Distributional treatment effect with latent rank invariance

Myungkou Shin*

December 18, 2025

Abstract

Treatment effect heterogeneity is of great concern when evaluating policy impacts, such as assessing the proportion of individuals who are better off under the treatment. However, existing analyses have mostly been limited to summary measures such as an average treatment effect, due to the fundamental limitation that we cannot simultaneously observe both a treated potential outcome and an untreated potential outcome for a given unit. In this paper, I propose a conditional independence framework to circumvent this limitation and estimate moment-identified distributional treatment effect (DTE) parameters, such as marginal distribution of treatment effect. The key identifying assumption is that the two potential outcomes are conditionally independent given a latent variable, which is informed by two proxy variables. Interpreting this latent variable as underlying individual-level heterogeneity, 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 nonnegative matrix factorization and plug-in GMM. Using Neyman orthogonality, I establish asymptotic normality of the estimator, enabling inference for DTE parameters.

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

JEL classification codes: C13

*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. Consider a potential outcome setup with a binary treatment:

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

$Y(1)$ is a treated potential outcome, $Y(0)$ is an untreated potential outcome, and $D \in \{0, 1\}$ is a binary treatment variable. In this setup, a researcher may be interested in distributional aspects of the treatment: 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? These questions correspond to testing $H_0 : \Pr\{Y(1) - Y(0) \leq 0\} = 0$ and estimating $\mathrm{Var}\big(Y(1) - Y(0)\big)$ and $\Pr\{Y(1) - Y(0) \geq c\}$. Note that these quantities, $\Pr\{Y(1) - Y(0) \leq 0\}, \Pr\{Y(1) - Y(0) \geq c\}$ and $\mathrm{Var}\big(Y(1) - Y(0)\big)$, 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 a class of distributional treatment effect (DTE) parameter θ, which is identified by a moment function defined over the two potential outcomes:

$$\mathbf{E}\left[m\big(Y(1), Y(0); \theta\big)\right] = 0. \tag{1}$$

The DTE parameter θ is identified if the joint distribution of the two potential outcomes

is identified.[1] Examples of θ include distribution of treatment effect, variance of treatment effect, and many more.

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], assuming that there exists a latent variable which captures the dependence between the two potential outcomes.

Consider a simple 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 ε^1 and ε^0:

$$Y(1) = \mu^1(U) + \varepsilon^1, \tag{2}$$

$$Y(0) = \mu^0(U) + \varepsilon^0. \tag{3}$$

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, μ^1 instead of μ^0, and there are two separate random noises drawn for each production function, ε^0 and ε^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:

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

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

[1] 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 use the expression 'distributional' only when the object involves the joint distribution of the two potential outcomes.

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 (2)-(3), the proxy variables X, Z will shift $\mu^d(U)$ independently of $\left(\varepsilon^1, \varepsilon^0\right)$, allowing us to decompose the variation of $Y(d)$ into the variation of U and the variation of ε^d. Additionally, to find a labeling on U, I assume that the conditional distribution of $Y(d)$ given U orders U: latent rank invariance.

In implementation, 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.[2] Under the finite support assumption, the conditional independence assumption can be interpreted as finite mixture whose properties are well-studied in the literature; the nonnegative matrix factorization serves as an estimator for the nonparametric finite mixture model. In the second step, I show that the infeasible moment function in (1) can be modified to a feasible moment function, with the outcome of the nonnegative matrix factorization algorithm as nuisance parameters. Asymptotic normality of the two-step plug-in DTE estimator is established, taking advantage of the Neyman orthogonality.

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. 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.

This paper makes a contribution to the distributional treatment effect literature: see Bedoya et al. [2018] for an overview. As a leading example of the DTE parameter, I point identify and estimate the marginal distribution of treatment effect, which contrasts the

[2]Though I assume that U is finitely discrete in the estimation, the identification result does not require such a restriction and I develop an alternative estimation method based on sieve maximum likelihood for a setup with continuous U, in the Online Appendix.

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]; both estimators build on an unconditional independence assumption, arguably more restrictive than this paper's conditional independence framework.

This paper also contributes to the nonclassical measurement error/proximal inference literature and the nonparametric finite mixture literature: Hu [2008], Hu and Schennach [2008], Henry et al. [2014], Bonhomme et al. [2016] and more. Unlike existing estimators based on the eigenvalue decomposition, the estimation strategy of this paper has guarantee that the estimated mixture probabilities are indeed nonnegative. Simulations show significant gain in the finite-sample performance, thanks to the additional regularization. Thus, the new estimator offers a promising alternative for nonparametric finite mixture estimation, especially when the target parameter is sensitive to the quality of mixture estimation. In addition, while establishing asymptotic normality of the DTE estimator, I provide an orthogonalization procedure applicable to GMM models where the mixture weights are used as nuisance parameters.

The rest of the paper is organized as follows. Section 2 identifies the joint distribution of the two potential outcomes and thus DTE parameters identified by a moment condition (1). In addition, Section 2 derives a testable implication of the framework. Section 3 explains the estimation method for moment-identified DTE parameters and develops asymptotic theory for the estimator. 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 of interest, D_i is a binary treatment and X_i, Z_i are two proxies for

individual-level heterogeneity. The outcome Y_i is constructed with two potential outcomes.

$$Y_i = D_i \cdot Y_i(1) + (1 - D_i) \cdot Y_i(0). \tag{4}$$

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 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: $(Y_i(1), Y_i(0), D_i, X_i, Z_i, U_i) \overset{iid}{\sim} \mathcal{F}$.

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)* $\big(Y_i(1), Y_i(0), X_i\big) \perp\!\!\!\perp \big(D_i, Z_i\big) \mid U_i$.

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 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 randomly assigned treatments, limiting my attention to randomized controlled trials.

Remark 1. A sufficient condition for Assumption 1 is

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

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. 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 DTE 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 independence 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, 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*, a binary treatment* D_i*, and two proxy variables* X_i *and* Z_i *which measure an innate ability* U_i*.*

$Y_i = D_i \cdot Y_i(1) + (1 - D_i) \cdot Y_i(0)$ *and*

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

$\varepsilon_i^0, \varepsilon_i^1, \varepsilon_i^X$ *and* ε_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 provided parenting advice 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 baseline measures of cognitive ability—such as the number of words the child can say and the number of complex phrases the child can say—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.[3] Compared to Attanasio et al. [2020], the only additional assumption that I impose here is that measurement errors ε_i^0 and ε_i^1 are conditionally independent of each other given the innate ability 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.[4]

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

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

[3]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.

[4]The idea of assuming a hidden Markov model for a panel dataset and using past and future outcomes 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.

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

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

The latent state process $\{V_{it}\}_{t=1}^{3}$ *is first-order Markovian given* D_i*. Also, the time-by-treatment-status shocks* $\varepsilon_{i1}^0, \varepsilon_{i2}^0, \varepsilon_{i2}^1, \varepsilon_{i3}^0, \varepsilon_{i3}^1$ *and* $\left(\{V_{it}\}_{t=1}^{3}, D_i\right)$ *are all mutually independent. When treatment* D_i *is randomly assigned at time* $t = 2$*, i.e.* $\{V_{it}\}_{t=1}^{2} \perp\!\!\!\perp D_i$*, Assumptions 1-2 are satisfied with* $Y_i = Y_{i2}, X_i = Y_{i1}, Z_i = Y_{i3}$ *and* $U_i = V_{i2}$*.*

In this example, I extend the hidden Markov model to a potential outcome setup. Firstly, I assume that each of the two potential outcomes follows a hidden Markov model. Then, I assume that the latent state process $\{V_{it}\}_{t=1}^{3}$ is shared across the two potential outcomes generating processes and first-order Markovian. 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 individual medical spending. The program included in-person classes on physical fitness and healthy workplace habits. The medical spending information was collected before, during, and after the treatment. Thus, the pretreatment and the follow-up medical spending can serve as the proxy variables X_i and Z_i and the concurrent medical spending as the outcome variable Y_i. In this context, the latent state $\{V_{it}\}_{t=1}^{3}$ can be thought of as a process of underlying health status.

The common feature shared 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 possible mechanisms of treatment effect: a change in production function itself and a change in the amount of inputs.[5] However, they abstract away from a comprehensive counterfactual

[5]This relates to the two different independence assumptions in the literature: $Y_i(1) \perp\!\!\!\perp Y_i(0)$ and $Y_i(0) \perp\!\!\!\perp \left(Y_i(1) - Y_i(0)\right)$. The latter approach would be plausible if the treatment mechanism works in an "input-changing" manner. That is, a treated potential outcome goes through the same outcome generating process

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 and therefore 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. When the parenting guidance systemically changed children's cognitive ability development process, Attanasio et al. [2020] would fit this framework. Similarly, if the wellness program systemically changed participants' health-related behaviors, Jones et al. [2019] would also fit the framework.[6]

The key assumption that the regime-specific error terms ε_i^0 and ε_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

$$\big(Y_i(1), Y_i(0), X_i, C_i, U_i\big) \perp\!\!\!\perp D_i \text{ and } \big(Y_i(1), Y_i(0), X_i\big) \perp\!\!\!\perp Z_i \mid (D_i, C_i, U_i)$$
$$Y_i(1), Y_i(0) \text{ and } X_i \text{ are all mutually independent given } \big(C_i, U_i\big).$$

In this setup, the scalar U_i only needs to explain remaining heterogeneity among units with

as an untreated potential outcome, but with an additional source of heterogeneity, which is independent of the existing source of heterogeneity. Suppose $V_i \perp\!\!\!\perp \varepsilon_i \mid U_i$ and

$$Y_i(d) = \alpha + \mu^0 U_i + d \cdot \mu^1 V_i + \varepsilon_i.$$

Then, $Y_i(0) \perp\!\!\!\perp \big(Y_i(1) - Y_i(0)\big) \mid U_i$. Here, V_i denotes the new source of individual-level heterogeneity, which contributes to the outcome only when treated. An example of such an empirical context is where the treatment provides a new infrastructure, but is not designed to affect individual behavior.

[6]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 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 $Y_i(0)$

In this subsection, I present the identification argument. A key step in identification is the diagonalization of observable conditional densities, drawn from Hu [2008], Hu and Schennach [2008]. For illustration purposes, suppose that Y_i, X_i, Z_i, U_i are discrete: $Y_i \in \{y^1, \cdots, y^{M_Y}\}, X_i \in \{x^1, \cdots, x^{M_X}\}, Z_i \in \{z^1, \cdots, z^{M_Z}\}$ and $U_i \in \{u^1, \cdots, u^K\}$. 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 $d = 0, 1$,

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

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

$$\mathbf{H}_d = \Gamma_d \cdot \Lambda_d \tag{5}$$

where

$$\Gamma_d = \begin{pmatrix} \Pr\left\{(Y_i(d), X_i) = (y^1, x^1) \,|U_i = u^1\right\} & \cdots & \Pr\left\{(Y_i(d), X_i) = (y^1, x^1) \,|U_i = u^K\right\} \\ \vdots & \ddots & \vdots \\ \Pr\left\{(Y_i(d), X_i) = \left(y^{M_Y}, x^{M_X}\right) |U_i = u^1\right\} & \cdots & \Pr\left\{(Y_i(d), X_i) = \left(y^{M_Y}, x^{M_X}\right) |U_i = u^K\right\} \end{pmatrix},$$

$$\Lambda_d = \begin{pmatrix} \Pr\left\{U_i = u^1 \middle| (D_i, Z_i) = (d, z^1)\right\} & \cdots & \Pr\left\{U_i = u^1 \middle| (D_i, Z_i) = \left(d, z^{M_Z}\right)\right\} \\ \vdots & \ddots & \vdots \\ \Pr\left\{U_i = u^K \middle| (D_i, Z_i) = (d, z^1)\right\} & \cdots & \Pr\left\{U_i = u^K \middle| (D_i, Z_i) = \left(d, z^{M_Z}\right)\right\} \end{pmatrix}. \tag{6}$$

The discreteness of Y_i, X_i, Z_i is nonbinding; we can use partitioning on $\mathbb{R}$ when they are continuous.[7]

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 Λ_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 Λ_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 representations across the subpopulations.

Under Assumption 2, Γ_0 and Γ_1 can be further decomposed. Let

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

Also, let $\Gamma_{d,k}$ denote the k-th column of Γ_d and similarly for Γ_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 $k = 1, \cdots, K$.

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 Γ_d is identified by collecting Γ_X and $\Delta(y)$

[7] Consider partitions on $\mathbb{R}$ such that

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

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

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

for each $d = 0, 1$. Γ_d and Λ_d are similarly constructed with partitioned Y_i, X_i and Z_i.

from the eigenvalue decompositions of

$$\mathbf{H}_d(y)\left(\sum_{y'}\mathbf{H}_d(y')\right)^{-1} = \Gamma_X \cdot \Delta(y) \cdot (\Gamma_X)^{-1} \tag{7}$$

across y, when Γ_X and Λ_d are full rank and no two columns of $\Gamma_{Y(d)}$ are identical.[8] Then, Λ_d is identified from the full rank of Γ_X. Thus Γ_d, Λ_d are identified from $\mathbf{H}_d$, for $d = 0, 1$.

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 Γ_1 and Γ_0 in the discretization. The last distribution is a function of Λ_1, Λ_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; Γ_X appears in both of the decompositions and thus we can connect the labeling on the latent variable U_i across the two subpopulations using Γ_X.

Assumption 3 assumes a discrete U_i, making the decomposition in (5) exact, and formally states the full rank condition and the no repeated eigenvalue condition.

Assumption 3.

a. *(finitely discrete U_i)* $\mathcal{U} = \{u^1, \cdots, u^K\}$.

b. *(full rank)* Λ_0, Λ_1 *and* Γ_X *have rank* K.

c. *(no repeated eigenvalue) For any* $k \neq k'$, *there exist some* $y, y' \in \{y^1, \cdots, y^{M_Y}\}$ *such that*

$$\Pr\left\{Y_i(0) = y \middle| U_i = u^k\right\} \neq \Pr\left\{Y_i(0) = y \middle| U_i = u^{k'}\right\},$$
$$\Pr\left\{Y_i(1) = y' \middle| U_i = u^k\right\} \neq \Pr\left\{Y_i(1) = y' \middle| U_i = u^{k'}\right\}.$$

[8] When Γ_X or Λ_d is not a square matrix, focusing on K linearly independent columns of Λ_d and using a pseudoinverse of Γ_X derives 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 $d = 0, 1$,

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

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 $u \neq u'$,*

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

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*

$$h(u) = M f_{Y(1)|U}(\cdot|u) \qquad or \qquad h(u) = M f_{Y(0)|U}(\cdot|u)$$

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 $Y_i(d) \mid U_i = u \sim \mathcal{N}(u^2 + d, \sigma^2)$. Neither $f_{Y(1)|U}$ nor $f_{Y(0)|U}$ helps us find an ordering between $f_{X|U}(\cdot|u)$ and $f_{X|U}(\cdot| - u)$.

As an example, let us revisit the simple example (2)-(3) and suppose that $\mathbf{E}[\varepsilon_i^1|U_i] = 0$ and $\mu^1(u)$ is strictly increasing in u. Then, Assumption 5 holds with $h(u) = \mu^1(u)$. In this example, U_i represents the rank of unit i in terms of the systemic part of their treated potential outcome $\mu^1(U_i)$, which is latent, and also pins down the rank in terms of the systemic part of their untreated potential outcome $\mu^0(U_i)$ as well. Thus, the latent variable U_i can be thought of as a 'latent rank,' a single latent variable that determines the ranks of the systemic parts of both potential outcomes, therefore inducing a deterministic relationship between the two systemic parts $\mu^1(U_i)$ and $\mu^0(U_i)$. Importantly, this does not imply that the two potential outcomes $Y_i(1)$ and $Y_i(0)$ be deterministically connected in the same manner. In this sense, I refer to the conditional independence framework of this paper as 'latent rank invariance,' which can be viewed as 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].

Theorem 1 formally states identification.

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

2.2 Distributional treatment effect parameters

It directly follows from Theorem 1 that any functional of the joint distribution of $Y_i(1)$ and $Y_i(0)$ is identified. In implementation, I focus on DTE parameters that are moment-identified and develop an estimation strategy based on two-step plug-in GMM. Thus, in this subsection, I present several examples of moment-identified DTE parameters: with $F_{Y(1)-Y(0)}$ denoting the marginal distribution of $Y_i(1) - Y_i(0)$,

- Conditional ATE given $Y_i(0)$: $\theta(\mathcal{Y}) = \mathbf{E}[Y_i(1) - Y_i(0)|Y_i(0) \in \mathcal{Y}]$.
- Variance-adjusted ATE: $\theta = \mathbf{E}[Y_i(1) - Y_i(0)] + \frac{\gamma}{2}\mathrm{Var}\big(Y_i(1) - Y_i(0)\big)$.
- Share of winners: $\theta = 1 - F_{Y(1)-Y(0)}(0)$.
- α-th quantile of treatment effect: $\theta(\alpha) = F^{-1}_{Y(1)-Y(0)}(\alpha)$.

For interpretation, suppose that a larger outcome is deemed more desirable. The conditional ATE discusses treatment effect heterogeneity in terms of the untreated potential outcome. This parameter has important policy implications in terms of fairness: by conditioning on the untreated outcome, we can determine whether the overall treatment effect is driven by gains on units with low baseline outcomes or those with high baseline outcomes.

The variance-adjusted ATE aggregates the treatment effect while allowing for inequality aversion. The risk aversion in expected utility maximization directly translates to the inequality aversion in social welfare. Thus, the quadratic approximation in utility maximization, e.g., [Levy and Markowitz, 1979], applies to the social welfare context: see Epstein and Segal [1992] for more. In the approximation, $\gamma \geq 0$ is the inequality aversion parameter:

$$\mathbf{E}\left[u(Y_i(1) - Y_i(0))\right] \approx u'(0) \cdot \mathbf{E}\left[(Y_i(1) - Y_i(0))\right] + \frac{1}{2}u''(0) \cdot \mathbf{E}\left[(Y_i(1) - Y_i(0))^2\right].$$

The share of winners connects to the marginal distribution of treatment effect and measures the proportion of units that are better off under the treatment. When the share of winners is one, the treatment is Pareto-improving, benefiting a treated unit with probability one. When the share of winners is bigger than 0.5, the median treatment effect is positive,

with the probability of the treatment benefiting a treated unit bigger than 0.5.[9]

On the flip side of the distribution function, the α-th quantile of treatment effect marks the threshold for the worst $100 \cdot \alpha\%$ of treatment effects. This parameter gives us a probabilistic guarantee that the probability of the treatment harming a unit by a margin greater than this threshold does not exceed α. This notion relates to the recent developments on probabilistic guarantee in optimal policy learning literature: Reeve et al. [2023].

2.3 Testable implication

When Assumption 5 extends to both $Mf_{Y(1)|U}(\cdot|u)$ and $Mf_{Y(0)|U}(\cdot|u)$, we have a testable implication of Assumptions 1-2 and 4-5, from over-identification. With the extension, we can find a labeling on $\{f_{X|U}(\cdot|u)\}_u$ within each subpopulation; the conditional densities $\left(f_{Y(1)|U}, f_{X|U}, f_{U|D=1,Z}\right)$ are identified from the treated subpopulation and the conditional densities $\left(f_{Y(0)|U}, f_{X|U}, f_{U|D=0,Z}\right)$ 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, under a common support assumption for $U_i \mid D_i = 0$ and $U_i \mid D_i = 1$, Assumption 1 imposes that

$$\min_{\tilde{g}:\text{monotone}} \mathbf{E}\left[\int_{\mathbb{R}} \left(f_{X|D=1,U}(x|U_i) - f_{X|D=0,U}(x|\tilde{g}(U_i))\right)^2 dx \middle| D_i = 1\right] = 0. \tag{8}$$

In (8), 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.[10] A test that uses (8) as a null can be used as a falsification test on the framework proposed in this paper.

What does a test on the null (8) 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

[9] The median treatment effect $F^{-1}_{Y(1)-Y(0)}(0.5)$ is different from the quantile treatment effect at median $F^{-1}_{Y(1)}(0.5) - F^{-1}_{Y(0)}(0.5)$.

[10] In the case of discrete U_i, Assumption 5 was not used in the identification. Based on the same reasoning, we get a testable implication without extending Assumption 5 to both $Mf_{Y(1)|U}(\cdot|u)$ and $Mf_{Y(0)|U}(\cdot|u)$:

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

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 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, (8) simplifies to

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

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

$$\int_{\mathbb{R}} \left(f_{U|D=1}(u) - f_{U|D=0}(u)\right)^2 du = 0. \tag{9}$$

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

3 Implementation

In implementation, 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 estimation of finite-dimensional objects Λ_0 and Λ_1. Lastly, with discrete U_i, the DTE parameters, which are nonlinear functionals of the densities $\left(f_{Y(1)|U}, f_{Y(0)|U}, f_U\right)$, becomes linear in quantities identified with quadratic moments.

The linearity induced from the 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 the appendix Section 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 problem, to estimate Λ_0 and Λ_1. Secondly, I estimate DTE parameters with a plug-in GMM estimation where Λ_0 and Λ_1 are nuisance parameters.

3.1 Nonnegative matrix factorization

To estimate the mixture weight matrices Λ_0 and Λ_1 from (6), I formulate a nonnegative matrix factorization (NMF) problem based on (5). 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

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

Each column of $\mathbb{H}_d$ is an empirical conditional distribution function of (Y_i, X_i) given $\left(D_i = d, Z_i\right)$. 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 Λ_0 and Λ_1 by solving a NMF problem:

with ι_x denoting a x-dimensional column vector of ones,

$$\min_{\Lambda_0,\Lambda_1,\Gamma_0,\Gamma_1} \|\mathbb{H}_0 - \Gamma_0\Lambda_0\|_F{}^2 + \|\mathbb{H}_1 - \Gamma_1\Lambda_1\|_F{}^2 \tag{10}$$

subject to linear constraints

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

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

$$\left(\sum_{m=1}^{M_Y} \Pr\left\{\big(Y_i(0), X_i\big) = (y^m, x)\,|U_i = u^k\right\}\right) = \left(\sum_{m=1}^{M_Y} \Pr\left\{\big(Y_i(1), X_i\big) = (y^m, x)\,|U_i = u^k\right\}\right) \tag{11}$$

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

$$\begin{aligned}
&\Pr\left\{(Y_i(d), X_i) = (y, x)|U_i = u^k\right\} \\
&= \left(\sum_{m=1}^{M_X} \Pr\left\{\big(Y_i(d), X_i\big) = (y, x^m)\,|U_i = u^k\right\}\right) \cdot \left(\sum_{m=1}^{M_Y} \Pr\left\{(Y_i(d), X_i) = \big(y^m, x\big)|U_i = u^k\right\}\right).
\end{aligned} \tag{12}$$

The objective (10) comes from the decomposition (5). The linear constraints impose that the columns of $\Gamma_0, \Gamma_1, \Lambda_0$ and Λ_1 are well-defined distributions and that $X_i \perp\!\!\!\perp D_i \;\big|\; U_i$ (Assumption 1). The quadratic constraints impose that $Y_i(d) \perp\!\!\!\perp X_i \;\big|\; U_i$ (Assumption 2). Theorem 1 says that the solution to the optimization problem is unique up to some permutation on the columns of Γ_0, Γ_1 and the rows of Λ_0, Λ_1, when $\mathbb{H}_0, \mathbb{H}_1$ are sufficiently close to $\mathbf{H}_0, \mathbf{H}_1$.

Note that the objective function in (10) is quadratic when we fix either (Λ_0, Λ_1) or (Γ_0, Γ_1). Moreover, recall that Γ_0 and Γ_1 can be further decomposed into three matrices $\Gamma_X, \Gamma_{Y(0)}, \Gamma_{Y(1)}$. Let $\Gamma(\cdot,\cdot)$ denote how Γ_X and $\Gamma_{Y(d)}$ recover Γ_d, using the column-wise Kronecker products: $\Gamma_d = \Gamma\Big(\Gamma_X, \Gamma_{Y(d)}\Big)$. The linear constraints (11) and the quadratic constraints (12) 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 Λ)* Given $\left(\Gamma_0^{(s)}, \Gamma_1^{(s)}\right)$, solve the following quadratic program:

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

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

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

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

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

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

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

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

5. Repeat **2-4** until convergence.

Each step of the iteration is a quadratic programming with a positive-(semi)definite Hessian matrix and linear constraints, which can be solved with a built-in optimization tool in most statistical software. The stepwise optimization assures a convergence to a local minimum. To find the global minimum, I consider various initial values $\left(\Gamma_0^{(0)}, \Gamma_1^{(0)}\right)$.[11]

Let $\widehat{\Lambda}_0$, $\widehat{\Lambda}_1$, $\widehat{\Gamma}_0$ and $\widehat{\Gamma}_1$ denote the solution to the minimization problem. Note that when Y_i is discrete, the estimates $\widehat{\Gamma}_0$ and $\widehat{\Gamma}_1$ directly estimate the conditional distribution of $Y_i(1)$

[11]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 ${\mathbb{H}_d}^\intercal\mathbb{H}_d$ as an initial value.

and $Y_i(0)$ given U_i. When Y_i is continuous and therefore partitioning was used in constructing $\mathbf{H}_0$ and $\mathbf{H}_1$, I use $\widehat{\Lambda}_0$ and $\widehat{\Lambda}_1$ to estimate the distribution of $Y_i(1)$ and $Y_i(0)$ given U_i.

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

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

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

as $n \to \infty$.

Proof. See Appendix. □

Theorem 2 is the key result in the asymptotic theory for the DTE estimator. In the proof, I also show that $\widehat{\Gamma}_0, \widehat{\Gamma}_1$ are $\sqrt{n}$-consistent estimators of Γ_0, Γ_1, up to some permutation.

The constraints in the NMF problem impose that the columns of Γ_d and Λ_d are probability mass functions. This is in contrast to an estimator based on eigenvalue decomposition (7), as suggested in Hu [2008]. In the eigenvalue decomposition-based estimation, there is no guarantee that the eigenvectors are all of the same sign; nonnegativity constraint is not imposed. Thanks to the additional regularization, the NMF estimator outperforms the eigenvalue decomposition estimator in finite sample: Table 3 of Section 4. In this sense, though computationally more demanding, the NMF estimator is an useful alternative to practitioners when their parameter of interest crucially depends on the estimation quality of a nonparametric finite mixture model.

3.2 Distributional treatment effect estimator

For the distributional treatment effect (DTE) estimation, adopt the following notation:

$$\begin{aligned}
\tilde{\Lambda}_d &= (\Lambda_d)^{-1} \quad \text{for } d = 0, 1 \\
p_{D,U}(d,k) &= \Pr\left\{D_i = d, U_i = u^k\right\} \quad \text{for } d = 0, 1 \text{ and } k = 1, \cdots, K \\
p_{D,Z}(d,j) &= \Pr\left\{D_i = d, Z_i = z^j\right\} \quad \text{for } d = 0, 1 \text{ and } j = 1, \cdots, K.
\end{aligned}$$

$p_{D,U}$ is the joint distribution of D_i and U_i and $p_{D,Z}$ is the joint distribution of D_i and Z_i. In addition, let $\tilde{\lambda}$ denote the vectorization of $\big(\tilde{\Lambda}_0, \tilde{\Lambda}_1\big)$ and p denote the collection of $\{p_{D,U}(0,k)\}_{k=1}^K$, $\{p_{D,U}(1,k)\}_{k=1}^K$, $\{p_{D,Z}(0,j)\}_{j=1}^K$ and $\{p_{D,Z}(1,j)\}_{j=1}^K$.

In this paper, I consider the class of DTE parameters identified by a moment function m defined on $\big(Y_i(1), Y_i(0), D_i, X_i\big)$: [12]

$$\mathbf{E}\left[m\big(Y_i(1), Y_i(0), D_i, X_i; \theta\big)\right] = 0. \tag{13}$$

From Theorem 1, we know that θ is identified under Assumptions 1-3. The following proposition further shows that the parameter of interest θ is identified with a feasible quadratic moment condition, with $\tilde{\lambda}, p$ as nuisance parameters. Let $W_i = (Y_i, D_i, X_i, Z_i)$.

Proposition 1. *Assumptions 1-3 hold. Suppose that a parameter of interest θ is identified by an infeasible moment condition* (13)*. Then, there is a moment function $\tilde{m}$ such that θ is identified by a feasible quadratic moment condition*

$$\mathbf{E}\left[\tilde{m}\big(W_i, W_{i'}; \theta, \tilde{\lambda}, p\big)\right] = 0 \tag{14}$$

where $\tilde{\lambda}, p$ are nuisance parameters and $i \neq i'$.

Proof. See Appendix. □

For example, the marginal distribution of treatment effect $\theta = F_{Y(1)-Y(0)}(\delta)$ for a fixed δ is identified with

$$\tilde{m}\big(W_i, W_{i'}; \theta, \tilde{\lambda}, p\big) = \sum_{j=1}^{K}\sum_{j'=1}^{K} \frac{\sum_{k=1}^K \big(p_{D,U}(0,k) + p_{D,U}(1,k)\big)\tilde{\lambda}_{jk,0}\tilde{\lambda}_{jk,1}}{p_{D,Z}(0,j)p_{D,z}(1,j')} \cdot \mathbf{1}\{Y_{i'} \leq Y_i + \delta, D_i = 0, Z_i = z^j, D_{i'} = 1, Z_{i'} = z^{j'}\} - \theta. \tag{15}$$

In Sections 4-5, the marginal distribution of treatment effect is used as a working example of

[12] Note that the moment condition (13) does not use Z_i. This is because the nuisance parameter $\tilde{\lambda}$ helps us with substituting for the conditional density of $\big(Y_i(1), Y_i(0), X_i\big)$ given U_i, but not for the conditional density of $\big(D_i, Z_i\big)$ given U_i. This does not contradict the general identification result of Theorem 1. We can extend Proposition 1 so that m in (13) also takes Z_i as an input, by using left inverses of Γ_0 and Γ_1 as additional nuisance parameters.

the DTE parameter. We can construct feasible moment functions for the DTE parameters discussed in Subsection 2.2 in a similar manner.

To estimate θ with the feasible moment condition (14), I estimate the nuisance parameters as follows: with the first-step NMF estimator $\widehat{\Lambda}_0$ and $\widehat{\Lambda}_1$,

$$\widehat{\tilde{\Lambda}}_d = \left(\widehat{\Lambda}_d\right)^{-1}, \tag{16}$$

$$\{\hat{p}_{D,U}(d,k)\}_{k=1}^K = \widehat{\Lambda}_d \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{D_i = d, Z_i = z^1\} \\ \vdots \\ \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{D_i = d, Z_i = z^K\} \end{pmatrix}, \tag{17}$$

$$\hat{p}_{D,Z}(d,j) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{D_i = d, Z_i = z^j\}. \tag{18}$$

In general, the estimation error in the nuisance parameter has first-order impact on the feasible moment $\tilde{m}$. Thus, to account for the first-step estimation error, I orthogonalize the moment function.

Even though the NMF estimators $\left(\widehat{\Lambda}_0, \widehat{\Lambda}_1\right)$ and the induced estimators $\left(\widehat{\tilde{\Lambda}}_0, \widehat{\tilde{\Lambda}}_1\right)$ are complex nonlinear functions of the data matrices $\mathbb{H}_0$ and $\mathbb{H}_1$, $\left(\tilde{\Lambda}_0, \tilde{\Lambda}_1\right)$ satisfy the following equations at their true values: for all (y, d, x, k),

$$\begin{aligned} &\sum_{j=1}^K \tilde{\lambda}_{jk,d} \Pr\left\{Y_i = y, X_i = x | D_i = d, Z_i = z^j\right\} \\ &\quad = \sum_{j=1}^K \tilde{\lambda}_{jk,d} \Pr\left\{Y_i = y | D_i = d, Z_i = z^j\right\} \cdot \sum_{j'=1}^K \tilde{\lambda}_{j'k,d} \Pr\left\{X_i = x | D_i = d, Z_i = z^{j'}\right\} \end{aligned} \tag{19}$$

$$\Pr\left\{D_i = d, X_i = x\right\} = \sum_{k=1}^K p_{D,U}(d,k) \sum_{j=1}^K \tilde{\lambda}_{jk,d} \Pr\left\{X_i = x | D_i = d, Z_i = z^j\right\}. \tag{20}$$

Equation (19) corresponds to (12) that $Y_i(d) \perp\!\!\!\perp X_i \mid U_i$ and Equation (20) corresponds to the law of iterated expectation that $\Pr\{D_i = d, X_i = x\} = \sum_{k=1}^K p_{D,U}(d,k) \Pr\{X_i = x | U_i = u^k\}$. Given $\{p_{D,Z}(d,j)\}_{d,j}$, Equation (19) can be written as a quadratic moment condition and

Equation (20) as a linear moment condition. In addition,

$$p_{D,Z}(d,j) = \Pr\left\{D_i = d, Z_i = z^j\right\} \tag{21}$$

gives us a moment condition for $p_{D,Z}(d,j)$. There are $2MK$ moments in the form of (19), $2M_X$ in the form of (20), and $2K$ in the form of (21). By collecting score functions for these additional moments across different values of (y,d,x,j,k), we get $\phi(W_i, W_{i'}; \tilde{\lambda}, p)$.

To use ϕ in the orthogonalization, I show that the Jacobian matrix of ϕ has full rank.

Lemma 1. *Assumptions 1-3 hold. Then, the following Jacobian matrix is full row rank:*

$$\begin{pmatrix} \mathbf{E}\left[\frac{\partial}{\partial \tilde{\lambda}}\phi(W_i, W_{i'}; \tilde{\lambda}, p)\right] \\ \mathbf{E}\left[\frac{\partial}{\partial p}\phi(W_i, W_{i'}; \tilde{\lambda}, p)\right] \end{pmatrix} \in \mathbb{R}^{(2K^2+2K+2K)\times(2MK+2M_X+2K)}.$$

The proof is provided in the Online Appendix. An important implication from the proof of Lemma 1 is that the Jacobian matrix in Lemma 1 is not full rank when Λ_0, Λ_1 have more than K columns; there are too many nuisance parameters. Thus, the support of the discretized Z_i has to be exactly K. Recall that Assumption 3.b imposes that we have at least K points in the support of X_i and Z_i. Now, Lemma 1 imposes that whenever we have more than K points in the support of Z_i, we need to use a partition to reduce it to K.

Then, with an additional nuisance parameter

$$\mu = \begin{pmatrix} \mathbf{E}\left[\frac{\partial}{\partial \tilde{\lambda}}\phi(W_i, W_{i'}; \tilde{\lambda}, p)\right] \\ \mathbf{E}\left[\frac{\partial}{\partial p}\phi(W_i, W_{i'}; \tilde{\lambda}, p)\right] \end{pmatrix}^{+} \begin{pmatrix} \mathbf{E}\left[\frac{\partial}{\partial \tilde{\lambda}}m(W_i, W_{i'}; \tilde{\lambda}, p)\right] \\ \mathbf{E}\left[\frac{\partial}{\partial p}m(W_i, W_{i'}; \tilde{\lambda}, p)\right] \end{pmatrix} \in \mathbb{R}^{2MK+2M_X+2K},$$

the score

$$\psi(W_i, W_{i'}; \theta, \tilde{\lambda}, p, \mu) = \tilde{m}(W_i, W_{i'}; \theta, \tilde{\lambda}, p) - \mu^{\intercal}\phi(W_i, W_{i'}; \tilde{\lambda}, p)$$

satisfies Neyman orthogonality. The orthogonalization procedure applies to any GMM estimation that uses the mixture component weight as nuisance parameters. Thus, Lemma 1 has broader applicability in deriving an asymptotic distribution for an estimator based on a nonparametric finite mixture model.

Let $\widehat{\tilde{\lambda}}$ and $\hat{p}$ denote the (vectorized) nuisance parameter estimators from (16)-(18) and $\hat{\mu}$ denote the plug-in, sample analogue estimator of μ. I estimate θ with

$$\binom{n}{2}^{-1} \sum_{i<i'} \frac{1}{2} \left(\psi \left(W_i, W_{i'}; \hat{\theta}, \widehat{\tilde{\lambda}}, \hat{p}, \hat{\mu} \right) + \psi \left(W_{i'}, W_i; \hat{\theta}, \widehat{\tilde{\lambda}}, \hat{p}, \hat{\mu} \right) \right) = 0.$$

Theorem 3 establishes the asymptotic normality of the DTE estimator.

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

$$\sqrt{n} \left(\hat{\theta} - \theta \right) \xrightarrow{d} \mathcal{N}(0, \Sigma)$$

as $n \to \infty$ with some consistently estimable Σ.

Proof. From Theorem 2, $\widehat{\tilde{\lambda}}$ is consistent for $\tilde{\lambda}$ at the rate of $n^{-\frac{1}{2}}$. Thus, $\left(\hat{p}, \hat{\mu}\right)$ are consistent for $\left(p, \mu\right)$ 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 ψ, the asymptotic normality is established. □

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

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

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

4 Simulation

In this section, I present Monte Carlo simulation results. I generated $1,000$ random samples $\{Y_i, D_i, X_i, Z_i\}_{i=1}^n$, from two data generating processes (DGP) that satisfy Assumption 1-3. In both of the two DGPs, I let Y_i be continuous and X_i, Z_i, U_i be discrete with three points in their support: $Y_i(1), Y_i(0) \in \mathbb{R}$ and $X_i, Z_i, U_i \in \{1, 2, 3\}$. The treatment $D_i \sim$ Bernoulli(0.5) was drawn independently of $\left(Y_i(1), Y_i(0), X_i, Z_i, U_i\right)$. Across the two DGPs, I varied the conditional distribution of Z_i given U_i to see how the estimation perfor-

	true value	bias				rMSE			
$\widehat{F}_{Y(1)-Y(0)}(0)$	0.084	0.000	0.000	-0.002	-0.001	0.014	0.009	0.011	0.007
$\widehat{F}_{Y(1)-Y(0)}(1)$	0.264	0.001	0.001	-0.001	-0.001	0.023	0.015	0.019	0.012
$\widehat{F}_{Y(1)-Y(0)}(2)$	0.536	0.001	0.000	0.000	-0.001	0.025	0.016	0.022	0.014
$\widehat{F}_{Y(1)-Y(0)}(3)$	0.775	0.002	0.000	0.002	0.000	0.020	0.012	0.018	0.011
$\widehat{F}_{Y(1)-Y(0)}(4)$	0.911	0.005	0.002	0.003	0.001	0.014	0.008	0.012	0.007
$\sigma_{\min}(\Lambda)$		0.337	0.337	0.806	0.806	0.337	0.337	0.806	0.806
n		750	2000	750	2000	750	2000	750	2000

Table 1: Bias and rMSE of DTE estimator $\widehat{F}_{Y(1)-Y(0)}(\delta)$ based on NMF.

mance depends on the informativeness of the proxy variable Z_i. The smallest singular value of the induced Λ is 0.377 and 0.806 across the two DGPs.[13]

Since Y_i is continuous, I used a three-way partition on $\mathbb{R}$ for Y_i in the first-step NMF: $(-\infty, 0], (0, 2], (2, \infty)$; the conditional probability matrices $\mathbf{H}_0$ and $\mathbf{H}_1$ were 9×3 matrices. For the choice of DTE parameter to assess the finite sample performance of the DTE estimator with, I used the marginal distribution of treatment effect $F_{Y(1)-Y(0)}(\delta)$ evaluated at $\delta = 0, 1, \cdots, 4$, using (15).

Table 1 contains the bias and the root mean squared error (rMSE) of the DTE estimators $\widehat{F}_{Y(1)-Y(0)}(\delta)$, across the two DGPs and sample size n. As Z_i becomes more informative for U_i, i.e. the smallest singular value $\sigma_{\min}(\Lambda)$ increases, the rMSE goes down. Additionally,

[13]The specifics of the DGPs are as follows. In the first DGP, $(p_U(1), p_U(2), p_U(3)) = (0.3, 0.3, 0.4)$,

$$\begin{pmatrix} \Pr\{X_i = 1|U_i = k\} \\ \Pr\{X_i = 2|U_i = k\} \\ \Pr\{X_i = 3|U_i = k\} \end{pmatrix} = \begin{pmatrix} \Pr\{Z_i = 1|U_i = k\} \\ \Pr\{Z_i = 2|U_i = k\} \\ \Pr\{Z_i = 3|U_i = k\} \end{pmatrix} = \begin{cases} \left(\frac{41}{45}, \frac{3}{45}, \frac{1}{45}\right)' & \text{if } k = 1 \\ \left(\frac{1}{20}, \frac{18}{20}, \frac{1}{20}\right)' & \text{if } k = 2 \\ \left(\frac{1}{45}, \frac{3}{45}, \frac{41}{45}\right)' & \text{if } k = 3 \end{cases}$$

and $Y_i(d) \mid U_i = k \sim \mathcal{N}\left(\mu^k(d), \sigma^k(d)^2\right)$ for $d = 0, 1$ where

$$\left(\mu^k(0), \sigma^k(0)\right) = \begin{cases} (-1, 1) & \text{if } k = 1 \\ (0, 1) & \text{if } k = 2 \\ (1, 1) & \text{if } k = 3 \end{cases} \quad \text{and} \quad \left(\mu^k(1), \sigma^k(1)\right) = \begin{cases} (1.5, 1.5) & \text{if } k = 1 \\ (2, 1) & \text{if } k = 2 \\ (2.5, 0.5) & \text{if } k = 3 \end{cases}.$$

For the second DGP, I only changed the conditional distribution of Z_i given U_i:

$$\begin{pmatrix} \Pr\{Z_i = 1|U_i = k\} \\ \Pr\{Z_i = 2|U_i = k\} \\ \Pr\{Z_i = 3|U_i = k\} \end{pmatrix} = \begin{cases} \left(\frac{62}{90}, \frac{21}{90}, \frac{7}{90}\right)' & \text{if } k = 1 \\ \left(\frac{7}{40}, \frac{26}{40}, \frac{7}{40}\right)' & \text{if } k = 2 \\ \left(\frac{7}{90}, \frac{21}{90}, \frac{62}{90}\right)' & \text{if } k = 3 \end{cases}$$

	true value	coverage probability			
$\widehat{F}_{Y(1)-Y(0)}(0)$	0.084	0.971	0.951	0.952	0.935
$\widehat{F}_{Y(1)-Y(0)}(1)$	0.264	0.975	0.959	0.958	0.952
$\widehat{F}_{Y(1)-Y(0)}(2)$	0.536	0.970	0.960	0.957	0.951
$\widehat{F}_{Y(1)-Y(0)}(3)$	0.775	0.962	0.959	0.943	0.951
$\widehat{F}_{Y(1)-Y(0)}(4)$	0.911	0.940	0.954	0.934	0.948
$\sigma_{\min}(\Lambda)$		0.337	0.337	0.806	0.806
n		750	2000	750	2000

Table 2: Coverage of 95% confidence interval based on NMF.

	true value	bias				rMSE			
$\widehat{F}_{Y(1)-Y(0)}(0)$	0.084	0.014	0.008	0.002	0.001	0.034	0.029	0.022	0.012
$\widehat{F}_{Y(1)-Y(0)}(1)$	0.264	0.006	0.004	0.002	0.000	0.030	0.021	0.024	0.014
$\widehat{F}_{Y(1)-Y(0)}(2)$	0.536	-0.006	-0.005	-0.001	0.000	0.037	0.029	0.025	0.015
$\widehat{F}_{Y(1)-Y(0)}(3)$	0.775	-0.009	-0.007	-0.001	-0.001	0.040	0.032	0.025	0.012
$\widehat{F}_{Y(1)-Y(0)}(4)$	0.911	-0.006	-0.004	0.000	-0.001	0.025	0.019	0.018	0.009
$\sigma_{\min}(\Lambda)$		0.337	0.337	0.806	0.806	0.337	0.337	0.806	0.806
n		750	2000	750	2000	750	2000	750	2000

Table 3: Bias and rMSE of DTE estimator $\widehat{F}_{Y(1)-Y(0)}(\delta)$ based on EVD.

Table 2 shows that the confidence intervals constructed with the asymptotic standard error achieve the target coverage.

Table 3 contains the bias and the rMSE of an alternative DTE estimator where the nuisance parameters Λ_0, Λ_1 are estimated with eigenvalue decomposition. The alternative estimation procedure failed for 15.4-45.2% of the simulated samples, due to nuisance parameter matrices being numerically singular.[14] Also, even conditioning on estimation success, the rMSE of the eigenvalue decomposition-based estimator is 1.25-4.77 times larger when $\sigma_{\min}(\Lambda) = 0.337$. The additional regularization in the NMF improves finite sample performance of estimation strategies based on a nonparametric finite mixture model, both on the extensive margin and the intensive margin.

[14]The failure rate of the eigenvalue decomposition estimation was 0.472, 0.334, 0.210, 0.154, for $\big(\sigma_{\min}(\Lambda), n\big) = (0.337, 750), (0.337, 2000), (0.806, 750), (0.806, 2000)$, respectively. The NMF estimation failed for one sample only, when $\big(\sigma_{\min}(\Lambda), n\big) = (0.337, 750)$.

5 Empirical illustration

As an empirical illustration, I revisit Jones et al. [2019] and estimate the effect of workplace wellness program on medical spending. As discussed in Example 2 of Section 2, Jones et al. [2019] fits the econometrics framework of this paper well. Firstly, the treatment, eligibility for a workplace wellness program, was randomly assigned. Secondly, it is plausible that the treatment mechanism is regime-changing in its nature since the workplace wellness program included information sessions on healthy lifestyle, which are designed to induce systemic changes in individuals' health-related behaviors including medical service-seeking and self-care practices. Lastly, the authors collected the outcome variable before the treatment assignment and after the treatment period, giving us two proxy variables.

Taking advantage of the random assignment, Jones et al. [2019] estimated the intent-to-treat type ATE of the workplace wellness program on the monthly medical spending. The ATE estimate 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.[15] 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."[16] I build onto this intuition and estimate the DTE parameters, to explore the treatment effect heterogeneity.

The dataset of Jones et al. [2019] contains monthly medical spending records for the following three time periods: July 2015-July 2016, August 2016-July 2017 and August 2017-January 2019. The experiment started in the summer of 2016 and the treated individuals

[15] 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.

[16] 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.

	(1)	(2)	(3)	(4)	(5)
$CATE(\mathcal{Y})$ (\$)	-20.43	164.17	297.49	-157.24	-732.94*
	(166.31)	(389.05)	(350.17)	(515.51)	(443.00)
$\mathcal{Y}$	$(-\infty, 42.7]$	$(42.7, 132.2]$	$(132.2, 286.8]$	$(286.8, 671.1]$	$(671.1, \infty]$

Table 4: Conditional average treatment effect $\mathbf{E}[Y_i(1) - Y_i(0)|Y_i(0) \in \mathcal{Y}]$.

Note: The five conditioning sets on $Y_i(0)$ as chosen with $F^{-1}_{Y(0)}(1/5), \cdots, F^{-1}_{Y(0)}(4/5)$.
The medical spending variable is measured in \$. A 10% significance level is denoted with $*$.

were offered to participate in the wellness program starting the fall semester of 2016:

Y_i : monthly medical spending during August 2016-July 2017

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

X_i : monthly medical spending during July 2015-July 2016

Z_i : monthly medical spending during August 2017-January 2019

X_i is the pretreatment outcome variable and Z_i is the post-treatment outcome variable. To connect the hidden Markov model in Example 2 to this empirical context, we can think of the common shock V_{it} as underlying health status. The first-order Markovian assumption on the underlying health status is consistent with the practices in health economics literature where the underlying health status variable is often modeled to be first-order autoregressive: Grossman [1972], Wagstaff [1993], Jacobson [2000], Yogo [2016] and more.

In formulating $\mathbf{H}_0, \mathbf{H}_1$ for the first-step NMF, I let $M_Y = 4$ and $M_X = 6$, using equal partitions based on quantiles. Thus, $\mathbf{H}_0, \mathbf{H}_1$ have 24 rows. For K, the number of points in the support of U_i, I turn to the rank estimator/test (appendix Section C) and the falsification tests (appendix Section B). The rank estimator and the rank test from Ahn and Horenstein [2013] and Kleibergen and Paap [2006] both suggest $K \geq 3$. Moreover, the falsification tests based on the testable implications (8) and (9) suggest $K = 5$. Thus, I let $M_Z = K = 5$, using an equal partition for Z_i. More discussion and robustness analyses with regard to K are provided in the Online Appendix.

Table 4 contains the estimates for conditional ATE, using $Y_i(0)$ as the conditioning vari-

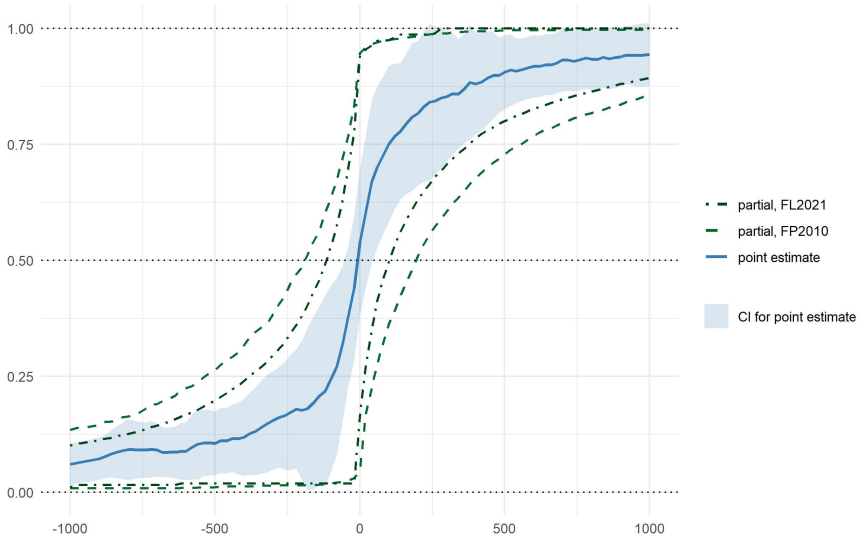

Figure 1: Marginal distribution of $Y_i(1) - Y_i(0)$.

Note: The distribution of treatment effect $Y_i(1) - Y_i(0)$ is estimated for $[-\$1000, \$1000]$. Additionally, 95% pointwise confidence intervals and partial bounds from Fan and Park [2010], Frandsen and Lefgren [2021] are provided.

able. The five conditioning sets are five quintiles of $Y_i(0)$. Conditioning on the fifth quintile of $Y_i(0)$, the CATE estimate is statistically significant at 10% significance level. For other quintiles, the estimates are not statistically significant. This suggests that the treatment may reduce the monthly medical spending among individuals with high baseline medical spending, while having little impact on the average.

Figure 1 contains the estimated marginal distribution of the treatment effect and its 95% pointwise confidence interval. Overall, it is unclear if the share of winners is bigger than 0.5; though the point estimate $\widehat{F}_{Y(1)-Y(0)}(0)$ is 0.540, the 95% confidence interval for $F_{Y(1)-Y(0)}(0)$ includes 0.5, not being able to reject the null $F_{Y(1)-Y(0)}(0) \leq 0.5$. As comparison, partial bounds from Fan and Park [2010], Frandsen and Lefgren [2021] are plotted in Figure 1. The DTE estimates are consistent with the partial identification, lying between the two bounds. At zero, the comparison highlights the information gain; the width of the 95% confidence interval for $F_{Y(1)-Y(0)}(0)$ using the DTE estimator is 0.315 while the width of the two partial bounds is 0.899.

6 Conclusion

An important avenue for future research is how we extend the current estimation strategy to account for a continuous latent variable U_i, while retaining the desirable properties of the discretization-based estimation strategy. For example, we may directly correct for the discretization bias by using some subsampling-based method such as the jackknife correction. Another important follow-up research direction would be to pursue a fully-developed policy learning framework based on the DTE estimation, providing distributional guarantees on policy recommendation.

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, Koen Jochmans, and Jean-Marc Robin. Estimating multivariate latent-structure models. *The Annals of Statistics*, pages 540–563, 2016.

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.

Larry G Epstein and Uzi Segal. Quadratic social welfare functions. *Journal of Political Economy*, 100(4):691–712, 1992.

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.

H Levy and HM Markowitz. Approximating expected utility by a function of mean and variance. *The American Economic Review*, 69(3):308–317, 1979.

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.

Henry WJ Reeve, Timothy I Cannings, and Richard J Samworth. Optimal subgroup selection. *The Annals of Statistics*, 51(6):2342–2365, 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 for 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 for multivariate U_i, X_i, and Z_i. (Theorem 1 of Hu and Schennach [2008])[17] 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.*

[17] This point is also utilized in Cunha et al. [2010] again. (Theorem 2 of Cunha et al. [2010])

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* $u \neq u'$,

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

for each $d = 0, 1$.

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

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

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, suppose that U_i, X_i, Z_i are all two-dimensional vectors, with a finite support:[18]

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

$|\mathcal{U}_1| = |\mathcal{X}_1| = |\mathcal{Z}_1| = K_1$ and $|\mathcal{U}_2| = |\mathcal{X}_2| = |\mathcal{Z}_2| = K_2$, with $K = K_1 \cdot K_2$. (X_{i1}, Z_{i1}) are

[18]This is a common practice in the early childhood development literature. Cunha et al. [2010], Attanasio et al. [2020] theorize U_i to be two-dimensional: 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.

proxies for U_{i1} and (X_{i2}, Z_{i2}) for U_{i2}. Then, we can modify the NMF problem (10) as follows:

1. Label the rows of Λ_0, Λ_1 and the columns of Γ_0, Γ_1 to each dimension of U_i. For example, the first K_1 rows of Λ_0, Λ_1 and the first K_1 columns of Γ_0, Γ_1 correspond to the same value of U_{i1} and so on.[19]

2. Add additional constraints on Γ_0 and Γ_1 such that each column of Γ_0 and Γ_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 Γ_0, Γ_1 that correspond to the same value of U_{i1} and likewise for X_{i2} and U_{i2}.

This will impose additional regularization on the NMF estimator.

B Falsification tests

In this section, I formally develop two falsification tests in a discrete U_i setup. Firstly, I construct a test statistic based on (8): $X_i \perp\!\!\!\perp D_i \mid U_i$. For the test statistic based on (8), we need to modify the NMF problem (10). 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 (8). Thus, I modify (10) by dropping the linear constraints (11). As long as we impose the quadratic constraints (12), the NMF optimization problem still admits a unique solution up to some permutation, when $\mathbb{H}_d$ is close to $\mathbf{H}_d$.

Let $\widehat{p}_{X|D=0,U}(\mathcal{X}|u)$ denote the plug-in GMM estimators for $\Pr\{X_i \in \mathcal{X}|D_i = 0, U_i = u\}$ and $\widehat{p}_{X|D=1,U}(\mathcal{X}|u)$ for $\Pr\{X_i \in \mathcal{X}|D_i = 1, U_i = u\}$, for a given $\mathcal{X} \subset \mathbb{R}$. Since the modified first-step NMF does not have built-in conditional independence between X_i and D_i, we can use $\widehat{p}_{X|D=0,U}(\mathcal{X}|u)$ and $\widehat{p}_{X|D=1,U}(\mathcal{X}|u)$ to test the distributional equivalence. Fix some

[19] Conversely, we can think of the Γ_d and Λ_d as an outcome of the "unfolding" procedure proposed in Bonhomme et al. [2016].

partition $\mathcal{X}^1, \cdots, \mathcal{X}^{\tilde{M}}$ such that $\cup_{m=1}^{\tilde{M}} \mathcal{X}^m = \mathbb{R}$. Then,

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

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

$$\begin{aligned} T_n^1 = n & \begin{pmatrix} \widehat{p}_{X|D=0,U}(\mathcal{X}^1|u^1) - \widehat{p}_{X|D=1,U}(\mathcal{X}^1|u^1) \\ \vdots \\ \widehat{p}_{X|D=0,U}(\mathcal{X}^{\tilde{M}}|u^K) - \widehat{p}_{X|D=1,U}(\mathcal{X}^{\tilde{M}}|u^K) \end{pmatrix}^{\intercal} \widehat{\Sigma^1}^{-1} \\ & \cdot \begin{pmatrix} \widehat{p}_{X|D=0,U}(\mathcal{X}^1|u^1) - \widehat{p}_{X|D=1,U}(\mathcal{X}^1|u^1) \\ \vdots \\ \widehat{p}_{X|D=0,U}(\mathcal{X}^{\tilde{M}}|u^K) - \widehat{p}_{X|D=1,U}(\mathcal{X}^{\tilde{M}}|u^K) \end{pmatrix} \end{aligned} \tag{22}$$

with $\widehat{\Sigma^1}$ being the plug-in estimator for the asymptotic variance Σ^1. The test statistic (22) is infeasible without further assumptions since I dropped the linear constraints (11) 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 (22) for every permutation on $\{1, \cdots, K\}$ and take the minimum, following the same spirit of minimizing over $\tilde{g}$ in (8).

Additionally, when D_i is randomly assigned as in Remark 1, we can directly test (9): $D_i \perp\!\!\!\perp U_i$. 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. Let $\widehat{p}_{U|D=0}(u)$ denote the plug-in GMM estimator for $\Pr\{U_i = u|D_i = 0\}$ and $\widehat{p}_{U|D=1}(u)$ for $\Pr\{U_i = u|D_i = 1\}$. Then,

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

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

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

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

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

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

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

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.[20] I estimate $\mathbf{H}_X$ with

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

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. Following Higgins [2025]'s treatment, I apply the eigenvalue ratio estimator of Ahn and Horenstein [2013] to $\mathbb{H}_X$.[21] Let $\nu_k(\mathbf{H})$ denote the k-th largest eigenvalue of $\mathbf{H}\mathbf{H}^\intercal$ when $M_X \leq 2M_Z$ and that of $\mathbf{H}^\intercal\mathbf{H}$ when $M_X \geq 2M_Z$. The eigenvalue ratio estimator is

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

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:

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

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

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

[20] 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.

[21] 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.

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. Kleibergen and Paap [2006] provides an explicit formula for the test statistic.

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 $\left\{f_{Y(1)|U}(\cdot|u), f_{Y(0)|U}(\cdot|u), f_{X|U}(\cdot|u)\right\}_{u\in\mathcal{U}}$, 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 $\tilde{U}_i = h(U_i) := Mf_{Y(d)|U}(\cdot|U_i)$ 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 $f_{Y(1)|\tilde{U}}, f_{Y(0)|\tilde{U}}$ and $f_{X|\tilde{U}}$. Note that both Assumptions 1-2 hold with $\tilde{U}_i$ in place of U_i since h is strictly increasing.

To complete the proof, let us show that the marginal distribution of $\tilde{U}_i$ is identified as well. For that, firstly I establish the injectivity of the integral operator based on the conditional density of X_i given $\tilde{U}_i$. Find that

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

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 identified. 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 $(Y_i(1), Y_i(0), D_i, X_i, Z_i)$. □

D.2 Proof for Theorem 2

The following proof is for Λ_0 and $K \geq 2$. The same proof applies to Λ_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,* $\left\|\Gamma_0\Lambda_0 - \widehat{\Gamma}_0\widehat{\Lambda}_0\right\|_F^2 = O_p\left(\frac{1}{\sqrt{n}}\right)$.

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

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

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* $n \to \infty$.

Lemma 2 shows that the NMF estimator retrieves the true conditional densities $\mathbf{H}_d = \Gamma_d\Lambda_d$ at the rate of $n^{-\frac{1}{2}}$. Lemma 3 shows that the estimator $\widehat{\Gamma}_0$ is consistent for some rotation of Γ_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 $\widehat{\Gamma}_0$ and the rows of $\widehat{\Lambda}_0$ so that A converges to $\mathbf{I}_K$,

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

□

D.3 Proof for Proposition 1

Firstly, using Assumptions 1-2, the joint density of $\left(Y_i(1), Y_i(0), D_i, X_i\right)$ factors into the conditional density of $\left(Y_i(1), Y_i(0), X_i\right)$ given U_i and the conditional density of U_i given D_i:

$$f_{Y(1),Y(0),D,X}(y, y', d, x) = \sum_{k=1}^{K} f_{Y(1),X|U}(y, x|u^k) \cdot f_{Y(0)|X}(y'|u^k) \cdot \underbrace{\frac{p_{D,U}(d,k)}{\sum_{j=1}^{K} p_{D,Z}(d,j)}}_{=f_{U|D}(u^k|d)}.$$

Secondly, recall the finite mixture representation:

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

Thus, the conditional density of $\left(Y_i(d), X_i\right)$ given U_i is linear in conditional densities of observable variables:

$$\begin{pmatrix} f_{Y(d),X|U}(y,x|u^1) & \cdots & f_{Y(d),X|U}(y,x|u^K) \end{pmatrix} = \begin{pmatrix} f_{Y,X|D=d,Z}(y,x|z^1) & \cdots & f_{Y,X|D=d,Z}(y,x|z^K) \end{pmatrix} \tilde{\Lambda}_d.$$

For notational convenience, let $\tilde{\lambda}_{jk,d}$ denote the j-th row and k-th column component of $\tilde{\Lambda}_d$. Then, $f_{Y(d),X|U}(y,x|u^k) = \sum_{j=1}^{K} \tilde{\lambda}_{jk,d} f_{Y|D=d,Z}(y|z^j)$.

Using the linear relationship, the infeasible moment condition (13) becomes

$$
\begin{aligned}
0 &= \mathbf{E}\left[m\big(Y_i(1), Y_i(0), D_i, X_i; \theta\big)\right] \\
&= \sum_{d=0,1} \int_{\mathbb{R}^3} m(y', y, d, x; \theta) \sum_{k=1}^{K} f_{Y(1),X|U}(y, x|u^k) \cdot f_{Y(0)|X}(y'|u^k) \cdot \frac{p_{D,U}(d,k)}{\sum_{j=1}^{K} p_{D,Z}(d,j)} dy dy' dx \\
&= \sum_{d=0,1} \sum_{k=1}^{K} \sum_{j=1}^{K} \sum_{j'=1}^{K} \frac{\tilde{\lambda}_{jk,0}\tilde{\lambda}_{jk,1}}{p_{D,Z}(0,j)p_{D,Z}(1,j')} \cdot \frac{p_{D,U}(d,k)}{\sum_{j=1}^{K} p_{D,Z}(d,j)} \\
&\qquad \int_{\mathbb{R}^3} m(y', y, d, x; \theta) \cdot f_{Y,D,Z}(y, 0, z^j) \cdot f_{Y,D,X,Z}(y', 1, x, z^{j'}) dy dy' dx \\
&= \sum_{d=0,1} \sum_{k=1}^{K} \frac{p_{D,U}(d,k)}{\sum_{j=1}^{K} p_{D,Z}(d,j)} \sum_{j=1}^{K} \sum_{j'=1}^{K} \frac{\tilde{\lambda}_{jk,0}\tilde{\lambda}_{jk,1}}{p_{D,Z}(0,j)p_{D,Z}(1,j')} \\
&\qquad \mathbf{E}\left[m(Y_{i'}, Y_i, d, X_{i'}; \theta)\mathbf{1}\{D_i = 0, Z_i = z^j\}\mathbf{1}\{D_{i'} = 1, Z_{i'} = z^{j'}\}\right].
\end{aligned}
$$

Thus, we can construct the feasible moment condition (14) with

$$
\begin{aligned}
&\tilde{m}\big((Y_i, D_i, X_i, Z_i), (Y_{i'}, D_{i'}, X_{i'}, Z_{i'}); \theta, \tilde{\lambda}, p\big) \\
&\quad = \sum_{d=0,1} \sum_{k=1}^{K} \frac{p_{D,U}(d,k)}{\sum_{j=1}^{K} p_{D,Z}(d,j)} \sum_{j=1}^{K} \sum_{j'=1}^{K} \frac{\tilde{\lambda}_{jk,0}\tilde{\lambda}_{jk,1}}{p_{D,Z}(0,j)p_{D,Z}(1,j')} \\
&\qquad\qquad \cdot m(Y_{i'}, Y_i, d, X_{i'}; \theta)\mathbf{1}\{D_i = 0, Z_i = z^j\}\mathbf{1}\{D_{i'} = 1, Z_{i'} = z^{j'}\}.
\end{aligned}
$$

□