the parenting guidance systemically changes children’s cognitive ability development process, Attanasio et al. [2020] would fit this framework well. Similarly, if the takeup was high and the healthy lifestyle information and the health risk assessment systemically changes participants’ health behaviors, Jones et al. [2019] would also fit the framework well. Other examples include a job training program where treated participants are assigned a new job, such as the National Supported Work Demonstration. Given a new job, the regime of how innate aptitude and skill lead to wage income will be shifted. Another example is educational intervention based on teaching methodology: e.g., Banerjee et al. [2007], Muralidharan et al. [2019] and more. By being taught with a different teaching methodology, the production function of students’ outcome shifts to a new regime.

The key assumption that the regime-specific error terms  $\varepsilon_i^0$  and  $\varepsilon_i^1$  are conditionally independent is most plausible when they represent purely random noises. This requires that  $U_i$  fully account for all of the individual-level heterogeneity and eliminate any rationale for dependence between random noises in two different regimes, making the assumption that the latent variable  $U_i$  is a scalar somewhat restrictive. With more information from observable data, this can be relaxed. Firstly, since the entire argument in this section can be conditional on control covariates, this problem is partially mitigated when given some additional observable control covariates  $C_i$  such that

$$(Y_i(1), Y_i(0), X_i, C_i, U_i) \perp\!\!\!\perp D_i \text{ and } (Y_i(1), Y_i(0), X_i) \perp\!\!\!\perp Z_i \mid (D_i, C_i, U_i)$$

$$Y_i(1), Y_i(0) \text{ and } X_i \text{ are all mutually independent given } (C_i, U_i).$$

In this setup, the scalar  $U_i$  only needs to explain remaining heterogeneity among individuals with the same level of  $C_i$ . Secondly, we can model the latent variable  $U_i$  to be multidimensional, as long as the two proxy variables  $X_i, Z_i$  are of the same dimension. For more discussion on multidimensional  $U_i$ , see the appendix Section A.

## 2.1 Identification of the joint distribution of $Y_i(1)$ and $Y_i(0)$

The remainder of this section outlines the identification argument. A key step in identification is the diagonalization of conditional densities, drawn from and illustrated in detail in Hu [2008], Hu and Schennach [2008]. I reiterate the diagonalization step here since it directly motivates the DTE estimation strategy in Section 3. For illustration purposes only, let  $Y_i, X_i, Z_i, U_i$  be discrete:  $Y_i \in \{y^1, \dots, y^{M_Y}\}, X_i \in \{x^1, \dots, x^{M_X}\}, Z_i \in \{z^1, \dots, z^{M_Z}\}$  and  $U_i \in \{u^1, \dots, u^K\}$ . With  $M = M_Y \cdot M_X$ , we can construct a  $M \times M_Z$  matrix  $\mathbf{H}_d$  that collects conditional probabilities of  $(Y_i, X_i)$  given  $(D_i = d, Z_i)$ : for  $d = 0, 1$ ,

$$\mathbf{H}_d = \begin{pmatrix} \Pr\{(Y_i, X_i) = (y^1, x^1) \mid (D_i, Z_i) = (d, z^1)\} & \cdots & \Pr\{(Y_i, X_i) = (y^1, x^1) \mid (D_i, Z_i) = (d, z^{M_Z})\} \\ \vdots & \ddots & \vdots \\ \Pr\{(Y_i, X_i) = (y^{M_Y}, x^{M_X}) \mid (D_i, Z_i) = (d, z^1)\} & \cdots & \Pr\{(Y_i, X_i) = (y^{M_Y}, x^{M_X}) \mid (D_i, Z_i) = (d, z^{M_Z})\} \end{pmatrix}.$$

From Assumption 1,  $\mathbf{H}_d$  decomposes into two matrices: for each  $d = 0, 1$ ,

$$\mathbf{H}_d = \Gamma_d \cdot \Lambda_d \tag{4}$$

where

$$\begin{aligned} \Gamma_d &= \begin{pmatrix} \Pr\{(Y_i(d), X_i) = (y^1, x^1) \mid U_i = u^1\} & \cdots & \Pr\{(Y_i(d), X_i) = (y^1, x^1) \mid U_i = u^K\} \\ \vdots & \ddots & \vdots \\ \Pr\{(Y_i(d), X_i) = (y^{M_Y}, x^{M_X}) \mid U_i = u^1\} & \cdots & \Pr\{(Y_i(d), X_i) = (y^{M_Y}, x^{M_X}) \mid U_i = u^K\} \end{pmatrix}, \\ \Lambda_d &= \begin{pmatrix} \Pr\{U_i = u^1 \mid (D_i, Z_i) = (d, z^1)\} & \cdots & \Pr\{U_i = u^1 \mid (D_i, Z_i) = (d, z^{M_Z})\} \\ \vdots & \ddots & \vdots \\ \Pr\{U_i = u^K \mid (D_i, Z_i) = (d, z^1)\} & \cdots & \Pr\{U_i = u^K \mid (D_i, Z_i) = (d, z^{M_Z})\} \end{pmatrix}. \end{aligned} \tag{5}$$

The discreteness of  $Y_i, X_i, Z_i$  is nonbinding; we can use partitioning on  $\mathbb{R}$  when they are continuous.<sup>4</sup>

The equation  $\mathbf{H}_d = \Gamma_d \cdot \Lambda_d$  shows us that the conditional density of  $(Y_i, X_i)$  given  $(D_i, Z_i)$  admits a mixture model. For each subpopulation  $\{i : (D_i, Z_i) = (d, z)\}$ , there is a column in the matrix  $\Lambda_d$  which denotes the subpopulation-specific distribution of  $U_i$ . Then, the density of  $(Y_i, X_i)$  in that subpopulation admits a mixture model with the aforementioned columns of  $\Lambda_d$  as mixture weights and the conditional density of  $(Y_i(d), X_i)$  given  $U_i$  as mixture component densities. The equation  $\mathbf{H}_d = \Gamma_d \cdot \Lambda_d$  aggregates the finite mixture formulations across the subpopulations.

Under Assumption 2,  $\Gamma_0$  and  $\Gamma_1$  can be further decomposed. Let

$$\Gamma_X = \left( \Pr \{X_i = x^m | U_i = u^k\} \right)_{m,k} \quad \text{and} \quad \Gamma_{Y(d)} = \left( \Pr \{Y_i(d) = y^m | U_i = u^k\} \right)_{m,k}.$$

Also, let  $\Gamma_{d,k}$  denote the  $k$ -th column of  $\Gamma_d$  and similarly for  $\Gamma_X$  and  $\Gamma_{Y(d)}$ . Then, with  $\otimes$  denoting the Kronecker product,  $\Gamma_{d,k} = \Gamma_{X,k} \otimes \Gamma_{Y(d),k}$  for  $d = 0, 1$  and  $k = 1, \dots, K$ .

Consider a submatrix of  $\mathbf{H}_d$  that stacks the rows of  $\mathbf{H}_d$  that correspond to a specific value of  $y$ :  $\mathbf{H}_d(y)$ . Then, from the two decompositions above,

$$\mathbf{H}_d(y) = \Gamma_X \cdot \Delta(y) \cdot \Lambda_d$$

where  $\Delta(y)$  is a diagonal matrix with the row of  $\Gamma_{Y(d)}$  corresponding to  $y$  as the diagonals. Hu [2008], Hu and Schennach [2008] show that  $\Gamma_d$  is identified by collecting  $\Gamma_X$  and  $\Delta(y)$

---

<sup>4</sup>Consider partitions on  $\mathbb{R}$  such that

$$\{\mathcal{Y}^m = (y^{m-1}, y^m)\}_{m=1}^{M_Y}, \quad \{\mathcal{X}^m = (x^{m-1}, x^m)\}_{m=1}^{M_X}, \quad \{\mathcal{Z}^m = (z^{m-1}, z^m)\}_{m=1}^{M_Z}$$

where  $y^0 = x^0 = z^0 = -\infty$  and  $y^{M_Y} = x^{M_X} = z^{M_Z} = \infty$ . Let  $\mathcal{W}^1 = \mathcal{Y}^1 \times \mathcal{X}^1, \mathcal{W}^2 = \mathcal{Y}^2 \times \mathcal{X}^1, \dots, \mathcal{W}^M = \mathcal{Y}^{M_Y} \times \mathcal{X}^{M_X}$ .  $\{\mathcal{W}^m\}_{m=1}^M$  is a partition on  $\mathbb{R}^2$ . Then,  $\mathbf{H}_d$  becomes

$$\mathbf{H}_d = \begin{pmatrix} \Pr \{(Y_i, X_i) \in \mathcal{W}^1 | D_i = d, Z_i \in \mathcal{Z}^1\} & \cdots & \Pr \{(Y_i, X_i) \in \mathcal{W}^1 | D_i = d, Z_i \in \mathcal{Z}^{M_Z}\} \\ \vdots & \ddots & \vdots \\ \Pr \{(Y_i, X_i) \in \mathcal{W}^M | D_i = d, Z_i \in \mathcal{Z}^1\} & \cdots & \Pr \{(Y_i, X_i) \in \mathcal{W}^M | D_i = d, Z_i \in \mathcal{Z}^{M_Z}\} \end{pmatrix}$$

for each  $d = 0, 1$ .  $\Gamma_d$  and  $\Lambda_d$  are similarly constructed with partitioned  $Y_i, X_i$  and  $Z_i$ .

from the eigenvalue decompositions of

$$\mathbf{H}_d(y) \left( \sum_{y'} \mathbf{H}_d(y') \right)^{-1} = \Gamma_X \cdot \Delta(y) \cdot (\Gamma_X)^{-1}$$

across  $y$ , when  $\Gamma_X$  and  $\Lambda_d$  are full rank and no two columns of  $\Gamma_{Y(d)}$  are identical.<sup>5</sup> Then,  $\Lambda_d$  is identified from the full rank of  $\Gamma_d$ , which itself follows from the full rank of  $\Gamma_X$ .

Recall that from Assumption 2, the joint distribution of  $Y_i(1)$  and  $Y_i(0)$  is identified once we identify the conditional distribution of  $Y_i(1)$  given  $U_i$ , the conditional distribution of  $Y_i(0)$  given  $U_i$ , and the marginal distribution of  $U_i$ . The first two distributions correspond to  $\Gamma_1$  and  $\Gamma_0$  in the discretization. The last distribution is a function of  $\Lambda_1, \Lambda_0$  and the distribution of  $(D_i, Z_i)$ , which is observed. Thus, the result of Hu and Schennach [2008] can be applied twice, firstly to  $\mathbf{H}_0$  and secondly to  $\mathbf{H}_1$ , to identify the joint distribution of  $Y_i(1)$  and  $Y_i(0)$ . The key condition which connects the decomposition of  $\mathbf{H}_0$  to that of  $\mathbf{H}_1$  is the part of Assumption 1 where I assume that the conditional distribution of  $X_i$  given  $(D_i, U_i)$  does not depend on  $D_i$ ;  $\Gamma_X$  appears in both of the decompositions and thus we can connect the labelings on the latent variable  $U_i$  across the two subpopulations using  $\Gamma_X$ .

Assumption 3 assumes a discrete  $U_i$ , making the decomposition in (4) exact, and formally states the full rank condition and the no repeated eigenvalue condition.

**Assumption 3.**

- a. (finitely discrete  $U_i$ )*  $\mathcal{U} = \{u^1, \dots, u^K\}$ .
- b. (full rank)*  $\Lambda_0, \Lambda_1$  and  $\Gamma_X$  have rank  $K$ .
- c. (no repeated eigenvalue)* For any  $k \neq k'$ , there exist some  $y, y' \in \{y^1, \dots, y^{M_Y}\}$  such that

$$\begin{aligned} \Pr \{Y_i(0) = y | U_i = u^k\} &\neq \Pr \{Y_i(0) = y | U_i = u^{k'}\}, \\ \Pr \{Y_i(1) = y' | U_i = u^k\} &\neq \Pr \{Y_i(1) = y' | U_i = u^{k'}\}. \end{aligned}$$

---

<sup>5</sup>When  $\Gamma_X$  or  $\Lambda_d$  is not a square matrix, focusing on  $K$  linearly independent columns of  $\Lambda_d$  and using a pseudoinverse of  $\Gamma_X$  derive the same result.

Assumption 3.b implicitly assumes that  $M_X, M_Z \geq K$ . The restriction that  $M_X, M_Z \geq K$  is sensible since I use the proxy variables to capture the variation in the latent variable  $U_i$ . The support for the two proxy variables has to be at least as rich as that of the latent variable. Assumption 3.c assumes that the eigenvalue decomposition does not have repeated eigenvalues.

Assumption 4 reiterates Assumption 3 for a setup where  $U_i$  is continuous. Let  $f_{Y(d)|U}$  denote the conditional density of  $Y_i(d)$  given  $U_i$ ,  $f_{X|U}$  denote the conditional density of  $X_i$  given  $U_i$ , and  $f_{U|D=d,Z}$  denote the conditional density of  $U_i$  given  $D_i = d$  and  $Z_i$ , for  $d = 0, 1$ . Define integral operators  $L_{X|U}$  and  $L_{Z|D=d,U}$  that map a function in  $\mathcal{L}^1(\mathbb{R})$  to a function in  $\mathcal{L}^1(\mathbb{R})$ : for  $d = 0, 1$ ,

$$\begin{aligned} [L_{X|U}g](x) &= \int_{\mathbb{R}} f_{X|U}(x|u)g(u)du, \\ [L_{Z|D=d,U}g](z) &= \int_{\mathbb{R}} f_{Z|D=d,U}(z|u)g(u)du. \end{aligned}$$

**Assumption 4.** *Assume*

- a.** *(continuous  $U_i$ )*  $\mathcal{U} = [0, 1]$ .
- b.** *(bounded density)* The conditional densities  $f_{Y(1)|U}, f_{Y(0)|U}, f_{X|U}, f_{U|D=1,Z}$  and  $f_{U|D=0,Z}$  and the marginal densities  $f_U, f_{Z|D=1}$  and  $f_{Z|D=0}$  are bounded.
- c.** *(completeness)* The integral operators  $L_{X|U}, L_{Z|D=1,U}$  and  $L_{Z|D=0,U}$  are injective on  $\mathcal{L}^1(\mathbb{R})$ .
- d.** *(no repeated eigenvalue)* For any  $u \neq u'$ ,

$$\Pr \{f_{Y(d)|U}(Y_i|u) \neq f_{Y(d)|U}(Y_i|u') | D_i = d\} > 0$$

for each  $d = 0, 1$ .

Assumption 4.c corresponds to Assumption 3.b and Assumption 4.d to Assumption 3.c.

When  $U_i$  is continuous, we need an additional assumption for the identification. This is because we need an ordering on the infinite collection  $\{f_{X|U}(\cdot|u)\}_u$  to connect  $u$  to  $f_{X|U}(\cdot|u)$  when  $U_i$  is continuous.

**Assumption 5.** (*latent rank*) *There exists a functional  $M$  defined on  $\mathcal{L}^1(\mathbb{R})$  such that either*

$$h(u) = Mf_{Y(1)|U}(\cdot|u) \quad \text{or} \quad h(u) = Mf_{Y(0)|U}(\cdot|u)$$

*defined on  $\mathcal{U}$  is strictly increasing and continuously differentiable.*

The functional  $M$  provides us an ordering on the infinite collection  $\{f_{X|U}(\cdot|u)\}_u$ , when applied to  $\{f_{Y(1)|U}(\cdot|u), f_{Y(0)|U}(\cdot|u)\}_u$ . A simple example where Assumption 5 fails is when  $\mathcal{U} = [-1, 1]$  and  $Y_i(d)|U_i = u \sim \mathcal{N}(u^2 + d, \sigma^2)$ . Neither  $f_{Y(1)|U}$  nor  $f_{Y(0)|U}$  helps us find an ordering between  $f_{X|U}(\cdot|u)$  and  $f_{X|U}(\cdot| -u)$ .

Under Assumption 5, the latent variable  $U_i$  is interpreted as a ‘latent rank.’ Suppose that  $\mathbf{E}[\varepsilon_i^1|U_i] = 0$  in (1) and that Assumption 5 holds with  $\mathbf{E}[Y_i(1)|U_i = u] = \mu^1(u)$ . Then,  $U_i$  represents the rank of the systemic part of the treated potential outcome generating process,  $\mu^1(U_i)$ , which is latent. If we extend Assumption 5 so that both  $Mf_{Y(1)|U}(\cdot|u)$  and  $Mf_{Y(0)|U}(\cdot|u)$  are strictly increasing in  $u$ , we get ‘latent rank invariance.’ Under the latent rank invariance, the relative positions of unit  $i$  in terms of the two systemic parts of the potential outcome generating processes are the same. However, given a realized outcome, the counterfactual potential outcome may still vary, depending on  $\varepsilon_i^0 | \mu^1(U_i) + \varepsilon_i^1$  or vice versa. In this sense, Assumption 5 is a direct relaxation of the rank invariance condition from the quantile treatment effect/IV literature ([Chernozhukov and Hansen, 2005, 2006, Athey and Imbens, 2006, Vuong and Xu, 2017, Callaway and Li, 2019, Han and Xu, 2023]) where the relative position of a realized outcome fixes the counterfactual potential outcome deterministically.

Theorem 1 formally states identification.

**Theorem 1.** *Either Assumptions 1-3 or Assumptions 1-2, 4-5 hold. Then, the joint density of  $(Y_i(1), Y_i(0), D_i, X_i, Z_i)$  is identified.*

*Proof.* See Appendix. □

Another interpretation of Theorem 1 under Assumption 3 is that it is a point identification adaptation of the set identification result from Henry et al. [2014]; the additional identifying power comes from the conditional independence between  $Y_i(d)$  and  $X_i$  given  $U_i$ . It directly

follows Theorem 1 that any functional of the joint distribution of  $Y_i(1)$  and  $Y_i(0)$  is identified: e.g.,  $\text{Var}(Y_i(1) - Y_i(0))$ ,  $\Pr\{Y_i(1) \geq Y_i(0)\}$ ,  $\Pr\{Y_i(1) \geq Y_i(0)|Y_i(0)\}$  and etc.

## 2.2 Testable implication

Under the latent rank invariance condition that both  $Mf_{Y(1)|U}(\cdot|u)$  and  $Mf_{Y(0)|U}(\cdot|u)$  are strictly increasing in  $u$ , we have a testable implication of Assumptions 1-2 and 4-5, from over-identification. With the latent rank invariance, we can find a labeling on  $\{f_{X|U}(\cdot|u)\}_u$  within each subpopulation; the conditional densities  $(f_{Y(1)|U}, f_{X|U}, f_{U|D=1,Z})$  are identified from the treated subpopulation and the conditional densities  $(f_{Y(0)|U}, f_{X|U}, f_{U|D=0,Z})$  are identified from the untreated subpopulation, separately. Let  $f_{X|D=1,U}$  denote the conditional density of  $X_i$  given  $U_i$ , identified from the treated subpopulation and likewise for  $f_{X|D=0,U}$ . Then, Assumption 1 imposes that

$$\min_{\tilde{g}: \text{monotone}} \mathbf{E} \left[ \int_{\mathbb{R}} (f_{X|D=1,U}(x|U_i) - f_{X|D=0,U}(x|\tilde{g}(U_i)))^2 dx \middle| D_i = 1 \right] = 0 \quad (6)$$

since  $X_i \perp\!\!\!\perp D_i \mid U_i$ . In (6), a monotone function  $\tilde{g}$  is used to connect the two identification results, now that  $f_{X|U}$  is not used to connect the two identification results.<sup>6</sup> A test that uses (6) as a null can be used as a falsification test on the framework proposed in this paper.

What does a test on the null (6) exactly test? The mixture model on the conditional density  $f_{Y,X|D=d,Z}$  assumes that conditioning on  $U_i$ , the potential outcome  $Y_i(d)$  and the proxy variable  $X_i$  are independent of each other. Recall that in Example 2, the proxy variable  $X_i$  is the past outcome. Thus, in the panel context, we can understand the falsification test as testing whether we can find a latent variable  $U_i$  conditioning on which the outcomes are *intertemporally* independent while satisfying  $X_i \perp\!\!\!\perp D_i \mid U_i$ . Note that Assumption 2 also includes that the potential outcomes are independent *across the treatment status*. While the conditional independence across the treatment status remains untestable due

---

<sup>6</sup>In the case of discrete  $U_i$ , Assumption 5 was not used in the identification. Thus, without invoking latent rank invariance, we have a testable implication:

$$\sum_{k=1}^K \min_{k'} \sum_{j=1}^{M_X} \left( \Pr\{X_i = x^j \mid (D_i, U_i) = (1, u^k)\} - \Pr\{X_i = x^j \mid (D_i, U_i) = (0, u^{k'})\} \right)^2 = 0.$$



to the limitation that we only observe either a treated potential outcome or a untreated potential outcome for a given unit, the falsification test in Example 2 tests if the outcomes are intertemporally independent, conditioning on some latent variable.

When  $D_i$  is assigned randomly as in Remark 1, (6) simplifies to

$$\mathbf{E} \left[ \int_{\mathbb{R}} (f_{X|D=1,U}(x|U_i) - f_{X|D=0,U}(x|U_i))^2 dx \right] = 0$$

since the distribution of  $U_i$  is identical across the two subpopulations. In addition, when  $D_i$  is assignly randomly, we can directly test the equivalence between the distribution of  $U_i$  in the treated subpopulation and that in the untreated subpopulation:

$$\mathbf{E} \left[ (f_{U|D=1}(U_i) - f_{U|D=0}(U_i))^2 \right] = 0.$$

Formal constructions of the two falsification tests are provided in appendix Section B.

### 3 Implementation

In this section, I propose an estimation strategy under the finite support assumption on  $\mathcal{U}$ . The focus on the case of discrete  $U_i$  has several reasons. Firstly, a discretization is often used in econometric models with latent heterogeneity as an approximation to a continuous latent heterogeneity space: see Bonhomme et al. [2022] for more. Secondly, with parametrization, the estimation of infinite-dimensional objects such as conditional densities  $f_{U|D=0,Z}$  and  $f_{U|D=1,Z}$  becomes an estimation of finite-dimensional objects  $\Lambda_0$  and  $\Lambda_1$ , giving us  $n^{-\frac{1}{2}}$  rate. Lastly, the linearity induced from discretization leads to a simple GMM estimation, reducing the computational burden substantially. Alternatively, we can construct a sieve maximum likelihood estimator, as suggested in the nonclassical measurement error literature, and let  $U_i$  be continuous under Assumptions 4-5. The specifics of the sieve MLE are discussed in the Online Appendix.

All of the discussions in this section assume that  $K$ , the number of points in the support of  $U_i$ , is known. In practice, we often do not have *a priori* choice of  $K$ . Thus, in Appendix C, I discuss how to apply the existing econometric methods such as the eigenvalue ratio

estimator from Ahn and Horenstein [2013] and the rank test from Kleibergen and Paap [2006] for guidance on the choice of  $K$ .

The estimation procedure of this paper is two-step. Firstly, I solve a nonnegative matrix factorization (NMF) problem, to estimate  $\Lambda_0$  and  $\Lambda_1$ . Secondly, I estimate the joint distribution of  $Y_i(1)$  and  $Y_i(0)$  and the marginal distribution of  $Y_i(1) - Y_i(0)$  with a plug-in GMM estimation where  $\Lambda_0$  and  $\Lambda_1$  are nuisance parameters:

$$F_{Y(0),Y(1)}(y, y') = \Pr \{Y_i(0) \leq y, Y_i(1) \leq y'\},$$

$$F_{Y(1)-Y(0)}(\delta) = \Pr \{Y_i(1) - Y_i(0) \leq \delta\}.$$

### 3.1 Nonnegative matrix factorization

To estimate the mixture weight matrices  $\Lambda_0$  and  $\Lambda_1$  from (5), I formulate a nonnegative matrix factorization problem based on (4). Since  $\mathbf{H}_d$  is not directly observed, I estimate  $\mathbf{H}_d$  with its sample analogue  $\mathbb{H}_d$ : for  $d = 0, 1$ , let

$$\mathbb{H}_d = \begin{pmatrix} \frac{\sum_{i=1}^n \mathbf{1}\{(Y_i, D_i, X_i, Z_i) = (y^1, d, x^1, z^1)\}}{\sum_{i=1}^n \mathbf{1}\{(D_i, Z_i) = (d, z^1)\}} & \cdots & \frac{\sum_{i=1}^n \mathbf{1}\{(Y_i, D_i, X_i, Z_i) = (y^1, d, x^1, z^K)\}}{\sum_{i=1}^n \mathbf{1}\{(D_i, Z_i) = (d, z^K)\}} \\ \vdots & \ddots & \vdots \\ \frac{\sum_{i=1}^n \mathbf{1}\{(Y_i, D_i, X_i, Z_i) = (y^{M_Y}, d, x^{M_X}, z^1)\}}{\sum_{i=1}^n \mathbf{1}\{(D_i, Z_i) = (d, z^1)\}} & \cdots & \frac{\sum_{i=1}^n \mathbf{1}\{(Y_i, D_i, X_i, Z_i) = (y^{M_Y}, d, x^{M_X}, z^K)\}}{\sum_{i=1}^n \mathbf{1}\{(D_i, Z_i) = (d, z^K)\}} \end{pmatrix}.$$

Each column of  $\mathbb{H}_d$  is an empirical conditional distribution function of  $(Y_i, X_i)$  given  $(D_i = d, Z_i)$ . In constructing  $\mathbb{H}_d$ , I impose that  $Z_i$  has  $K$  points in its support. Whenever  $M_Z \geq K$ , this is nonbinding since I can simply use partitioning on  $\mathbb{R}$ . Similarly, when  $Y_i$  or  $X_i$  is continuous, I use partitioning on  $\mathbb{R}$ .

Given the estimates of  $\mathbf{H}_0$  and  $\mathbf{H}_1$ , I estimate  $\Lambda_0$  and  $\Lambda_1$  by solving a nonnegative matrix factorization problem: with  $\iota_x$  denoting a  $x$ -dimensional column vector of ones,

$$\min_{\Lambda_0, \Lambda_1, \Gamma_0, \Gamma_1} \|\mathbb{H}_0 - \Gamma_0 \Lambda_0\|_F^2 + \|\mathbb{H}_1 - \Gamma_1 \Lambda_1\|_F^2 \quad (7)$$

subject to linear constraints

$$\begin{aligned}\Lambda_0 &\in \mathbb{R}_+^{K \times K}, \quad \Lambda_1 \in \mathbb{R}_+^{K \times K}, \quad \Gamma_0 \in \mathbb{R}_+^{M \times K}, \quad \Gamma_1 \in \mathbb{R}_+^{M \times K}, \\ \iota_K^\top \Lambda_0 &= \iota_K^\top, \quad \iota_K^\top \Lambda_1 = \iota_K^\top, \quad \iota_M^\top \Gamma_0 = \iota_K^\top, \quad \iota_M^\top \Gamma_1 = \iota_K^\top,\end{aligned}$$

linear constraints on  $(\Gamma_0, \Gamma_1)$  that for every  $(x, k)$

$$\left( \sum_{m=1}^{M_Y} \Pr \{ (Y_i(0), X_i) = (y^m, x) \mid U_i = u^k \} \right) = \left( \sum_{m=1}^{M_Y} \Pr \{ (Y_i(1), X_i) = (y^m, x) \mid U_i = u^k \} \right) \quad (8)$$

and quadratic constraints on  $(\Gamma_0, \Gamma_1)$  that for every  $(y, d, x, k)$

$$\Pr \{ (Y_i(d), X_i) = (y, x) \mid U_i = u^k \} \quad (9)$$

$$= \left( \sum_{m=1}^{M_X} \Pr \{ (Y_i(d), X_i) = (y, x^m) \mid U_i = u^k \} \right) \cdot \left( \sum_{m=1}^{M_Y} \Pr \{ (Y_i(d), X_i) = (y^m, x) \mid U_i = u^k \} \right) \quad (10)$$

The objective (7) comes from the decomposition (4). The linear constraints impose that the columns of  $\Gamma_0, \Gamma_1, \Lambda_0$  and  $\Lambda_1$  are well-defined distributions and that  $X_i \perp\!\!\!\perp D_i \mid U_i$  (Assumption 1). The quadratic constraints impose that  $Y_i(d) \perp\!\!\!\perp X_i \mid U_i$  (Assumption 2). Theorem 1 says that the solution to the optimization problem is unique up to some permutation on the columns of  $\Gamma_0, \Gamma_1$  and the rows of  $\Lambda_0, \Lambda_1$ , when  $\mathbb{H}_0$  and  $\mathbb{H}_1$  are sufficiently close to  $\mathbf{H}_0$  and  $\mathbf{H}_1$ .

Note that the objective function in (7) is quadratic when we fix either  $(\Lambda_0, \Lambda_1)$  or  $(\Gamma_0, \Gamma_1)$ . Moreover, recall that  $\Gamma_0$  and  $\Gamma_1$  can be further decomposed into three matrices  $\Gamma_X, \Gamma_{Y(0)}, \Gamma_{Y(1)}$ . Let  $\Gamma(\cdot, \cdot)$  denote how  $\Gamma_X$  and  $\Gamma_{Y(d)}$  recover  $\Gamma_d$ , using the column-wise Kronecker products:  $\Gamma_d = \Gamma(\Gamma_X, \Gamma_{Y(d)})$ . The quadratic constraints are trivially imposed by optimizing over  $\Gamma_X, \Gamma_{Y(0)}$  and  $\Gamma_{Y(1)}$ . Using these, I propose an iterative algorithm to solve the minimization problem.

1. Initialize  $\Gamma_0^{(0)}, \Gamma_1^{(0)}$ .

2. (*Update*  $\Lambda$ ) Given  $(\Gamma_0^{(s)}, \Gamma_1^{(s)})$ , solve the following quadratic program:

$$(\Lambda_0^{(s+1)}, \Lambda_1^{(s+1)}) = \arg \min_{\Lambda_0, \Lambda_1} \left\| \mathbb{H}_0 - \Gamma_0^{(s)} \Lambda_0 \right\|_F^2 + \left\| \mathbb{H}_1 - \Gamma_1^{(s)} \Lambda_1 \right\|_F^2$$

subject to  $\Lambda_0 \in \mathbb{R}_+^{K \times K}$ ,  $\Lambda_1 \in \mathbb{R}_+^{K \times K}$ ,  $\iota_K^\top \Lambda_0 = \iota_K^\top$  and  $\iota_K^\top \Lambda_1 = \iota_K^\top$ .

3. (*Update*  $\Gamma_X$ ) Given  $(\Lambda_0^{(s+1)}, \Lambda_1^{(s+1)}, \Gamma_{Y(0)}^{(s)}, \Gamma_{Y(1)}^{(s)})$ , solve the following quadratic program:

$$(\Gamma_X^{(s+1)}) = \arg \min_{\Gamma_X} \left\| \mathbb{H}_0 - \Gamma \left( \Gamma_X, \Gamma_{Y(0)}^{(s)} \right) \Lambda_0^{(s+1)} \right\|_F^2 + \left\| \mathbb{H}_1 - \Gamma \left( \Gamma_X, \Gamma_{Y(1)}^{(s)} \right) \Lambda_1^{(s+1)} \right\|_F^2$$

subject to  $\Gamma_X \in \mathbb{R}_+^{M_X \times K}$ ,  $\iota_{M_X}^\top \Gamma_X = \iota_K^\top$ .

4. (*Update*  $\Gamma_Y$ ) Given  $(\Lambda_0^{(s+1)}, \Lambda_1^{(s+1)}, \Gamma_X^{(s+1)})$ , solve the following quadratic program:

$$\begin{aligned} & (\Gamma_{Y(0)}^{(s+1)}, \Gamma_{Y(1)}^{(s+1)}) \\ &= \arg \min_{\Gamma_{Y(0)}, \Gamma_{Y(1)}} \left\| \mathbb{H}_0 - \Gamma \left( \Gamma_X^{(s+1)}, \Gamma_{Y(0)} \right) \Lambda_0^{(s+1)} \right\|_F^2 + \left\| \mathbb{H}_1 - \Gamma \left( \Gamma_X^{(s+1)}, \Gamma_{Y(1)} \right) \Lambda_1^{(s+1)} \right\|_F^2 \end{aligned}$$

subject to  $\Gamma_{Y(0)} \in \mathbb{R}_+^{M_Y \times K}$ ,  $\Gamma_{Y(1)} \in \mathbb{R}_+^{M_Y \times K}$ ,  $\iota_{M_Y}^\top \Gamma_{Y(0)} = \iota_K^\top$ ,  $\iota_{M_Y}^\top \Gamma_{Y(1)} = \iota_K^\top$ .

5. Repeat 2-4 until convergence.

Each step of the iteration is a quadratic programming with linear constraints, which can be solved with a built-in optimization tool in most statistical softwares. The stepwise optimization assures a convergence to a local minimum. To find the global minimum, I consider various initial values  $(\Gamma_0^{(0)}, \Gamma_1^{(0)})$ .<sup>7</sup>

Let  $\hat{\Lambda}_0$ ,  $\hat{\Lambda}_1$ ,  $\hat{\Gamma}_0$  and  $\hat{\Gamma}_1$  denote the solution to the minimization problem. Note that when  $Y_i$  and  $X_i$  are discrete, the estimates  $\hat{\Gamma}_0$  and  $\hat{\Gamma}_1$  directly estimate the conditional distribution of  $Y_i(1)$  and  $Y_i(0)$  given  $U_i$ . When  $Y_i$  and  $X_i$  are continuous and therefore partitioning was used in constructing  $\mathbf{H}_0, \mathbf{H}_1$ , I use  $\hat{\Lambda}_0$  and  $\hat{\Lambda}_1$  to estimate the distribution of  $Y_i(1)$  and  $Y_i(0)$  given  $U_i$ .

---

<sup>7</sup>To initialize  $\Gamma_0^{(0)}, \Gamma_1^{(1)}$ , I use columns of  $\mathbb{H}_d$  and weighted sums of columns of  $\mathbb{H}_d$  with randomly drawn  $K$  sets of weights that sum to one as initial values. Alternatively, we can select the eigenvectors associated with the first  $K$  largest eigenvalues of  $\mathbb{H}_d^\top \mathbb{H}_d$  as an initial value.

### 3.2 Distributional treatment effect estimator

Before formally constructing the estimators for the DTE parameters, let us first establish that the DTE parameters are functions of quantities that are directly identified from the distribution of  $(Y_i, D_i, X_i, Z_i)$  and the nuisance parameters  $\Lambda_0, \Lambda_1$  estimated from the first-step nonnegative matrix factorization.

Firstly, let us see that the conditional distribution of  $Y_i(d)$  given  $U_i$  is a function of  $\Lambda_0, \Lambda_1$  and some observed quantities. Find that for any  $y \in \mathbb{R}$ ,

$$\begin{pmatrix} F_{Y|D=d,Z}(y|z^1) & \cdots & F_{Y|D=d,Z}(y|z^K) \end{pmatrix} = \begin{pmatrix} F_{Y(d)|U}(y|u^1) & \cdots & F_{Y(d)|U}(y|u^K) \end{pmatrix} \Lambda_d.$$

Since  $\Lambda_d$  is full rank, we have

$$\begin{pmatrix} F_{Y(d)|U}(y|u^1) & \cdots & F_{Y(d)|U}(y|u^K) \end{pmatrix} = \begin{pmatrix} F_{Y|D=d,Z}(y|z^1) & \cdots & F_{Y|D=d,Z}(y|z^K) \end{pmatrix} (\Lambda_d)^{-1}.$$

Thus, the conditional distribution  $F_{Y(d)|U}(\cdot|u)$  is characterized as a linear combination of the observed distributions  $\{F_{Y|D=d,Z}(\cdot|z)\}_{z=1}^K$  with  $(\Lambda_d)^{-1}$  as weights. For notational convenience, let  $\tilde{\Lambda}_d = (\Lambda_d)^{-1}$  for  $d = 0, 1$  and let  $\tilde{\lambda}_{jk,d}$  denote the  $j$ -th row and  $k$ -th column component of  $\tilde{\Lambda}_d$ . Then,  $(\tilde{\lambda}_{1k,d}, \dots, \tilde{\lambda}_{Kk,d})^\top$ , the  $k$ -th column of  $\tilde{\Lambda}_d$ , is the set of linear coefficients on  $\{F_{Y|D=d,Z}(\cdot|z)\}_{z=1}^K$  to retrieve the conditional distribution of  $Y_i(d)$  given  $U_i = u^k$ .

Secondly, the distribution of  $U_i$  is also a function of  $\Lambda_0, \Lambda_1$  and some observed quantities:

$$\begin{pmatrix} \Pr\{U_i = u^1\} \\ \vdots \\ \Pr\{U_i = u^K\} \end{pmatrix} = \Lambda_0 \begin{pmatrix} \Pr\{D_i = 0, Z_i = z^1\} \\ \vdots \\ \Pr\{D_i = 0, Z_i = z^K\} \end{pmatrix} + \Lambda_1 \begin{pmatrix} \Pr\{D_i = 1, Z_i = z^1\} \\ \vdots \\ \Pr\{D_i = 1, Z_i = z^K\} \end{pmatrix}. \quad (11)$$

Let  $p_U(k)$  denote  $\Pr\{U_i = u^k\}$  for  $k = 1, \dots, K$  and let  $p_{D,Z}(d, j)$  denote  $\Pr\{D_i = d, Z_i = z^j\}$  for  $d = 0, 1$  and  $j = 1, \dots, K$ .

By combining the two characterizations, we get that for any  $y, y' \in \mathbb{R}$ ,

$$\begin{aligned} F_{Y(0), Y(1)}(y, y') &= \sum_{k=1}^K p_U(k) F_{Y(0)}(y) F_{Y(1)}(y') \\ &= \sum_{j=1}^K \sum_{j'=1}^K \left( \sum_{k=1}^K p_U(k) \tilde{\lambda}_{jk,0} \tilde{\lambda}_{j'k,1} \right) F_{Y|D=0,Z}(y|z^j) \cdot F_{Y|D=1,Z}(y'|z^{j'}). \end{aligned}$$

$F_{Y(0), Y(1)}$  is a linear combination of  $\{F_{Y|D=0,Z}(y|z^j) \cdot F_{Y|D=1,Z}(y'|z^{j'})\}_{j,j'}$  where the weights are functions of  $\tilde{\Lambda}_0, \tilde{\Lambda}_1$  and  $\{p_U(u)\}_u$ . We can derive a characterization for  $F_{Y(1)-Y(0)}$  in a similar manner: for any  $\delta \in \mathbb{R}$ ,

$$F_{Y(1)-Y(0)}(\delta) = \sum_{j=1}^K \sum_{j'=1}^K \left( \sum_{k=1}^K p_U(k) \tilde{\lambda}_{jk,0} \tilde{\lambda}_{j'k,1} \right) \int_{\mathbb{R}} F_{Y|D=1,Z}(y + \delta|z^j) \cdot f_{Y|D=0,Z}(y|z^{j'}) dy.$$

There are two important observations to make here. Firstly, both of the DTE parameters are characterized as a weighted sum of quantities that are indexed by a pair of subpopulations  $\{i : D_i = 0, Z_i = z^j\}$  and  $\{i : D_i = 1, Z_i = z^{j'}\}$ . Secondly, each of the pairwise-indexed quantities is identified with a quadratic moment condition. I will build on these observations and develop an estimator for  $F_{Y(1)-Y(0)}$ . The adaptation for  $F_{Y(0), Y(1)}$  is straightforward.

Fix some  $\delta$  and let

$$\theta_{jj'} = \int_{\mathbb{R}} F_{Y|D=1,Z}(y + \delta|z^j) \cdot f_{Y|D=0,Z}(y|z^{j'}) dy$$

for  $j, j' = 1, \dots, K$ . Then,  $F_{Y(1)-Y(0)}(\delta) = \sum_{j=1}^K \sum_{j'=1}^K \left( \sum_{k=1}^K p_U(k) \tilde{\lambda}_{jk,0} \tilde{\lambda}_{j'k,1} \right) \theta_{jj'}$  and

$$\begin{aligned} \theta_{jj'} &= \mathbf{E} \left[ \frac{\mathbf{E}[\mathbf{1}\{Y_{i'} \leq Y_i + \delta, D_{i'} = 1, Z_{i'} = z^{j'}\}]}{\mathbf{E}[\mathbf{1}\{D_{i'} = 1, Z_{i'} = z^{j'}\}]} \middle| D_i = 0, Z_i = z^j \right] \\ &= \frac{\mathbf{E}[\mathbf{1}\{Y_{i'} \leq Y_i + \delta, D_i = 0, Z_i = z^j, D_i = 1, Z_{i'} = z^{j'}\}]}{\mathbf{E}[\mathbf{1}\{D_i = 0, Z_i = z^j, D_{i'} = 1, Z_{i'} = z^{j'}\}]} \end{aligned}$$

with  $(Y_i, D_i, Z_i) \perp\!\!\!\perp (Y_{i'}, D_{i'}, Z_{i'})$ . For notational simplicity, let  $\tilde{\lambda}$  denote the vectorized  $(\tilde{\Lambda}_0, \tilde{\Lambda}_1)$  and  $p$  denote the vector of  $\{p_U(k)\}_k$  and  $\{p_{D,Z}(d, j)\}_{d,j}$ . Then,  $\theta_{jj'}$  is identified

from a quadratic moment

$$\mathbf{E} \left[ m_{jj'} \left( W_i, W_{i'}; \theta_{jj'}, \tilde{\lambda}, p \right) \right] = 0$$

where  $W_i = (Y_i, D_i, X_i, Z_i)$  and

$$\begin{aligned} & m_{jj'} \left( W_i, W_{i'}; \theta_{jj'}, \tilde{\lambda}, p \right) \\ &= \frac{1}{p_{D,Z}(0, j) \cdot p_{D,Z}(1, j')} \cdot \left( \frac{1}{2} \mathbf{1}\{Y_{i'} \leq Y_i + \delta, D_i = 0, Z_i = z^j, D_i = 1, Z_{i'} = z^{j'}\} \right. \\ & \quad \left. + \frac{1}{2} \mathbf{1}\{Y_i \leq Y_{i'} + \delta, D_i = 1, Z_i = z^{j'}, D_i = 0, Z_{i'} = z^j\} \right) - \theta_{jj'}. \end{aligned}$$

Note that  $m_{jj'}$  is symmetric. By summing over  $j$  and  $j'$ , we can construct a moment function  $m = \sum_{j=1}^K \sum_{j'=1}^K \left( \sum_{k=1}^K p_U(k) \tilde{\lambda}_{jk,0} \tilde{\lambda}_{j'k,1} \right) m_{jj'}$  such that

$$\mathbf{E} \left[ m \left( W_i, W_{i'}; F_{Y(1)-Y(0)}(\delta), \tilde{\lambda}, p \right) \right] = 0 \quad (12)$$

identifies  $F_{Y(1)-Y(0)}(\delta)$ .

Thus, if the nuisance parameters  $\tilde{\lambda}, p$  were known, we can construct a GMM estimator using (12), and the standard asymptotic theory of  $U$ -statistic would apply. However, in practice, the nuisance parameters are estimated: with the first-step estimate  $\hat{\Lambda}_0$  and  $\hat{\Lambda}_1$ ,

$$\hat{\Lambda}_d = \left( \hat{\Lambda}_d \right)^{-1} \quad \text{for } d = 0, 1, \quad (13)$$

$$\hat{p}_U = \hat{\Lambda}_0 \begin{pmatrix} \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{D_i = 0, Z_i = z^1\} \\ \vdots \\ \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{D_i = 0, Z_i = z^K\} \end{pmatrix} + \hat{\Lambda}_1 \begin{pmatrix} \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{D_i = 1, Z_i = z^1\} \\ \vdots \\ \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{D_i = 1, Z_i = z^K\} \end{pmatrix}, \quad (14)$$

$$\hat{p}_{D,Z}(d, j) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{D_i = d, Z_i = z^j\}. \quad (15)$$

Thus, to account for the first step estimation error, I orthogonalize the moment function.

Even though the NMF estimators  $(\hat{\Lambda}_0, \hat{\Lambda}_1)$  and the induced estimators  $(\hat{\tilde{\Lambda}}_0, \hat{\tilde{\Lambda}}_1)$  are complex nonlinear functions of the data matrix  $\mathbb{H}_0$  and  $\mathbb{H}_1$ ,  $(\tilde{\Lambda}_0, \tilde{\Lambda}_1)$  satisfy the following

equations at their true values: for all  $(y, d, x, k)$ ,

$$\begin{aligned} \sum_{j=1}^K \tilde{\lambda}_{jk,d} \Pr \{Y_i = y, X_i = x | D_i = d, Z_i = z^j\} \\ = \sum_{j=1}^K \tilde{\lambda}_{jk,d} \Pr \{Y_i = y | D_i = d, Z_i = z^j\} \cdot \sum_{j=1}^K \tilde{\lambda}_{jk,d} \Pr \{X_i = x | D_i = d, Z_i = z^j\} \end{aligned} \quad (16)$$

$$\Pr \{X_i = x\} = \sum_{k=1}^K p_U(k) \sum_{j=1}^K \tilde{\lambda}_{jk,d} \Pr \{X_i = x | D_i = d, Z_i = z^j\}. \quad (17)$$

Equation (16) corresponds to (10) that  $Y_i(d) \perp\!\!\!\perp X_i \mid U_i$  and Equation (17) corresponds to the law of iterated expectation that  $\Pr\{X_i = x\} = \sum_{k=1}^K p_U(k) \Pr\{X_i = x | U_i = u^k\}$ . Given  $\{p_{D,Z}(d, j)\}_{d,j}$ , Equation (16) can be written as a quadratic moment condition and Equation (17) as a linear moment condition. The score function for these additional moments is

$$\begin{aligned} \phi(W_i, W_{i'}; \tilde{\lambda}, p) = \\ \left( \begin{aligned} & \sum_j \frac{\tilde{\lambda}_{j1,0}}{p_{D,Z}(0,j)} \cdot \frac{\mathbf{1}\{Y_i=y^1, D_i=0, X_i=x^1, Z_i=z^j\} + \mathbf{1}\{Y_{i'}=y^1, D_{i'}=0, X_{i'}=x^1, Z_{i'}=z^j\}}{2} - \\ & \sum_{j,j'} \frac{\tilde{\lambda}_{j1,0} \tilde{\lambda}_{j'1,0}}{p_{D,Z}(0,j) \cdot p_{D,Z}(0,j')} \cdot \frac{1}{2} \left( \mathbf{1}\{Y_i = y^1, D_i = 0, Z_i = z^j, X_{i'} = x^1, D_{i'} = 0, Z_{i'} = z^{j'}\} + \right. \\ & \quad \left. \mathbf{1}\{X_i = x^1, D_i = 0, Z_i = z^{j'}, Y_{i'} = y^1, D_{i'} = 0, Z_{i'} = z^j\} \right) \\ & \quad \vdots \\ & \sum_j \frac{\tilde{\lambda}_{jK,1}}{p_{D,Z}(1,j)} \cdot \frac{\mathbf{1}\{Y_i=y^{M_Y}, D_i=1, X_i=x^{M_X}, Z_i=z^j\} + \mathbf{1}\{Y_{i'}=y^{M_Y}, D_{i'}=1, X_{i'}=x^{M_X}, Z_{i'}=z^j\}}{2} - \\ & \sum_{j,j'} \frac{\tilde{\lambda}_{jK,1} \tilde{\lambda}_{j'K,1}}{p_{D,Z}(1,j) \cdot p_{D,Z}(1,j')} \cdot \frac{1}{2} \left( \mathbf{1}\{Y_i = y^{M_Y}, D_i = 1, Z_i = z^j, X_{i'} = x^{M_X}, D_{i'} = 1, Z_{i'} = z^{j'}\} + \right. \\ & \quad \left. \mathbf{1}\{X_i = x^{M_X}, D_i = 1, Z_i = z^{j'}, Y_{i'} = y^{M_Y}, D_{i'} = 1, Z_{i'} = z^j\} \right) \\ & \quad \frac{\mathbf{1}\{X_i=x^1\} + \mathbf{1}\{X_{i'}=x^1\}}{2} - \sum_k p_U(k) \sum_j \frac{\tilde{\lambda}_{jk,0}}{p_{D,Z}(0,j)} \cdot \frac{\mathbf{1}\{D_i=0, X_i=x^1, Z_i=z^j\} + \mathbf{1}\{D_{i'}=0, X_{i'}=x^1, Z_{i'}=z^j\}}{2} \\ & \quad \vdots \\ & \frac{\mathbf{1}\{X_i=x^{M_X}\} + \mathbf{1}\{X_{i'}=x^{M_X}\}}{2} - \sum_k p_U(k) \sum_j \frac{\tilde{\lambda}_{jk,1}}{p_{D,Z}(1,j)} \cdot \frac{\mathbf{1}\{D_i=1, X_i=x^{M_X}, Z_i=z^j\} + \mathbf{1}\{D_{i'}=1, X_{i'}=x^{M_X}, Z_{i'}=z^j\}}{2} \\ & \quad \frac{\mathbf{1}\{D_i=0, Z_i=z^1\} + \mathbf{1}\{D_{i'}=0, Z_{i'}=z^1\}}{2} - p_{D,Z}(0, 1) \\ & \quad \vdots \\ & \frac{\mathbf{1}\{D_i=1, Z_i=z^K\} + \mathbf{1}\{D_{i'}=1, Z_{i'}=z^K\}}{2} - p_{D,Z}(1, K) \end{aligned} \right). \end{aligned}$$

$\phi$  collects the quadratic moments from (16) across  $(y, d, x, k)$ , the linear moments from (17) across  $(d, x)$ , and the linear moments  $p_{D,Z}(d, j) = \mathbf{E}[\mathbf{1}\{D_i = d, Z_i = z^j\}]$  across  $(d, j)$ .

To use  $\phi$  to orthogonalize  $m$ , I show that the Jacobian matrix of  $\phi$  has full rank.



**Lemma 1.** *Assumptions 1-3 hold. Then,*

$$\begin{pmatrix} \mathbf{E} \left[ \frac{\partial}{\partial \lambda} \phi(W_i, W_{i'}; \tilde{\lambda}, p) \right] \\ \mathbf{E} \left[ \frac{\partial}{\partial p} \phi(W_i, W_{i'}; \tilde{\lambda}, p) \right] \end{pmatrix}$$

*has a full rank.*

The proof is provided in the Online Appendix. Then, with an additional nuisance parameter

$$\begin{aligned} \mu = & \begin{pmatrix} \mathbf{E} \left[ \frac{\partial}{\partial \lambda} \phi(W_i, W_{i'}; \tilde{\lambda}, p) \right] \\ \mathbf{E} \left[ \frac{\partial}{\partial p} \phi(W_i, W_{i'}; \tilde{\lambda}, p) \right] \end{pmatrix}^\top \\ & \cdot \left( \begin{pmatrix} \mathbf{E} \left[ \frac{\partial}{\partial \lambda} \phi(W_i, W_{i'}; \tilde{\lambda}, p) \right] \\ \mathbf{E} \left[ \frac{\partial}{\partial p} \phi(W_i, W_{i'}; \tilde{\lambda}, p) \right] \end{pmatrix} \begin{pmatrix} \mathbf{E} \left[ \frac{\partial}{\partial \lambda} \phi(W_i, W_{i'}; \tilde{\lambda}, p) \right] \\ \mathbf{E} \left[ \frac{\partial}{\partial p} \phi(W_i, W_{i'}; \tilde{\lambda}, p) \right] \end{pmatrix}^\top \right)^{-1} \\ & \cdot \begin{pmatrix} \mathbf{E} \left[ \frac{\partial}{\partial \lambda} m(W_i, W_{i'}; \tilde{\lambda}, p) \right] \\ \mathbf{E} \left[ \frac{\partial}{\partial p} m(W_i, W_{i'}; \tilde{\lambda}, p) \right] \end{pmatrix}, \end{aligned}$$

the score

$$\psi(W_i, W_{i'}; F_{Y(1)-Y(0)}(\delta), \tilde{\lambda}, p, \mu) = m(W_i, W_{i'}; F_{Y(1)-Y(0)}(\delta), \tilde{\lambda}, p) - \mu^\top \phi(W_i, W_{i'}; \tilde{\lambda}, p)$$

satisfies the Neyman orthogonality.

Let  $\hat{\lambda}$  and  $\hat{p}$  denote the (vectorized) nuisance parameter estimators from (13)-(15) and  $\hat{\mu}$  denote the plug-in, sample analogue estimator of  $\mu$ . I estimate  $F_{Y(1)-Y(0)}(\delta)$  with

$$\binom{n}{2}^{-1} \sum_{i < i'} \psi \left( W_i, W_{i'}; \hat{F}_{Y(1)-Y(0)}(\delta), \hat{\lambda}, \hat{p}, \hat{\mu} \right) = 0.$$

The DTE estimator  $\hat{F}_{Y(0), Y(1)}(y, y')$  is constructed in a similar manner.

### 3.3 Asymptotic properties

Theorem 2 establishes the consistency of the first-step estimators  $\hat{\Lambda}_0$  and  $\hat{\Lambda}_1$ .

**Theorem 2.** *Assumptions 1-3 hold. Up to some permutation on  $\{u^1, \dots, u^K\}$ ,*

$$\left\| \widehat{\Lambda}_0 - \Lambda_0 \right\|_F = O_p \left( \frac{1}{\sqrt{n}} \right) \quad \text{and} \quad \left\| \widehat{\Lambda}_1 - \Lambda_1 \right\|_F = O_p \left( \frac{1}{\sqrt{n}} \right)$$

as  $n \rightarrow \infty$ .

*Proof.* See Appendix. □

Theorem 2 is the key result in the asymptotic theory for the DTE estimator:  $\widehat{\Lambda}_0, \widehat{\Lambda}_1$  are  $\sqrt{n}$ -consistent. In the proof, I also show that  $\widehat{\Gamma}_0, \widehat{\Gamma}_1$  are  $\sqrt{n}$ -consistent estimators of  $\Gamma_0, \Gamma_1$ , up to some permutation. A direct corollary of Theorem 2 is that  $\widehat{\lambda}$  is consistent for  $\tilde{\lambda}$  at the rate of  $n^{-\frac{1}{2}}$ .

Theorem 3 establishes the asymptotic normality of the distributional treatment effect estimators.

**Theorem 3.** *Assumptions 1-3 hold. Then, for any  $y, y', \delta \in \mathbb{R}^2$  and  $(x^1, \dots, x^M) \subset \mathbb{R}$ ,*

$$\begin{aligned} \sqrt{n} \left( \widehat{F}_{Y(0), Y(1)}(y, y') - F_{Y(0), Y(1)}(y, y') \right) &\xrightarrow{d} \mathcal{N}(0, \sigma(y, y')^2) \\ \sqrt{n} \left( \widehat{F}_{Y(1) - Y(0)}(\delta) - F_{Y(1) - Y(0)}(\delta) \right) &\xrightarrow{d} \mathcal{N}(0, \sigma(\delta)^2) \end{aligned}$$

as  $n \rightarrow \infty$  with some consistently estimable  $\sigma(y, y')^2$  and  $\sigma(\delta)^2$ .

*Proof.* From Theorem 2,  $\widehat{\lambda}$  is consistent for  $\tilde{\lambda}$  at the rate of  $n^{-\frac{1}{2}}$ . Thus,  $(\widehat{p}, \widehat{\mu})$  are consistent for  $(p, \mu)$  at the rate of  $n^{-\frac{1}{2}}$  as well. Then, from the central limit theorem for  $U$ -statistics and the orthogonality of the score function  $\psi$ , the asymptotic normality is established. □

The asymptotic variances are computed from a projection of the orthogonal scores:

$$\tilde{\psi}(w) = \mathbf{E} [\psi(W_i, w)] \quad \text{and} \quad \sigma^2 = 4\mathbf{E} \left[ \tilde{\psi}(W_i)^2 \right].$$

In Sections 4-5, the standard errors are obtained by estimating the asymptotic variance with plug-in estimators.

## 4 Simulation

In this section, I discuss Monte Carlo simulation results. I generated  $B = 200$  random samples from DGPs with discrete  $Y_i(1), Y_i(0), X_i, Z_i$  and  $U_i$  where  $M_Y = 3, M_X = 6, M_Z = 3$  and  $K = 3$ :  $Y_i \in \{1, 2, 3\}$ ,  $X_i \in \{1, 2, 3, 4, 5, 6\}$  and  $Z_i \in \{1, 2, 3\}$ .<sup>8</sup> The treatment  $D_i$  was drawn randomly, independent of  $Y_i(1), Y_i(0), X_i, Z_i$ . In the first step nonnegative matrix factorization, I collapsed the support of  $X_i$  so that the effective number of points in the support of  $X_i$  is three. Thus, the conditional probability matrix  $\mathbb{H}_0$  and  $\mathbb{H}_1$  were  $9 \times 3$  matrices. Across difference DGPs, I varied  $\Lambda$ , the conditional probability of  $U_i$  given  $Z_i$  which is shared across treated and untreated subpopulation, to vary the informativeness of the proxy variable  $Z_i$  with regard to the latent variable  $U_i$ .

Table 1 contains the bias and the root mean squared error (rMSE) of the distributional treatment effect estimators  $\hat{F}_{Y(1)-Y(0)}$ . As  $\Lambda$  becomes less informative about the distribution of  $U_i$ , i.e. the smallest singular value  $\sigma_{\min}(\Lambda)$  decreases, the rMSE goes up. This suggests that

<sup>8</sup>The specifics of the DGPs are as follows:  $p_U = (0.286, 0.286, 0.438)$ ,

$$\Gamma_X = \begin{pmatrix} 0.778 & 0.028 & 0.022 \\ 0.067 & 0.050 & 0.033 \\ 0.056 & 0.422 & 0.044 \\ 0.044 & 0.422 & 0.056 \\ 0.033 & 0.050 & 0.067 \\ 0.022 & 0.028 & 0.778 \end{pmatrix}, \quad \Gamma_{Y(1)} = \begin{pmatrix} 0.656 & 0.022 & 0.000 \\ 0.117 & 0.706 & 0.117 \\ 0.228 & 0.272 & 0.883 \end{pmatrix}, \quad \Gamma_{Y(0)} = \begin{pmatrix} 0.756 & 0.122 & 0.078 \\ 0.167 & 0.756 & 0.167 \\ 0.078 & 0.122 & 0.756 \end{pmatrix},$$

and  $\Lambda$ s in the order of decreasing smallest singular value are

$$\Lambda = \begin{pmatrix} 0.840 & 0.091 & 0.040 \\ 0.077 & 0.772 & 0.056 \\ 0.083 & 0.137 & 0.905 \end{pmatrix}, \quad \begin{pmatrix} 0.722 & 0.134 & 0.078 \\ 0.124 & 0.665 & 0.095 \\ 0.154 & 0.201 & 0.827 \end{pmatrix}, \quad \begin{pmatrix} 0.611 & 0.175 & 0.120 \\ 0.168 & 0.563 & 0.137 \\ 0.221 & 0.262 & 0.744 \end{pmatrix}.$$

$\hat{F}_{Y(1)-Y(0)}$						
$\delta$	$\sigma_{\min}(\Lambda) = 0.701$		$\sigma_{\min}(\Lambda) = 0.501$		$\sigma_{\min}(\Lambda) = 0.310$	
	bias	rMSE	bias	rMSE	bias	rMSE
-2	0.000	0.006	0.001	0.010	0.001	0.025
-1	-0.000	0.017	0.000	0.025	-0.002	0.052
0	-0.007	0.028	-0.012	0.040	-0.014	0.076
1	-0.009	0.025	-0.014	0.040	-0.015	0.084

Table 1: Bias and rMSE of DTE estimator,  $B = 200$ .

	$\widehat{F}_{Y(1)-Y(0)}$		
	$\sigma_{\min}(\Lambda) = 0.701$	$\sigma_{\min}(\Lambda) = 0.501$	$\sigma_{\min}(\Lambda) = 0.310$
$\Pr \{F_{Y(1)-Y(0)}(-2) \in \widehat{CI}\}$	0.968	0.970	0.990
$\Pr \{F_{Y(1)-Y(0)}(-1) \in \widehat{CI}\}$	0.978	0.960	0.970
$\Pr \{F_{Y(1)-Y(0)}(0) \in \widehat{CI}\}$	0.960	0.975	0.990
$\Pr \{F_{Y(1)-Y(0)}(1) \in \widehat{CI}\}$	0.970	0.970	0.980
$\Pr \{\text{reject } F_{X D=1,U} = F_{X D=0,U}\}$	0.070	0.063	0.049

Table 2: Coverage of CI and type I error of falsification test,  $B = 200$ .

the first step nonnegative matrix factorization estimation quality depends on how informative the proxy variables  $X_i$  and  $Z_i$  are for the latent variable  $U_i$ . Additionally, Table 2 contains the coverage probability of the confidence interval constructed with the asymptotic standard error and the type  $I$  error of the falsification test proposed in Subsection 2.2. The 95% confidence interval shows mostly correct coverage, sometimes slightly too conservative, and the falsification test is valid.

## 5 Empirical illustration

In this section, we revisit Jones et al. [2019] and estimate the distributional treatment effect of workplace wellness program on medical spending. Jointly with the Campus Well-being Services at the University of Illinois Urbana-Champaign, the authors of Jones et al. [2019] conducted a large-scale randomized controlled trials. The experiment started in July 2016, by inviting 12,459 eligible university employees to participate in an online survey. Of 4,834 employees who completed the survey, 3,300 employees were randomly selected into treatment, being offered to participate in a workplace wellness program names iThrive. The participation itself was not enforced; the treated individuals were merely financially incentivized to participate by being offered monetary reward for completing each step of the wellness program. Thus, the main treatment effect parameter of Jones et al. [2019] is the ‘intent-to-treat’ effect. The workplace wellness program consisted of various activities such as chronic disease management, weight management, and etc. The treated individuals were offered to participate in the wellness program starting the fall semester of 2016, until the

spring semester of 2018.

One of the main outcome variables that Jones et al. [2019] studied is the monthly medical spending. Since the authors had access to the university-sponsored health insurance data, they had detailed information on the medical spending behaviors of the participants. Taking advantage of the randomness in assigning eligibility to the participants, Jones et al. [2019] estimated the intent-to-treat type ATE of the workplace wellness program on the monthly medical spending. The ATE estimate on the first-year monthly medical spending, from August 2016 to July 2017, showed that the eligibility for the wellness program raised the monthly medical spending by \$10.8, with  $p$ -value of 0.937, finding no significant intent-to-treat effect.

In Jones et al. [2019], the authors acknowledge that the null effect on the mean does not necessarily mean null effect everywhere, though they themselves do not explore the treatment effect heterogeneity in the paper.<sup>9</sup> On page 1890, Jones et al. [2019] state “there may exist subpopulations who did benefit from the intervention or who would have benefited had they participated.”<sup>10</sup> I build onto this observation and estimate the distributional treatment effect of the randomly assigned eligibility for the wellness program. By looking at the distribution, I find the proportion of the subpopulation among treated population that benefited from the treatment.

The dataset built by the authors of Jones et al. [2019] fits the context of the short panel model in Example 2. For each individual, the dataset contains monthly medical spending records for the following three time durations: July 2015-July 2016, August 2016-July 2017 and August 2017-January 2019. Since the experiment started in the summer of 2016 and the treated individuals were offered to participate in the wellness program starting the fall semester of 2016, the monthly medical spending record for July 2015-July 2016 could be thought of as a ‘pretreatment’ outcome variable. Thus, we could use the information from

---

<sup>9</sup>In the original dataset used in Jones et al. [2019], the authors had connected the medical spending variables to additional survey variables such as age, health behavior, salary, etc. They did not explore how the treatment effect interacts with the additional characteristics, but they did add these additional control variables through double Lasso. Adding the control variables increased the point estimate for the ATE (\$34.9) but the estimate still remained insignificant, with  $p$ -value being 0.859.

<sup>10</sup>Damon Jones, David Molitor, and Julian Reif, “What do workplace wellness programs do? Evidence from the Illinois workplace wellness study,” *The Quarterly Journal of Economics*, vol. 134, no.4 (2019): 1747-1791.

the distribution of the pretreatment outcome variable to connect the treated subpopulation and the untreated subpopulation. The followings are the variables taken from the dataset.

$Y_i$  : monthly medical spending for August 2016-July 2017

$D_i$  : a binary variable for whether eligible to participate in the wellness program

$X_i$  : monthly medical spending for July 2015-July 2016

$Z_i$  : monthly medical spending for August 2017-January 2019

In this specific empirical context, the common shock  $V_{it}$  can be thought of as underlying health status and the regime-specific shocks  $(\varepsilon_{it}(1), \varepsilon_{it}(0))$  could be thought of as additional random shocks such as susceptibility to the workplace wellness program or transient health shock which does not persist over time.<sup>11</sup> The first-order Markovian assumption in Example 2 is consistent with the health economics literature and broader economics literature of modeling household choices regarding health expenditure: Grossman [1972], Wagstaff [1993], Jacobson [2000], Yogo [2016] and more. Applying assumptions in Example 2, the treatment is allowed to affect the underlying health status in the post-treatment period of August 2017-January 2019, but is assumed to be independent of the underlying health status in July 2015-July 2017.

To choose,  $K$ , the size of the support for the latent variable  $U_i$ , I applied the eigenvalue ratio estimator and the Kleibergen-Paap rank test to a  $12 \times 10$  matrix of  $\mathbf{H}_X$ .<sup>12</sup> Under Assumptions 1-3, the matrix is at most rank  $K$ . Given a matrix, the eigenvalue ratio estimator from Ahn and Horenstein [2013] estimates  $K$  by choosing the value of  $K$  maximizes eigenvalue/growth ratio and the Kleibergen-Paap test statistic tests the null hypothesis of rank  $K$  against the alternative of rank being bigger than  $K$ . For detailed discussion, see Appendix C.

---

<sup>11</sup>As discussed in Section 2, the framework of this paper is best suited for empirical contexts where a treatment is regime-changing and introduces a new source of heterogeneity in outcome. The first condition may hold when the information given during the workplace wellness program systemically changes one's health behaviors by motivating them to seek out advice from medical professionals or schedule checkups. Also, by limiting our attention to individuals with no prior exposure to activities offered in the workplace wellness program, the second condition can be satisfied.

<sup>12</sup>The partition for  $X_i$  is constructed with  $F_X^{-1}(1/12), \dots, F_X^{-1}(11/12)$  and the partition for  $Z_i$  is constructed with  $F_Z^{-1}(1/5), \dots, F_Z^{-1}(4/5)$ .

$K$	1	2	3	4	5	6	7	8
eigenvalue ratio	3.505	3.991	4.029	2.721	1.653	1.863	1.418	3.309
growth ratio	0.964	1.135	1.472	1.353	0.893	0.956	0.580	1.035

Table 3: Eigenvalue ratios and growth ratios

$K$	1	2	3	4	5	6
test statistic	884.82	116.23	35.75	20.08	13.80	7.94
$p$ -value	0.000	0.001	0.984	0.998	0.995	0.992

Table 4: Kleibergen-Paap rank test statistics for  $H_0 : \text{rank} = K$  and their  $p$ -values

Both the eigenvalue ratio estimator and the Kleibergen-Paap rank test suggest  $K = 3$ . To be conservative, I used  $K = 5$  for the main estimation specification. Thus, when constructing  $\mathbb{H}_0$  and  $\mathbb{H}_1$  to be used in the first step nonnegative matrix factorization, I used the quintiles of the marginal distributions:  $(-\infty, F_Y^{-1}(0.2), F_Y^{-1}(0.4), F_Y^{-1}(0.6), F_Y^{-1}(0.8), \infty)$  and similarly for  $X_i$ . The matrices  $\mathbb{H}_0$  and  $\mathbb{H}_1$  are  $25 \times 5$ -dimensional.

With  $K = 5$ , the conditional independence framework of the paper passes the falsification test.  $T_n$  from (20) is computed with  $(x^1, \dots, x^4) = (F_X^{-1}(0.2), F_X^{-1}(0.4), F_X^{-1}(0.6), F_X^{-1}(0.8))$  and therefore

$$T_n \xrightarrow{d} \chi^2(20)$$

as  $n \rightarrow \infty$ .  $T_n$  is 16.435 and its  $p$ -value is 0.689, passing the falsification test.

Figure 1 contains the estimated joint distribution of the two potential outcomes from the nonnegative matrix factorization algorithm with  $K = 5$ . For visibility, I first partitioned the potential outcome variable with quantiles  $F_Y^{-1}(1/7), \dots, F_Y^{-1}(6/7)$  and plotted the joint distribution of partitioned potential outcomes. Since the treated potential outcomes are plotted on the vertical axis, higher mass on the left-upper triangle means that the treatment reduces the medical spending. Overall, there is no definitive pattern. One notable observation is that the joint density is higher where  $F_Y(Y_i(1)) \approx F_Y(Y_i(0)) \approx 0$  and  $F_Y(Y_i(1)) \approx F_Y(Y_i(0)) \approx 1$ . This is intuitive since on the two ends of the underlying health status spectrum, the effectiveness of the workplace wellness program must be limited.

Figure 2 contains the estimated marginal distribution of the treatment effect and its 95% pointwise confidence interval. Note that the point estimates are mostly upward-sloping and lie between zero and one. Though the quadratic moment representation used in the DTE estimators does not impose any monotonicity or nonnegativity restrictions, the estimated marginal distribution violates these constraints only on a small subset of the range  $[-1000, 1000]$ . Overall, it is unclear if more than half of the people would be better off from the treatment; the confidence interval for  $\Pr\{Y_i(1) - Y_i(0) \geq 0\}$  contains 0.5, not being able to reject the null  $\Pr\{Y_i(1) - Y_i(0) \geq 0\} \leq 0.5$ .

As comparison, estimates for the upper bound and the lower bound from Makarov [1982], Fan and Park [2010] are also provided in Figure 2, as green dotted lines. The point estimates are consistent with the partial identification result, lying between the lower bound and the upper bound. The comparison highlights the gain of the point identification result, at the cost of assuming stronger identifying assumptions. For  $\delta \in [-500, 600]$ , the 95% confidence interval is included in the partially identified set, giving us much bigger power in inference.

Lastly, the point identification helps us analyze the pattern of the treatment heterogeneity. Recall that the ATE estimate was inconclusive about the effectiveness of the treatment. However, the DTE estimates on  $\Pr\{Y_i(1) - Y_i(0) \leq \delta\}$  for  $\delta \leq -600$  and the DTE estimates on  $\Pr\{Y_i(1) - Y_i(0) \leq \delta\}$  for  $\delta \geq 400$  shows us interesting treatment effect het-

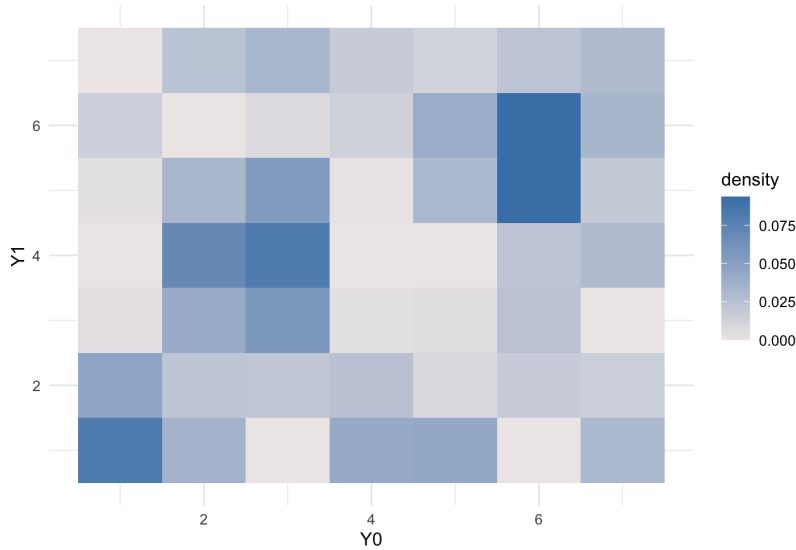


Figure 1: Joint density of  $F_Y(Y_i(1))$  and  $F_Y(Y_i(0))$ ,  $K = 5$ .



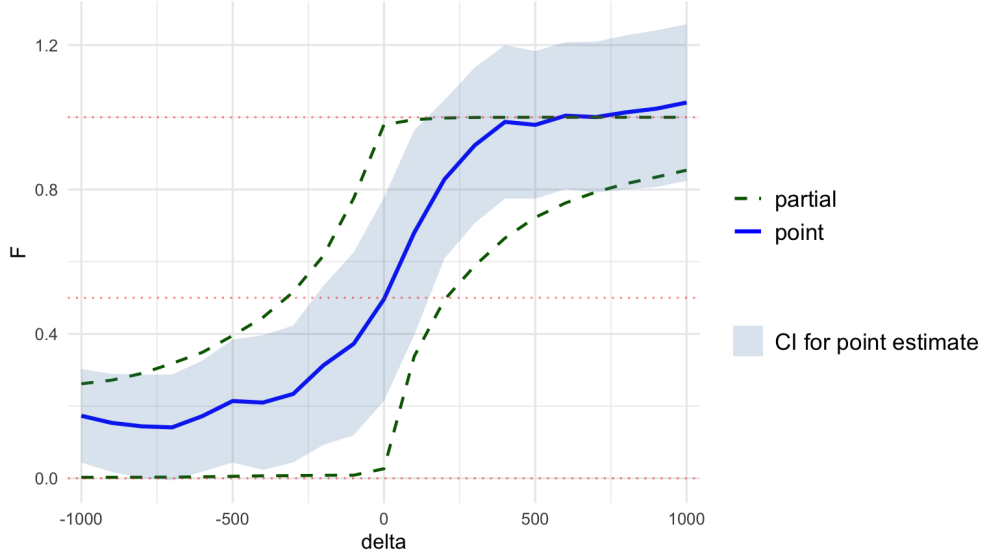


Figure 2: Marginal distribution of  $Y_i(1) - Y_i(0)$ ,  $K = 5$ .

erogeneity patterns, in favor of implementing the treatment. The negative impact of the treatment, i.e. how much more money you spend under the treatment, is capped at \$400:  $\hat{F}_{Y(1)-Y(0)}(400) \approx 1$ . On the other hand, the left tail of the treatment effect distribution is thicker, implying that some people are greatly benefited from participating in the program:  $\hat{F}_{Y(1)-Y(0)}(-600) \approx 0.15$ .

## 6 Conclusion

An important avenue for future research is how to extend the current framework to account for a continuous latent variable  $U_i$ , while retaining the desirable properties of the discretization-based estimation strategy. Firstly, the conditional independence with a discrete  $U_i$  could be relaxed by adopting concepts of partial independence, which would lead to a partial identification of the DTE parameters. Alternatively, the discretization bias could be explicitly addressed. Once we establish the asymptotic behavior of the discretization bias, we may be able to use a subsampling-based method such as the jackknife correction to remove the bias coming from discretization.

## References

- Seung C Ahn and Alex R Horenstein. Eigenvalue ratio test for the number of factors. *Econometrica*, 81(3):1203–1227, 2013.
- Peter Arcidiacono and Robert A Miller. Conditional choice probability estimation of dynamic discrete choice models with unobserved heterogeneity. *Econometrica*, 79(6):1823–1867, 2011.
- Susan Athey and Guido W Imbens. Identification and inference in nonlinear difference-in-differences models. *Econometrica*, 74(2):431–497, 2006.
- Orazio Attanasio, Sarah Cattan, Emla Fitzsimons, Costas Meghir, and Marta Rubio-Codina. Estimating the production function for human capital: results from a randomized controlled trial in colombia. *American Economic Review*, 110(1):48–85, 2020.
- Abhijit V Banerjee, Shawn Cole, Esther Duflo, and Leigh Linden. Remedying education: Evidence from two randomized experiments in india. *The quarterly journal of economics*, 122(3):1235–1264, 2007.
- Guadalupe Bedoya, Luca Bittarello, Jonathan Davis, and Nikolas Mittag. Distributional impact analysis: Toolkit and illustrations of impacts beyond the average treatment effect. Technical report, IZA Discussion Papers, 2018.
- Stéphane Bonhomme, Thibaut Lamadon, and Elena Manresa. Discretizing unobserved heterogeneity. *Econometrica*, 90(2):625–643, 2022.
- Brantly Callaway and Tong Li. Quantile treatment effects in difference in differences models with panel data. *Quantitative Economics*, 10(4):1579–1618, 2019.
- Pedro Carneiro, Karsten T. Hansen, and James J. Heckman. 2001 lawrence r. klein lecture estimating distributions of treatment effects with an application to the returns to schooling and measurement of the effects of uncertainty on college choice\*. *International Economic Review*, 44(2):361–422, 2003.

- Victor Chernozhukov and Christian Hansen. An iv model of quantile treatment effects. *Econometrica*, 73(1):245–261, 2005.
- Victor Chernozhukov and Christian Hansen. Instrumental quantile regression inference for structural and treatment effect models. *Journal of Econometrics*, 132(2):491–525, 2006.
- Flavio Cunha and James J Heckman. Formulating, identifying and estimating the technology of cognitive and noncognitive skill formation. *Journal of human resources*, 43(4):738–782, 2008.
- Flavio Cunha, James J Heckman, and Susanne M Schennach. Estimating the technology of cognitive and noncognitive skill formation. *Econometrica*, 78(3):883–931, 2010.
- Ben Deaner. Proxy controls and panel data, 2023.
- Yanqin Fan and Sang Soo Park. Sharp bounds on the distribution of treatment effects and their statistical inference. *Econometric Theory*, 26(3):931–951, 2010.
- Yanqin Fan, Robert Sherman, and Matthew Shum. Identifying treatment effects under data combination. *Econometrica*, 82(2):811–822, 2014.
- Sergio Firpo and Cristine Pinto. Identification and estimation of distributional impacts of interventions using changes in inequality measures. *Journal of Applied Econometrics*, 31(3):457–486, 2016.
- Sergio Firpo and Geert Ridder. Partial identification of the treatment effect distribution and its functionals. *Journal of Econometrics*, 213(1):210–234, 2019.
- Brigham R Frandsen and Lars J Lefgren. Partial identification of the distribution of treatment effects with an application to the knowledge is power program (kipp). *Quantitative Economics*, 12(1):143–171, 2021.
- Michael Grossman. On the concept of health capital and the demand for health. *Journal of Political economy*, 80(2):223–255, 1972.
- Sukjin Han and Haiqing Xu. On quantile treatment effects, rank similarity, and variation of instrumental variables. *arXiv preprint arXiv:2311.15871*, 2023.

- James J Heckman, Jeffrey Smith, and Nancy Clements. Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts. *The Review of Economic Studies*, 64(4):487–535, 1997.
- Marc Henry, Yuichi Kitamura, and Bernard Salanié. Partial identification of finite mixtures in econometric models. *Quantitative Economics*, 5(1):123–144, 2014.
- Ayden Higgins. Panel data models with interactive fixed effects and relatively small t. *working paper*, 2025.
- Yingyao Hu. Identification and estimation of nonlinear models with misclassification error using instrumental variables: A general solution. *Journal of Econometrics*, 144(1):27–61, 2008.
- Yingyao Hu and Yuya Sasaki. Closed-form identification of dynamic discrete choice models with proxies for unobserved state variables. *Econometric Theory*, 34(1):166–185, 2018.
- Yingyao Hu and Susanne M Schennach. Instrumental variable treatment of nonclassical measurement error models. *Econometrica*, 76(1):195–216, 2008.
- Yingyao Hu and Matthew Shum. Nonparametric identification of dynamic models with unobserved state variables. *Journal of Econometrics*, 171(1):32–44, 2012.
- Lena Jacobson. The family as producer of health—an extended grossman model. *Journal of health economics*, 19(5):611–637, 2000.
- Damon Jones, David Molitor, and Julian Reif. What do workplace wellness programs do? evidence from the illinois workplace wellness study. *The Quarterly Journal of Economics*, 134(4):1747–1791, 2019.
- Tetsuya Kaji and Jianfei Cao. Assessing heterogeneity of treatment effects, 2023.
- Hiroyuki Kasahara and Katsumi Shimotsu. Nonparametric identification of finite mixture models of dynamic discrete choices. *Econometrica*, 77(1):135–175, 2009.
- Desire Kedagni. Identifying treatment effects in the presence of confounded types. *Journal of Econometrics*, 234(2):479–511, 2023.

- Frank Kleibergen and Richard Paap. Generalized reduced rank tests using the singular value decomposition. *Journal of econometrics*, 133(1):97–126, 2006.
- GD Makarov. Estimates for the distribution function of a sum of two random variables when the marginal distributions are fixed. *Theory of Probability & its Applications*, 26(4):803–806, 1982.
- Wang Miao, Zhi Geng, and Eric J Tchetgen Tchetgen. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993, 2018.
- Karthik Muralidharan, Abhijeet Singh, and Alejandro J Ganimian. Disrupting education? experimental evidence on technology-aided instruction in india. *American Economic Review*, 109(4):1426–1460, 2019.
- Kenichi Nagasawa. Treatment effect estimation with noisy conditioning variables. *arXiv preprint arXiv:1811.00667*, 2022.
- Sungho Noh. Nonparametric identification and estimation of heterogeneous causal effects under conditional independence. *Econometric Reviews*, 42(3):307–341, 2023.
- Quang Vuong and Haiqing Xu. Counterfactual mapping and individual treatment effects in nonseparable models with binary endogeneity. *Quantitative Economics*, 8(2):589–610, 2017.
- Adam Wagstaff. The demand for health: an empirical reformulation of the grossman model. *Health Economics*, 2(2):189–198, 1993.
- Ximing Wu and Jeffrey M Perloff. Information-theoretic deconvolution approximation of treatment effect distribution. *Available at SSRN 903982*, 2006.
- Motohiro Yogo. Portfolio choice in retirement: Health risk and the demand for annuities, housing, and risky assets. *Journal of monetary economics*, 80:17–34, 2016.

# APPENDIX

## A Multidimensional $U_i$

The identification result of this paper holds with a multidimensional latent variable  $U_i$ , given that the proxy variables  $X_i$  and  $Z_i$  are at least of the same dimension. This is because the spectral decomposition result of Hu and Schennach [2008] that this paper builds on holds also for multivariate  $U_i$ ,  $X_i$ , and  $Z_i$ . (Theorem 1 of Hu and Schennach [2008])<sup>13</sup> Suppose that  $U_i, X_i, Z_i \in \mathbb{R}^p$  with some  $p \geq 1$ . The following assumption collects Assumptions 1 and 3-5 of Hu and Schennach [2008], replacing Assumption 4 for multidimensional  $U_i$ .

**Assumption 6.** *Assume*

- a. (multidimensional  $U_i$ )  $\mathcal{U} \subset \mathbb{R}^p$ .*
- b. (bounded density) The conditional densities  $f_{Y(1)|U}$ ,  $f_{Y(0)|U}$ ,  $f_{X|U}$ ,  $f_{U|D=1,Z}$  and  $f_{U|D=0,Z}$  and the marginal densities  $f_U$ ,  $f_{Z|D=1}$  and  $f_{Z|D=0}$  are bounded.*
- c. (completeness) The integral operators  $L_{X|U}$ ,  $L_{Z|D=1,U}$  and  $L_{Z|D=0,U}$  are injective on  $\mathcal{L}^1(\mathbb{R}^p)$ .*
- d. (no repeated eigenvalue) For any  $u \neq u'$ ,*

$$\Pr \{f_{Y(d)|U}(Y_i|u) \neq f_{Y(d)|U}(Y_i|u') | D_i = d\} > 0$$

*for each  $d = 0, 1$ .*

- e. (measurement error) There exists a functional  $M$  defined on  $\mathcal{L}^1(\mathbb{R}^p)$  such that*

$$M f_{X|U}(\cdot|u) = u \quad \text{for all } u \in \mathcal{U}.$$

Under Assumptions 1-2 and 6, the joint distribution of  $(Y_i(1), Y_i(0), D_i, X_i, Z_i, U_i)$  is identified.

---

<sup>13</sup>This point is also utilized in Cunha et al. [2010] again. (Theorem 2 of Cunha et al. [2010])

An important caveat of modeling the latent  $U_i$  to be multidimensional is that we cannot use the information from the conditional distribution of  $Y_i(d)$  given  $U_i$  to find a labeling on  $U_i$ , since  $Y_i(0)$  and  $Y_i(1)$  are univariate while  $U_i$  is not. Thus, in Assumption 6.e, I fully adopt the ‘repeated measure’ interpretation on the proxy variables  $X_i$  and  $Z_i$ , as in Cunha et al. [2010], Attanasio et al. [2020] and Example 1, to find a labeling on  $U_i$ .

Importantly, Assumption 6 does not impose any restriction on the dependence structure within  $U_i$ ,  $X_i$  and  $Z_i$ . Different dimensions of the latent variable  $U_i$  may be correlated. In addition, Assumption 6 does not impose any relationship between different dimensions of  $U_i$  and those of  $X_i$  and  $Z_i$ . We do not need each dimension of  $X_i$  and  $Z_i$  to correspond to a specific dimension of  $U_i$ . However, in practice, such knowledge may help in terms of estimation. For example, Cunha et al. [2010], Attanasio et al. [2020] theorize two dimensions of  $U_i$ : cognitive ability and noncognitive/socio-emotional ability. Given information on what the available measures of latent ability are designed to measure, both of the papers match a subset of the proxy variables to cognitive ability and another to noncognitive/socio-emotional ability.

Similarly, the estimation strategy proposed in this paper may benefit from such knowledge on the proxy variable. For example, suppose that  $U_i, X_i, Z_i$  are all two-dimensional vectors, with a finite support:

$$U_i = \left( U_{i1}, U_{i2} \right)^\top, \quad X_i = \left( X_{i1}, X_{i2} \right)^\top, \quad Z_i = \left( Z_{i1}, Z_{i2} \right)^\top.$$

$|\mathcal{U}_1| = |\mathcal{X}_1| = |\mathcal{Z}_1| = K_1$  and  $|\mathcal{U}_2| = |\mathcal{X}_2| = |\mathcal{Z}_2| = K_2$ , with  $K = K_1 \cdot K_2$ .  $(X_{i1}, Z_{i1})$  are proxies for  $U_{i1}$  and  $(X_{i2}, Z_{i2})$  for  $U_{i2}$ . Then, we can modify the nonnegative matrix factorization problem (7) as follows:

1. Label the rows of  $\Lambda_0, \Lambda_1$  and the columns of  $\Gamma_0, \Gamma_1$  to each dimension of  $U_i$ . For example, the first  $K_1$  rows of  $\Lambda_0, \Lambda_1$  and the first  $K_1$  columns of  $\Gamma_0, \Gamma_1$  correspond to the same value of  $U_{i1}$  and so on.
2. Add additional constraints on  $\Gamma_0$  and  $\Gamma_1$  such that each column of  $\Gamma_0$  and  $\Gamma_1$  satisfy

- (a)  $Y_i(d)$ ,  $X_{i1}$ , and  $X_{i2}$  are mutually independent of each other given  $U_i$ ;

- (b) Conditional distribution of  $X_{i1}$  is equal across the columns of  $\Gamma_0, \Gamma_1$  that correspond to the same value of  $U_{i1}$  and likewise for  $X_{i2}$  and  $U_{i2}$ .

These additional constraints from the knowledge on the measures  $X_i$  induce more regularization on the finite mixture model. Note that the matching between the dimensions of  $U_i$  and those of  $Z_i$  does not impose any constraints on  $\Gamma_0, \Gamma_1$  since  $U_{i1}$  and  $U_{i2}$  can be arbitrarily correlated.

## B Falsification tests

In this section, I formally develop two test statistics, in the context of Assumption 3.a. Firstly, I construct a test statistic based on (6). For the test statistic based on (6), we need to modify the nonnegative matrix factorization problem (7). By construction, the NMF algorithm described in Subsection 3.1 imposes the conditional independence between  $X_i$  and  $D_i$  given  $U_i$ , invalidating the falsification test based on (6). Thus, I modify (7) by dropping the linear constraints (8). As long as we impose the quadratic constraints (10), the NMF optimization problem still admits a unique solution, when  $\mathbb{H}_d$  is close to  $\mathbf{H}_d$ .

Given  $\tilde{\Lambda}_0, \tilde{\Lambda}_1$ , find that

$$F_{X|U}(x|u^k) = \sum_{j=1}^j \tilde{\lambda}_{jk,0} \Pr \{X_i = x | D_i = 0, Z_i = z^j\} \quad (18)$$

$$= \sum_{j=1}^j \tilde{\lambda}_{jk,1} \Pr \{X_i = x | D_i = 1, Z_i = z^j\}. \quad (19)$$

Thus, from  $\Pr \{X_i = x | D_i = d, Z_i = z^j\} = \frac{1}{p_{D,Z}(d,j)} \mathbf{E} [\mathbf{1}\{D_i = d, X_i = x, Z_i = z^j\}]$ , the use of  $\tilde{\Lambda}_d$  as the nuisance parameter and the orthogonalization procedure from Subsection 3.2 apply here. Since we have two characterizations for the same quantity  $F_{X|U}(x|u^k)$ , we can build two estimators, one from the orthogonalized moment based on (18) and another from the orthogonalized moment based on (19). Let  $\hat{F}_{X|D=0,U}(x|u)$  and  $\hat{F}_{X|D=1,U}(x|u)$  denote the estimators, respectively. Since the modified first-step NMF does not have built-in conditional independence between  $X_i$  and  $D_i$ , we can use  $\hat{F}_{X|D=0,U}(x|u)$  and  $\hat{F}_{X|D=1,U}(x|u)$  to test the distributional equivalence as a falsification test.



Fix some  $x^1 < \dots < x^{\tilde{M}}$  such that  $F_{X|U}(x^1|u) > 0$  and  $F_{X|U}(x^{\tilde{M}}|u) < 1$  for every  $u$ . Then,

$$\sqrt{n} \left( \begin{pmatrix} \hat{F}_{X|D=0,U}(x^1|u^1) \\ \hat{F}_{X|D=0,U}(x^2|u^1) \\ \vdots \\ \hat{F}_{X|D=0,U}(x^{\tilde{M}}|u^K) \end{pmatrix} - \begin{pmatrix} \hat{F}_{X|D=1,U}(x^1|u^1) \\ \hat{F}_{X|D=1,U}(x^2|u^1) \\ \vdots \\ \hat{F}_{X|D=1,U}(x^{\tilde{M}}|u^K) \end{pmatrix} \right) \xrightarrow{d} \mathcal{N}(\mathbf{0}_{\tilde{M}K}, \Sigma^1)$$

as  $n \rightarrow \infty$ , under Assumptions 1-3. The (infeasible) test statistic is

$$T_n^1 = n \begin{pmatrix} \hat{F}_{X|D=1,U}(x^1|u^1) - \hat{F}_{X|D=0,U}(x^1|u^1) \\ \vdots \\ 1 - \hat{F}_{X|D=1,U}(x^{\tilde{M}}|u^K) - (1 - \hat{F}_{X|D=0,U}(x^{\tilde{M}}|u^K)) \end{pmatrix}^{\top} \widehat{\Sigma}^1{}^{-1} \cdot \begin{pmatrix} \hat{F}_{X|D=1,U}(x^1|u^1) - \hat{F}_{X|D=0,U}(x^1|u^1) \\ \vdots \\ 1 - \hat{F}_{X|D=1,U}(x^{\tilde{M}}|u^K) - (1 - \hat{F}_{X|D=0,U}(x^{\tilde{M}}|u^K)) \end{pmatrix} \quad (20)$$

with  $\widehat{\Sigma}^1$  being the plug-in estimator for the asymptotic variance  $\Sigma^1$ .

The test statistic (20) is infeasible without further assumptions since I dropped the linear constraints (8) in the NMF step. The labeling on  $U_i$  from the treated subpopulation and the labeling on  $U_i$  from the untreated subpopulation are not connected. Since  $K$  is finite, one may compute (20) for every permutation on  $\{1, \dots, K\}$  and take the minimum, following the same spirit of minimizing over  $\tilde{g}$  in (6). The feasible test statistic based on the minimum will be asymptotically valid, but may have low power. Thus, one may alternatively assume Assumption 5 for each subpopulation, and connect the two sets of estimators  $\{\hat{F}_{X|D=0,U}(x|u)\}_{x,u}$  and  $\{\hat{F}_{X|D=1,U}(x|u)\}_{x,u}$  using the latent rank interpretation. For example, in the NMF step, one can reorder the columns of  $\Gamma_d$  and the rows of  $\Lambda_d$  so that  $\mathbf{E}[Y_i(d)|U_i = u]$  is increasing in  $u$ .

Additionally, when  $D_i$  is randomly assigned as in Remark 1, we can directly test if the distribution of  $U_i$  is identical across the two subpopulations. Since this distributional equivalence does not hold by construction in the first-step NMF, I do not modify the NMF algorithm for this test statistic. Recall the construction of the marginal distribution of  $U_i$

from (11). Find similarly that

$$\begin{pmatrix} \Pr\{U_i = u^1 | D_i = 0\} \\ \vdots \\ \Pr\{U_i = u^K | D_i = 0\} \end{pmatrix} = \Lambda_0 \begin{pmatrix} \Pr\{Z_i = z^1 | D_i = 0\} \\ \vdots \\ \Pr\{Z_i = z^K | D_i = 0\} \end{pmatrix}, \quad (21)$$

$$\begin{pmatrix} \Pr\{U_i = u^1 | D_i = 1\} \\ \vdots \\ \Pr\{U_i = u^K | D_i = 1\} \end{pmatrix} = \Lambda_1 \begin{pmatrix} \Pr\{Z_i = z^1 | D_i = 1\} \\ \vdots \\ \Pr\{Z_i = z^K | D_i = 1\} \end{pmatrix}. \quad (22)$$

When  $D_i$  is randomly assigned as in Remark 1, the LHSs of (21) and (22) should be the same. Since  $\Pr\{Z_i = z | D_i = d\} = \frac{1}{\sum_j p_{D,Z}(d,j)} \mathbf{E}[\mathbf{1}\{D_i = d, Z_i = z\}]$ , the discussion in Subsection 3.2 applies here as well. Let  $\hat{p}_{U|D=0}(k)$  denote the estimator from the orthogonalized moment based on (21) and  $\hat{p}_{U|D=1}(k)$  denote the estimator from the orthogonalized moment based on (22). Then,

$$\sqrt{n} \left( \begin{pmatrix} \hat{p}_{U|D=0}(1) \\ \vdots \\ \hat{p}_{U|D=0}(K) \end{pmatrix} - \begin{pmatrix} \hat{p}_{U|D=1}(1) \\ \vdots \\ \hat{p}_{U|D=1}(K) \end{pmatrix} \right) \xrightarrow{d} \mathcal{N}(\mathbf{0}_K, \Sigma^2)$$

as  $n \rightarrow \infty$ , under Assumptions 1-3 and Remark 1. The test statistic is

$$T_n^2 = n \begin{pmatrix} \hat{p}_{U|D=0}(1) - \hat{p}_{U|D=1}(1) \\ \vdots \\ \hat{p}_{U|D=0}(K) - \hat{p}_{U|D=1}(K) \end{pmatrix}^\top \widehat{\Sigma}^2{}^{-1} \begin{pmatrix} \hat{p}_{U|D=0}(1) - \hat{p}_{U|D=1}(1) \\ \vdots \\ \hat{p}_{U|D=0}(K) - \hat{p}_{U|D=1}(K) \end{pmatrix} \quad (23)$$

with  $\widehat{\Sigma}^2$  being the plug-in estimator for the asymptotic variance  $\Sigma^2$ .

The following theorem establishes the asymptotic validity of the two test statistics.

**Theorem 4.** *Let Assumptions 1-3 hold. Then,  $T_n^1 \xrightarrow{d} \chi^2(K\tilde{M})$  as  $n \rightarrow \infty$ . In addition, let Remark 1 hold. Then,  $T_n^2 \xrightarrow{d} \chi^2(K-1)$  as  $n \rightarrow \infty$ .*

*Proof.* This is a straightforward adaptation of the proof for Theorem 3. □

The test statistic  $T_N^1$  is constructed with  $K(\tilde{M} + 1)$ -dimensional vector but the degree of

freedom is  $K\tilde{M}$  since the conditional probabilities sum to one. Similarly, the degree of freedom for  $T_N^2$  is  $K - 1$ .

## C Choice of $K$ for a discrete $U_i$

The finite support assumption in Assumption 3 has definite merits such as significantly alleviating the computational burden compared to Assumption 4. However, the finite support assumption requires researcher to commit to a specific value of  $K$  in the estimation. In the framework of this paper,  $K$ , the number of points in the support of  $U_i$ , is equivalent to the rank of the matrix  $\mathbf{H}_0$  and  $\mathbf{H}_1$ . Thus, in this section, I introduce existing rank estimation and inference methods in the literature and connect them to the choice of  $K$  in the NMF.

Consider a  $M_X \times 2M_Z$  matrix  $\mathbf{H}_X$ :

$$\mathbf{H}_X = \begin{pmatrix} \Pr\{X_i = x^1 | (D_i, Z_i) = (0, z^1)\} & \cdots & \Pr\{X_i = x^1 | (D_i, Z_i) = (1, z^{M_Z})\} \\ \vdots & \ddots & \vdots \\ \Pr\{X_i = x^{M_X} | (D_i, Z_i) = (0, z^1)\} & \cdots & \Pr\{X_i = x^{M_X} | (D_i, Z_i) = (1, z^{M_Z})\} \end{pmatrix}.$$

Again, when  $X_i$  and  $Z_i$  are continuous random variables, we may use partitioning on  $\mathbb{R}$  to construct such  $\mathbf{H}_X$ . From Assumptions 1 and 3, the rank of  $\mathbf{H}_X$  is  $K$ .  $\mathbf{H}_X$  pools information from the treated subpopulation and the untreated subpopulation. We cannot use the column-wise concatenation of  $\mathbf{H}_0$  and  $\mathbf{H}_1$  to pool information here since the conditional distribution of  $Y_i(1)$  given  $U_i$  is likely different from that of  $Y_i(0)$  given  $U_i$  and thus rank of the concatenated matrix may be bigger than  $K$ . I estimate  $\mathbf{H}_X$  with

$$\mathbb{H}_X = \begin{pmatrix} \frac{\sum_{i=1}^n \mathbf{1}\{(D_i, X_i, Z_i) = (0, x^1, z^1)\}}{\sum_{i=1}^n \mathbf{1}\{(D_i, Z_i) = (0, z^1)\}} & \cdots & \frac{\sum_{i=1}^n \mathbf{1}\{(D_i, X_i, Z_i) = (1, x^1, z^{M_Z})\}}{\sum_{i=1}^n \mathbf{1}\{(D_i, Z_i) = (1, z^{M_Z})\}} \\ \vdots & \ddots & \vdots \\ \frac{\sum_{i=1}^n \mathbf{1}\{(D_i, X_i, Z_i) = (0, x^{M_X}, z^1)\}}{\sum_{i=1}^n \mathbf{1}\{(D_i, Z_i) = (0, z^1)\}} & \cdots & \frac{\sum_{i=1}^n \mathbf{1}\{(D_i, X_i, Z_i) = (1, x^{M_X}, z^{M_Z})\}}{\sum_{i=1}^n \mathbf{1}\{(D_i, Z_i) = (1, z^{M_Z})\}} \end{pmatrix}.$$

Firstly, to estimate  $\text{rank}(\mathbf{H}_X)$  using  $\mathbb{H}_X$ , I use the eigenvalue ratio estimator from Ahn and Horenstein [2013]. The eigenvalue ratio estimator is developed for a setup where a low-rank matrix of growing dimension is estimated, which is different from the setup of this paper; as

$n$  increases, the dimension of  $\mathbf{H}_X$  stays the same while the estimation error of  $\mathbb{H}_X$  decreases. Higgins [2025] discusses a similar setup to mine where a factor model is assumed for short  $T$  panel data and the dimension of the factor model is estimated with a reduced/collapsed data matrix whose dimension is fixed. Following Higgins [2025]’s treatment, I apply the eigenvalue ratio estimator of Ahn and Horenstein [2013] to  $\mathbb{H}_X$ . Let  $\nu_k(\mathbf{H})$  denote the  $k$ -th largest eigenvalue of  $\mathbf{H}\mathbf{H}^\top$  when  $M_X \leq 2M_Z$  and that of  $\mathbf{H}^\top\mathbf{H}$  when  $M_X \geq 2M_Z$ . The eigenvalue ratio estimator is

$$\hat{K}_{ER} = \max_{K \leq K_{\max}} \frac{\nu_K(\mathbb{H}_X)}{\mu_{K+1}(\mathbb{H}_X)}$$

with  $K_{\max} = \min\{M_X, 2M_Z\} - 1$ . Similarly, Ahn and Horenstein [2013] also proposes an estimator based on the growth rate of eigenvalues:

$$\hat{K}_{GR} = \max_{K \leq K_{\max}} \frac{\log \left( 1 + \frac{\nu_K(\mathbb{H}_X)}{\sum_{k=K+1}^{M_X} \nu_k(\mathbb{H}_X)} \right)}{\log \left( 1 + \frac{\nu_{K+1}(\mathbb{H}_X)}{\sum_{k=K+2}^{M_X} \nu_k(\mathbb{H}_X)} \right)}$$

with  $K_{\max} = \min\{M_X, 2M_Z\} - 2$ . Both estimators are consistent for true  $K$  as  $n \rightarrow \infty$ .

Secondly, to infer on  $\text{rank}(\mathbf{H}_X)$ , I use the Kleibergen-Paap (KP) rank test: Kleibergen and Paap [2006]. For a given  $K$ , the KP rank test tests the null hypothesis  $H_0 : \text{rank}(\mathbf{H}_X) = K$  against the alternative hypothesis  $H_1 : \text{rank}(\mathbf{H}_X) \geq K + 1$ . The KP rank test is developed for a general setup where a low-dimensional, low-rank matrix is estimated with estimators that are asymptotically normal. Since  $\text{vec}(\mathbb{H}_X)$  is asymptotically normal around  $\text{vec}(\mathbf{H}_X)$ , we can directly apply the KP rank test. The construction of the KP rank test statistic is notationally complex but is implemented easily with singular value decomposition. The key element of the KP test statistic is the construction of a  $(M_X - K) \times (2M_Z - K)$  matrix that is a linear transformation of  $\text{vec}(\mathbb{H}_X)$  and therefore asymptotically normal. The matrix is all zeros when  $\text{rank}(\mathbf{H}_X) = K$  and contains nonzero element when  $\text{rank}(\mathbf{H}_X) \geq K + 1$ . Kleibergen and Paap [2006] provides an explicit formula for the matrix through singular value decomposition. With the KP rank test, I test if there is significant evidence against  $K$  chosen to be used in the estimation.

## D Proofs

### D.1 Proof for Theorem 1

The proof for Theorem 1 under Assumptions 1-3 is straightforward from the discussion in the main text. Thus, I present the proof for Theorem 1 under Assumptions 1-2, 4-5 here. By repeating the spectral decomposition of Hu and Schennach [2008] twice, firstly for the treated subpopulation and secondly for the untreated subpopulation, we have a collection of  $\{f_{Y(1)|U}(\cdot|u), f_{Y(0)|U}(\cdot|u), f_{X|U}(\cdot|u)\}_{u \in \mathcal{U}}$ , without a labeling on  $u$ ; we have separated the triads of conditional densities for each value of  $u$ , but we have not labeled each triad with their respective values of  $u$  yet. To find an ordering on the infinite number of triads, let  $\tilde{U}_i = h(U_i) := M f_{Y(d)|U}(\cdot|U_i)$  from Assumption 5. Also, let  $\tilde{\mathcal{U}} = h(\mathcal{U})$ . Now, we have labeled each triad with  $\tilde{u} = h(u)$  and therefore identified  $f_{Y(1)|\tilde{U}}(\cdot|\cdot)$ ,  $f_{Y(0)|\tilde{U}}(\cdot|\cdot)$  and  $f_{X|\tilde{U}}(\cdot|\cdot)$ . Note that both Assumptions 1-2 hold with  $\tilde{U}_i$  in place of  $U_i$  since  $h$  is strictly increasing.

It remains to show that the marginal distribution of  $\tilde{U}_i$  is identified. For that, let us establish the injectivity of the integral operator based on the conditional density of  $X_i$  given  $\tilde{U}_i$ . Find that

$$\begin{aligned} f_{X|\tilde{U}}(x|\tilde{u}) &= f_{X|U}(x|h^{-1}(\tilde{u})) \\ \left[ L_{X|\tilde{U}} g \right](x) &= \int_{\tilde{\mathcal{U}}} f_{X|\tilde{U}}(x|\tilde{u}) g(\tilde{u}) d\tilde{u} = \int_{\tilde{\mathcal{U}}} f_{X|U}(x|h^{-1}(\tilde{u})) g(\tilde{u}) d\tilde{u} \\ &= \int_{\tilde{\mathcal{U}}} f_{X|U}(x|h^{-1}(\tilde{u})) g(h(h^{-1}(\tilde{u}))) d\tilde{u} \\ &= \int_{\mathcal{U}} f_{X|U}(x|u) g(h(u)) h'(u) du, \quad \text{by letting } \tilde{u} = h(u). \end{aligned}$$

From the completeness of  $f_{X|U}$ ,  $L_{X|\tilde{U}} g = 0$  implies that  $g(h(u))h'(u) = 0$  for almost everywhere on  $\mathcal{U}$ . Since  $h$  is strictly increasing,  $h'(u) > 0$ . Thus, we have  $g(\tilde{u}) = 0$  almost everywhere on  $\tilde{\mathcal{U}}$ : the completeness of  $f_{X|\tilde{U}}$  follows. Using the completeness, we identify  $f_{\tilde{U}|D=d,Z}$  from

$$f_{X|D=d,Z} = \int_{\mathbb{R}} f_{X|\tilde{U}}(x|\tilde{u}) f_{\tilde{U}|D=d,Z}(\tilde{u}|z) d\tilde{u}.$$

Since the conditional density of  $Z_i$  given  $D_i = d$  is directly observed, the marginal density of  $\tilde{U}_i$  is identified and therefore the conditional density of  $(D_i, Z_i)$  given  $\tilde{U}_i$  is also identi-

fied. Since Assumptions 1-2 hold with  $\tilde{U}_i$ , the joint density of  $(Y_i(1), Y_i(0), D_i, X_i, Z_i, \tilde{U}_i)$  is identified. Integrating out  $\tilde{U}_i$  gives us the joint density of  $(Y_i(1), Y_i(0), D_i, X_i, Z_i)$ .  $\square$

## D.2 Proof for Theorem 2

The following proof is for  $\Lambda_0$  and  $K \geq 2$ . The same proof applies to  $\Lambda_1$ . The proof is trivial when  $K = 1$ . The proofs for the lemmas are provided in the Online Appendix.

**Lemma 2.** *Let Assumptions 1-2, 3.a hold. Then,  $\left\| \Gamma_0 \Lambda_0 - \hat{\Gamma}_0 \hat{\Lambda}_0 \right\|_F^2 = O_p \left( \frac{1}{\sqrt{n}} \right)$ .*

**Lemma 3.** *Let Assumptions 1-3 hold. Then,  $\left\| \hat{\Gamma}_0 - \Gamma_0 A \right\|_F = O_p \left( \frac{1}{\sqrt{n}} \right)$  with some  $K \times K$  matrix*

$$A = \begin{cases} \Lambda_0 \left( \hat{\Lambda}_0 \right)^{-1}, & \text{if } \Gamma_0^\top \hat{\Gamma}_0 \hat{\Lambda}_0 \text{ is invertible} \\ \mathbf{I}_K, & \text{if } \Gamma_0^\top \hat{\Gamma}_0 \hat{\Lambda}_0 \text{ is not invertible} \end{cases} \quad (24)$$

with  $\mathbf{I}_K$  being the  $K \times K$  identity matrix.

**Lemma 4.** *Let Assumptions 1-3 hold. Then, the  $K \times K$  matrix  $A$  defined in (24) converges to a permutation matrix at the rate of  $n^{-\frac{1}{2}}$ , as  $n \rightarrow \infty$ .*

Lemma 2 shows that the NMF estimator retrieves the true conditional densities at the rate of  $n^{-\frac{1}{2}}$ . Lemma 3 shows that the estimator  $\hat{\Gamma}_0$  is consistent for some rotation of  $\Gamma_0$  at the rate of  $n^{-\frac{1}{2}}$ , where the rotation is denoted with the matrix  $A$ . Lemma 4 shows that the rotation matrix  $A$  converges to a permutation matrix at the rate of  $n^{-\frac{1}{2}}$ . By rearranging the columns of  $\hat{\Gamma}_0$  and the rows of  $\hat{\Lambda}_0$  so that  $A$  converges to  $\mathbf{I}_K$ ,

$$\begin{aligned} \left\| \Lambda_0 - \hat{\Lambda}_0 \right\|_F &\leq \left\| \Lambda_0 - (\Gamma_0^\top \Gamma_0)^{-1} \Gamma_0^\top \hat{\Gamma}_0 \hat{\Lambda}_0 \right\|_F + \left\| (\Gamma_0^\top \Gamma_0)^{-1} \Gamma_0^\top \hat{\Gamma}_0 \hat{\Lambda}_0 - \hat{\Lambda}_0 \right\|_F \\ &\leq \left\| (\Gamma_0^\top \Gamma_0)^{-1} \Gamma_0^\top \right\|_F \cdot \left\| \Gamma_0 \Lambda_0 - \hat{\Gamma}_0 \hat{\Lambda}_0 \right\|_F \\ &\quad + \left\| \hat{\Lambda}_0 \right\|_F \cdot \left( \left\| (\Gamma_0^\top \Gamma_0)^{-1} \Gamma_0^\top \left( \hat{\Gamma}_0 - \Gamma_0 A \right) \right\|_F + \left\| (\Gamma_0^\top \Gamma_0)^{-1} \Gamma_0^\top \Gamma_0 (A - \mathbf{I}_K) \right\|_F \right) \\ &= \left( \left\| (\Gamma_0^\top \Gamma_0)^{-1} \Gamma_0^\top \right\|_F + \left\| \hat{\Lambda}_0 \right\|_F \cdot \left\| (\Gamma_0^\top \Gamma_0)^{-1} \Gamma_0^\top \right\|_F + \left\| \hat{\Lambda}_0 \right\|_F \right) \cdot O_p \left( \frac{1}{\sqrt{n}} \right). \end{aligned}$$

$\square$