Distributional Treatment Effect with Latent Rank Invariance*

Myungkou Shin†

June 20, 2025

Abstract

Treatment effect heterogeneity is of a great concern when evaluating policy impact. However, even in the simple case of a binary random treatment, researchers' interests have been mostly limited to an average treatment effect or a quantile treatment effect, due to the fundamental limitation that we cannot simultaneously observe both treated potential outcome and untreated potential outcome for a given unit. This paper assumes a conditional independence assumption that the two potential outcomes are independent of each other given a scalar latent variable. With a specific example of strictly increasing conditional expectation, we label the latent variable as 'latent rank' and motivate the identifying assumption as 'latent rank invariance.' In implementation, we assume a finite support on the latent variable and propose an estimation strategy based on a nonnegative matrix factorization. A limiting distribution is derived for the distributional treatment effect estimator, using Neyman orthogonality.

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

JEL classification codes: C13

*I acknowledge the support from the European Research Council through the grant ERC-2018-CoG-819086-PANEDA. Any and all errors are my own.

†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 whole distribution of treatment effect, researchers often try to estimate some summary measures of the treatment effect, 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 only be answered with the *distribution* of the treatment effect; e.g., is the treatment Pareto improving? what is the share of people who are worse off under the treatment regime? This paper aims to answer these questions, by identifying the distribution of treatment effect.

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. 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 resort to *conditional* independence, by assuming a scalar latent variable that captures the individual-level heterogeneity in terms of the dependence between the two potential outcomes. For illustration, consider a simple additive model as in Heckman et al. (1997). The observed outcome

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

is constructed with a treated potential outcome $Y(1)$, an untreated potential outcome $Y(0)$, and a binary treatment variable $D \in \{0, 1\}$ and the two potential outcomes

$$Y(1) = \mu_1(U) + \varepsilon(1) \tag{1}$$

$$Y(0) = \mu_0(U) + \varepsilon(0) \tag{2}$$

are constructed as follows with a unit-level latent variable $U \in \mathcal{U} \subset \mathbb{R}$ and two treatment-status-specific random shocks $\varepsilon(1)$ and $\varepsilon(0)$. When

$$\varepsilon(1) \perp\!\!\!\perp \varepsilon(0) \mid U, \tag{3}$$

we can characterize the joint distribution of the two potential outcome as follows:

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

Thus, the task of identifying the joint distribution of the two potential outcomes becomes that of identifying the condtional distribution of $\varepsilon(1)$ given U, the conditional distribution of $\varepsilon(0)$ given U, and the marginal distribution of U.

The assumption that there exists a latent variable U that captures the dependence between the two potential outcomes is not new to the literature. A condition in the same spirit as (3) has been widely used in quantile treatment effect estimation and quantile regression, in the name of rank invariance/similarity. (See Chernozhukov and Hansen (2005, 2006); Athey and Imbens (2006); Vuong and Xu (2017); Callaway and Li (2019); Han and Xu (2023) and more.) The rank invariance condition assumes that the rank of one potential outcome is equal to the rank of the other potential outcome, in their respective distributions. In (1)-(2), this assumption can be characterized as $\varepsilon(1)$ and $\varepsilon(0)$ being nonrandom. The usage of the rank invariance condition is mostly limited to the quantile treatment effect and not applied to the distributional treatment effect,[1] due to the fact that it imposes excessive restriction on the joint distribution of the two potential outcomes; the distribution of the quantile treatment effect is equal to the distribution of the individual-level treatment effect. On the other hand, in this paper, I do not directly assume rank invariance. While I assume that U plays a similar role as the rank in the rank invariance assumption in restricting the dependence between the two potential outcomes, I let U be a latent variable and $\varepsilon(1), \varepsilon(0)$ be random variables.

In the conditional independence framework, the key step in identifying the joint distribution of the two potential outcomes is to identify the conditional distributions and the marginal distributions involving the latent variable. For that, I assume that there are two additional proxy variables X, Z, observable to the econometrician (see Hu and Schennach (2008); Miao et al. (2018); Deaner (2023);

[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 reserve the expression 'distributional' to only when the object involves the joint distribution of the two potential outcomes.

Kedagni (2023); Nagasawa (2022) and more). In the simple example (1)-(2), the proxy variables will shift $\mu(U)$ independently of $\big(\varepsilon(1), \varepsilon(0)\big)$, allowing us to decompose the variation in $Y(d)$ into the variation from U and that from $\varepsilon(d)$. Additionally, since I do not adopt the 'measurement error' interpretation on the proxy variables, I assume that there exists a functional of the joint distribution of the potential outcomes given the latent variable, which strictly increases in the latent variable U. Using the strict monotonicity, I can identify the marginal distribution of U and integrate the product of the conditional densities across the support of U. An example of such a functional is conditional expectation. Suppose that the two conditional expectations $\mathbf{E}\left[Y(1)|U=u\right]$ and $\mathbf{E}\left[Y(0)|U=u\right]$ are strictly increasing in u. Within this example, the latent variable U can be thought of as the rank of the conditional means $\mathbf{E}[Y(1)|U]$ and $\mathbf{E}[Y(0)|U]$: hence the terminology 'latent rank invariance.'

To develop on the identification result and estimate a distributional treatment effect parameter, I assume a finite support on U and uses a nonnegative matrix factorization algorithm. Though I assume that U is finitely discrete in the estimation, the identification result does not require such an assumption and I develop an alternative estimation method based on sieve maximum likelihood for a setup with continuous U in the Appendix subsection A.2. The finite support assumption in the main estimation strategy is for computational practicality and draws from recent developments in the literature regarding discretizing a continuous latent heterogeneity: see Bonhomme et al. (2022).

The estimation procedure is two-step. In the first step, I estimate the conditional probability $\Pr\{U=u|Z=z\}$, using the nonnegative matrix factorization. In the second step, I characterize the distributional treatment effect (DTE) parameters

$$\Pr\{Y(1) \leq y_1, Y(0) \leq y_0\}$$
$$\Pr\{Y(1) - Y(0) \leq \delta\}$$

as a function of the probabilities $\Pr\{Y \leq y|D=d, Z=z\}$ and $\Pr\{U=u|Z=z\}$. The former probability is directly observed from the dataset and the latter is estimated in the first step. Thus, the estimator can be thought of as a plug-in GMM estimator, where nuisance parameters are estimated in the first-step nonnegative matrix factorization. Asymptotic normality of the distributional treatment effect parameters is established. In deriving asymptotic normality, I construct a moment condition that satisfies Neyman orthogonality to be robust to the first-step estimation error from the nonnegative matrix factorzation.

This paper makes contribution to the distributional treatment effect literature by proposing a framework where the joint distribution of the potential outcomes and thus the marginal distribution of treatment effect are point identified, without imposing any parametric structure on the potential outcomes. This is in contrast to the partial identification results in the literature: Heckman et al. (1997); Fan and Park (2010); Fan et al. (2014); Firpo and Ridder (2019); Frandsen and Lefgren (2021) and more. There exist several notable point identification results, which use a type of (conditional) independence assumption as I do: Heckman et al. (1997); Carneiro et al. (2003); Gautier and Hoderlein (2015); Noh (2023). Gautier and Hoderlein (2015) use assumptions on the treatment assignment model, arguably more restrictive than the restrictions I impose on the treatment assignment, to identify the distribution of the latent variable. Heckman et al. (1997) discusses an identification result similar to this paper while assuming a parametric structure. In this sense, the identification result in this paper can be thought of as a nonparametric version of Heckman et al. (1997), at the cost of assuming a univariate U_i.

The rest of the paper is organized as follows. Section 2 discusses the identification result for the joint distribution of the two potential outcomes. Section 3 explains the estimation method for the two distributional treatment effect parameters and develops asymptotic theory for the estimators. Section 4 contains Monte Carlo simulation restuls and Section 5 applies the estimation procedure to an empirical dataset from Jones et al. (2019).

2 Identification

An econometrican observes a dataset $\{Y_i, D_i, X_i, Z_i\}_{i=1}^n$ where $Y_i, X_i, Z_i \in \mathbb{R}$ and $D_i \in \{0, 1\}$. Y_i is an outcome variable, D_i is a binary treatment variable and X_i, Z_i are two proxy variables. 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 $U_i \in \mathcal{U} \subset \mathbb{R}$. 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. 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 that the treatment D_i is random conditioning on the latent variable U_i and the other proxy variable Z_i, with regard to the potential outcomes and one of the proxy variables.

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

This assumption assumes that the proxy variable Z_i and the latent variable U_i contain sufficient information on the treatment assignment process so that any remaining variation in D_i is as good as random with regard to the potential outcomes and X_i. In this sense, Assumption 1 is a restriction on treatment endogeneity.

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 from Assumption 1 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 $Y_i(0)$, I assume that the latent variable U_i captures all of the dependence between the two potential outcomes and the proxy variables.

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

Note that the latent variable U_i lies in $\mathbb{R}$ as do $Y_i(1)$ and $Y_i(0)$. This excludes the non-binding case where $U_i = \big(Y_i(1), Y_i(0)\big)$. Assumption 2 assumes that $Y_i(1)$ and $Y_i(0)$ are independent of each other conditioning on U_i. In addition, Assumption 2 assumes that the two proxy variables X_i and Z_i satisfy the standard exclusion restriction; given U_i, the proxy variables do not give us additional information on the distribution of the potential outcomes.

When U_i is observed, Assumptions 1-2 identify the joint distribution of the two potential outcomes and various distributional treatment effect parameters. Examples include the variance of the treatment effect $\mathrm{Var}\big(Y_i(1) - Y_i(0)\big)$ and the marginal distribution of the treatment effect $\Pr\{Y_i(1) - Y_i(0) \leq \delta\}$. 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 in the identification result of this paper.

Assumption 2 plays a key role in the identification result. To motivate the conditional independence assumption, I present two examples of models with proxy variables that satisfy Assumption 2. The first example is a short panel with past and future outcomes and proxy variables.

Example 1. For $t = 1, 2, 3$ and $d = 0, 1$,

$$Y_{it}(d) = g_d\big(V_{it}, \varepsilon_{it}(d)\big), \tag{5}$$
$$Y_{it} = D_{it} \cdot Y_{it}(1) + (1 - D_{it}) \cdot Y_{it}(0).$$

$\{V_{it}\}_{t=1}^{3}$ is first-order Markovian and $\{V_{it}\}_{t=1}^{3}, \varepsilon_{i1}(0), \varepsilon_{i2}(1), \varepsilon_{i2}(0)$ and $\varepsilon_{i3}(1)$ are mutually independent. $D_{i1} = 0$ and $D_{i3} = 1$ with probability one and $0 < \Pr\{D_{i2} = 1\} < 1$.

A potential outcome $Y_{it}(d)$ is a function of a latent variable V_{it} and an error term $\varepsilon_{it}(d)$. Note that V_{it} appears in the model twice; for $Y_{it}(1)$ and for $Y_{it}(0)$. In this sense, V_{it} is a common shock to the potential outcomes where $\varepsilon_{it}(d)$ is a treatment-status-specific shock. In this example, Assumption 2 is satisfied by letting $Y_i = Y_{i2}, X_i = Y_{i1}, Z_i = Y_{i3}$ and $U_i = V_{i2}$. When the treatment D_{i2} is assigned randomly, Assumption 1 holds as well. The key elements of Example 1 are that the common shock process $\{V_{it}\}_{t=1}^{3}$ and the treatment-status-specific shocks $\varepsilon_{i1}(1), \ldots, \varepsilon_{i3}(0)$ are all mutually independent and that dependence within $\{V_{it}\}_{t=1}^{3}$ themselves is restricted to be first-order Markovian. Thus, V_{i2} has sufficient information on the dependence between $Y_{i2}(1)$ and $Y_{i2}(0)$ and the past and the future outcomes Y_{i1} and Y_{i3} can be used as a proxy for V_{i2}.

Even when the treatment D_{i2} is not random, Assumption 1 may not be too restrictive an assumption in the context of Example 1. Suppose that the common shock V_{it} is drawn first and then the treatment-status-specific shocks $\varepsilon_{it}(1)$ and $\varepsilon_{it}(0)$ are drawn subsequently and that at time $t = 2$, individuals select into treatment by comparing their expected gain from being treated with their costs η_i before the treatment-status-specific shocks are realized:

$$D_{i2} = \mathbf{1}\{\mathbf{E}\left[Y_{i2}(1) - Y_{i2}(0) | V_{i2}\right] \geq \eta_i\}$$

The assignment model above assumes that at the timing of selection, individuals are only aware of their common shock V_{i2} and thus their (conditionally) expected gain $\mathbf{E}\left[Y_{i2}(1) - Y_{i2}(0) | V_{i2}\right]$, but not the realized gain $Y_i(1) - Y_i(0)$. When η_i, the idiosyncratic shock in the assignment model, is independent of the shocks in the outcome model, Assumption 1 is satisfied.

The second example is repeated measurements on the latent variable, as in Carneiro et al. (2003). Here I present a simplied version of their model. X_i and Z_i are measurements on the latent variable U_i and both the potential outcomes and the repeated measurements follow a factor model.

Example 2. *(Carneiro et al., 2003) For* $d = 0, 1$,

$$\begin{aligned}
Y_i(d) &= U_i^\intercal \alpha^d + \varepsilon_i(d), \\
X_i &= U_i^\intercal \alpha^X + \varepsilon_{X,i}, \\
Z_i &= U_i^\intercal \alpha^Z + \varepsilon_{Z,i}.
\end{aligned}$$

$U_i \perp\!\!\!\perp (\varepsilon_i(0), \varepsilon_i(1), \varepsilon_{X,i}, \varepsilon_{Z,i})$ *and* $(\varepsilon_i(0), \varepsilon_i(1), \varepsilon_{X,i}, \varepsilon_{Z,i})$ *are mutually independent.*

In this example of repeated measurements, we often look to an economic model to have an interpretation on the latent variable U_i and find measurements on the latent variable. For example, in Carneiro et al. (2003), the model assumed U_i to be an innate ability of an individual and test scores were used as proxy variable. Assuming a linear structure, Carneiro et al. (2003) allows U_i to be multidimensional factor. In this paper, I allow for more flexibility on the dependence of $Y_i(d), X_i, Z_i$ on U_i, while restricting U_i to be one-dimensional.

The remainder of this section outlines the identification argument. For illustration purposes only, let Y_i, X_i, Z_i, U_i be 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 $U_i \in \{u^1, \cdots, u^K\}$. Then, we can construct a $M \times M_Z$ matrix $\mathbf{H}_d$ of conditional probabilities as follows: with $M = M_Y \cdot M_X$,

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

for each $d = 0, 1$. $\mathbf{H}_0$ is the conditional probability of (Y_i, X_i) given Z_i in the untreated subsample and $\mathbf{H}_1$ is the conditional probability in the treated subsample. From Assumptions 1-2, both $\mathbf{H}_0$ and $\mathbf{H}_1$ decompose into a multiplication of two matrices: for each $d = 0, 1$,

$$\mathbf{H}_d = \Gamma_d \cdot \Lambda_d \tag{6}$$

where

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

$$\Lambda_d = \begin{pmatrix} \Pr\left\{U_i = u^1 | \left(D_i, Z_i\right) = \left(d, z^1\right)\right\} & \cdots & \Pr\left\{U_i = u^1 | \left(D_i, Z_i\right) = \left(d, z^{M_Z}\right)\right\} \\ \vdots & \ddots & \vdots \\ \Pr\left\{U_i = u^K | \left(D_i, Z_i\right) = \left(d, z^1\right)\right\} & \cdots & \Pr\left\{U_i = u^K | \left(D_i, Z_i\right) = \left(d, z^{M_Z}\right)\right\} \end{pmatrix}. \tag{7}$$

Note that the discreteness of Y_i, X_i, Z_i is nonbinding; we can use partitioning on $\mathbb{R}$ when they are continuous.[2] Thus, the only effective discretization is on U_i, which is imposed only for the expositional brevity. The identification argument does not hinge on the discreteness of U_i; a continuous version of the identification follows the same argument and uses one additional assumption to find a labeling on the infinite number of functions: Assumption 5. I present more discussion on Assumption 5 later in this section and a full identification argument for continuous U_i is provided in Subsection A.1 of Appendix.

The equation $\mathbf{H}_d = \Gamma_d \cdot \Lambda_d$ shows us that the conditional density model in (15) is indeed 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 across the subpopulations.

From Assumption 2, 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 identify the joint distribution of $Y_i(1)$ and $Y_i(0)$. Note that $(\Gamma_0, \Gamma_1, \Lambda_0, \Lambda_1)$ contain those distributions of interest. To decompose $\mathbf{H}_d$ into Γ_d and Λ_d,

[2] 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 $\mathcal{W}^1 = \mathcal{Y}^1 \times \mathcal{X}^1, \mathcal{W}^2 = \mathcal{Y}^2 \times \mathcal{X}^1, \cdots, \mathcal{W}^M = \mathcal{Y}^{M_Y} \cdot \mathcal{X}^{M_X}$. $\{\mathcal{W}^m\}_{m=1}^M$ 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.

first fix $y \in \{y^1, \cdots, y^{M_Y}\}$ and extract rows of $\mathbf{H}_d$ and Γ_d that correspond to $(y, x^1), \cdots, (y, x^{M_X})$:

$$\mathbf{H}_d(y) = \Big(\Pr\big\{(Y_i, X_i) = \big(y, x^j\big) \,\big|\, (D_i, Z_i) = \big(d, z^k\big)\big\}\Big)_{1\le j\le M_X, 1\le k\le M_Z},$$
$$\Gamma_d(y) = \Big(\Pr\big\{(Y_i(d), X_i) = \big(y, x^j\big) \,\big| U_i = u^k\big\}\Big)_{1\le j\le M_X, 1\le k\le K}.$$

for $d = 0, 1$. From Assumption 2, the mixture component density matrix $\Gamma_d(y)$ can be further decomposed:

$$\begin{aligned}\Gamma_d(y) = &\begin{pmatrix} \Pr\big\{X_i = x^1 \big| U_i = u^1\big\} & \cdots & \Pr\big\{X_i = x^1 \big| U_i = u^K\big\} \\ \vdots & \ddots & \vdots \\ \Pr\big\{X_i = x^{M_X} \big| U_i = u^1\big\} & \cdots & \Pr\big\{X_i = x^{M_X} \big| U_i = u^K\big\} \end{pmatrix} \\ &\cdot \operatorname{diag}\big(\Pr\big\{Y_i(d) = y \big| U_i = u^1\big\}, \cdots, \Pr\big\{Y_i(d) = y \big| U_i = u^K\big\}\big) \\ =: &\, \Gamma_X \cdot \Delta_d(y).\end{aligned}$$

Now, sum $\mathbf{H}_d(y)$ across $y^1, \cdots, y^{M_Y}$:

$$\sum_y \mathbf{H}_d(y) = \Gamma_X \cdot \sum_y \Delta_d(y) \cdot \Lambda_d = \Gamma_X \cdot \Lambda_d.$$

Find that when $M_X = M_Z = K$ and both Γ_X and Λ_d have full rank,

$$\begin{aligned}\mathbf{H}_d(y) \left(\sum_y \mathbf{H}_d(y)\right)^{-1} &= \Gamma_X \cdot \Delta_d(y) \cdot \Lambda_d \left(\Gamma_X \cdot \Lambda_d\right)^{-1} \\ &= \Gamma_X \cdot \Delta_d(y) \cdot {\Gamma_X}^{-1}.\end{aligned}$$

Given a no repeated eigenvalue condition that for any $u \neq u'$ there exist some (y, d) such that $\Pr\{Y_i(d) = y | U_i = u\} \neq \Pr\{Y_i(d) = y | U_i = u'\}$, diagonalization of $\mathbf{H}_d(y) \left(\sum_y \mathbf{H}_d(y)\right)^{-1}$ across different y and d identifies Γ_X and $\{\Delta_d(y)\}_{y^1 \le y \le y^{M_Y}}$.[3] Once Γ_X is identified, the identification of Λ_0, Λ_1 follows from Γ_X having full rank. When M_X or M_Z is bigger than K, we may stack some of the rows or the columns of $\sum_y \mathbf{H}_d(y)$ to make it into a square matrix.

Assumption 3 formally states the full rank condition and the no repeated eigenvalue condition for discrete U_i.

[3]Eigenvalue decomposition on its own is not unique but we have sufficiently many constraints on Γ_X for unqiueness; Γ_X, the eigenvector matrix, is nonnegative and its column-wise sums are one since they are conditional probabilities. See Hu and Schennach (2008) for more.

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*

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

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 variation in the conditional density of X_i given $Z_i = z$ across z 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 the support 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 are 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 $L_{U|D=d,Z}$ that map a function in $\mathcal{L}^1(\mathbb{R})$ to a function in $\mathcal{L}^1(\mathbb{R})$: for $d = 0, 1$,

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

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* $L_{X|U}, L_{X|D=1,Z}$ *and* $L_{X|D=0,Z}$ *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 addtional assumption for the identification. This is because when U_i is discrete and finite, a bijection between u and $\Pr\{X_i = \cdot|U_i = u\}$ needs not be specified. However, when U_i is continuous, we need an ordering on the infinite collection $\{f_{X|U}(\cdot|u)\}_u$ to connect u to $f_{X|U}(\cdot|u)$.

Assumption 5. *(latent rank) There exists a functional M defined on $\mathcal{L}^1(\mathbb{R}^2)$ and a strictly increasing and continuously differentiable function h defined on $\mathcal{U}$ such that*

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

The functional M provides us an ordering on the infinite collection $\{f_{X|U}(\cdot|u)\}_u$, by applying the functional to $\{f_{Y(1)|U}(\cdot|u) \cdot f_{Y(0)|U}(\cdot|u)\}_u$.

Along with Assumptions 1-2, Assumption 5 is a key identifying assumption in the case of continuous U_i. As hinted by its label, Assumption 5 draws the inspiration from the rank invariance assumption in the quantile treatment effect literature. Suppose that Assumption 5 holds true for two functionals $\mathbf{E}[Y_i(1)|U_i = u]$ and $\mathbf{E}[Y_i(0)|U_i = u]$. Then, the two potential outcomes of a given unit have the same 'latent rank' in the sense that their expected values $\mathbf{E}[Y_i(1)|U_i]$ and $\mathbf{E}[Y_i(0)|U_i]$ have the same rank in their respective distributions. A similar motivational assumption can be made with other summary measures such as median or mode. In this sense, I do not assume the rank invariance in its stronger form by assuming that the rank of a unit in terms of their realized potential outcome and that in terms of their counterfactual potential outcome are the same, but assume that there is a latent variable U_i conditioning on which some summary measures applied to the conditional distributions of each potential outcome have the same rank.

Theorem 1 formally states the identification result.

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. □

It directly follows that any functional of the joint distribution of $Y_i(1)$ and $Y_i(0)$ is identified: e.g., $\mathrm{Var}\big(Y_i(1) - Y_i(0)\big), \Pr\{Y_i(1) \geq Y_i(0)\}, \Pr\{Y_i(1) \geq Y_i(0)|Y_i(0)\}$ and etc. Importantly, note that the second set of the assumptions allow U_i to be continuous. The rest of the section discusses the restrictions on the joint distribution of $Y_i(1)$ and $Y_i(0)$ implied by the identifying assumptions and a testable implication of the identifying assumptions which proposes a falsification test.

2.1 Restriction on the joint distribution

Assumption 2 assumes that there exists a latent variable U_i which contains sufficient information on the dependence between a treated potential outcome and an untreated potential outcome. By assuming that U_i is a scalar, the assumption imposes restrictions on the joint distribution of the potential outcomes. Find that for any $\mathcal{Y}, \mathcal{Y}' \subset \mathbb{R}$, we have

$$\Pr\{Y_i(1) \in \mathcal{Y}, Y_i(0) \in \mathcal{Y}'\} = \mathbf{E}\left[\Pr\{Y_i(1) \in \mathcal{Y}|U_i\} \cdot \Pr\{Y_i(0) \in \mathcal{Y}'|U_i\}\right]$$

from the conditional independence part of Assumption 2: $Y_i(1) \perp\!\!\!\perp Y_i(0) \mid U_i$. Then, from the exclusion restriction part of Assumption 2, i.e. $(Y_i(1), Y_i(0)) \perp\!\!\!\perp Z_i \mid U_i$, and the full rank/completeness condition from Assumption 3.b or 4.c, we get

$$\begin{aligned} &\Pr\{Y_i(1) \in \mathcal{Y}|Y_i(0) \in \mathcal{Y}'\} \\ &\quad = \frac{\int_{\mathbb{R}^2} \Pr\{Y_i \in \mathcal{Y}|D_i = 1, Z_i = z\} \cdot \Pr\{Y_i \in \mathcal{Y}'|D_i = 0, Z_i = z'\} w(z, z') dz dz'}{\int_{\mathbb{R}^2} \Pr\{Y_i \in \mathcal{Y}'|D_i = 0, Z_i = z'\} w(z, z') dz dz'} \end{aligned}$$

with some weighting function w. The conditional distribution of $Y_i(1)$ given $Y_i(0) \in \mathcal{Y}'$ is a weighted average of the conditional distributions of Y_i given $(D_i = 1, Z_i)$, where the weights depend on $\mathcal{Y}'$.

The implied linear characterization imposes restrictions on the joint distribution of $Y_i(1)$ and $Y_i(0)$; the conditional distribution of $Y_i(1)$ given $Y_i(0)$ must be spanned from the conditional distribution of Y_i given $(D_i = 1, Z_i)$ and vice versa. For that, the conditional distribution of $Y_i(1)$ given $Y_i(0)$ inherits any linear relationship that holds uniformly across the conditional distributions of Y_i given $(D_i = 1, Z_i = z)$. For example, if there is any $\mathcal{Y} \subset \mathbb{R}$ such that $\Pr\{Y_i \in \mathcal{Y}|D_i = 1, Z_i = z\}$ is constant in z, then the conditional distribution of $Y_i(1)$ given $Y_i(0)$ must satisfy

$$\Pr\{Y_i(1) \in \mathcal{Y}|Y_i(0) \in \mathcal{Y}'\} = \Pr\{Y_i \in \mathcal{Y}|D_i = 1\} \quad \forall \mathcal{Y}'.$$

If there are some $\mathcal{Y}, \tilde{\mathcal{Y}} \subset \mathbb{R}$ such that $\Pr\{Y_i \in \mathcal{Y}|D_i = 1, Z_i = z\} = c \Pr\{Y_i \in \tilde{\mathcal{Y}}|D_i = 1, Z_i = z\}$ for all z, the conditional distribution must satisfy

$$\Pr\{Y_i(1) \in \mathcal{Y} | Y_i(0) \in \mathcal{Y}'\} = c \Pr\left\{Y_i(1) \in \tilde{\mathcal{Y}} | Y_i(0) \in \mathcal{Y}'\right\} \quad \forall \mathcal{Y}'.$$

If there are some $\mathcal{Y}, \tilde{\mathcal{Y}}, \tilde{\mathcal{Y}}' \subset \mathbb{R}$ such that $\Pr\{Y_i \in \mathcal{Y} | D_i = 1, Z_i = z\} + \Pr\{Y_i \in \tilde{\mathcal{Y}} | D_i = 1, Z_i = z\} = \Pr\{Y_i \in \tilde{\mathcal{Y}}' | D_i = 1, Z_i = z\}$ for all z, the conditional distribution must satisfy

$$\Pr\{Y_i(1) \in \mathcal{Y} | Y_i(0) \in \mathcal{Y}'\} + \Pr\left\{Y_i(1) \in \tilde{\mathcal{Y}} | Y_i(0) \in \mathcal{Y}'\right\} = \Pr\left\{Y_i(1) \in \tilde{\mathcal{Y}}' | Y_i(0) \in \mathcal{Y}'\right\} \quad \forall \mathcal{Y}'.$$

In this sense, the point identification result can be interpreted as follows: when Assumptions 1-2 hold, the joint distribution of $Y_i(1)$ and $Y_i(0)$ can be written as a linear combination of the conditional distribution of (Y_i, X_i) given (D_i, Z_i) and the linear coefficients are uniquely decided from the observed distribution of (Y_i, D_i, X_i, Z_i).

2.2 Testable implication

Note that Assumption 5 only assumes a single functional M on the joint density of $(Y_i(1), Y_i(0))$ given U_i. When we assume Assumption 5 for two different functionals, one on $f_{Y(1)|U}$ and another on $f_{Y(0)|U}$, we have a testable implication of Assumptions 1-2 and 4-5, from over-identification. Suppose that $\mathbf{E}[Y_i(1)|U_i = u]$ and $\mathbf{E}[Y_i(0)|U_i = u]$ are strictly increasing in u. Then, the conditional densities $\left(f_{Y(1)|U}, f_{X|U}, f_{U|D=1,Z}\right)$ are identified in the treated subsample and the conditional densities $\left(f_{Y(0)|U}, f_{X|U}, f_{U|D=0,Z}\right)$ are identified in the untreated subsample. Let $f_{X|D=1,U}$ denote the conditional density of X_i given U_i, identified from the treated subsample and likewise for $f_{X|D=0,U}$. Then, Assumption 1 imposes that

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

since $f_{X|D=1,U} = f_{X|D=0,U}$. In (8), a monotone function $\tilde{g}$ is used to connect the identification result from the treated subpopulation to the untreated subpopulation, now that Assumption 5 is assumed separately for each subpopulation. 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 equation (8) is true if the conditional density of (Y_i, X_i) given $(D_i = d, Z_i)$ admits a mixture model where the mixture component densities satisfy

$$f_{Y(d),X|U}(y,x|u) = f_{Y(d)|U}(y|u) \cdot f_{X|U}(x|u).$$

Recall that in Example 1, the two proxy variables are past and future outcomes. Thus, in the panel context, the falsification test can be understood as testing whether there is a latent variable U_i conditioning on which the outcomes are *intertemporally* independent. Note that the key identifying assumption is that the potential outcomes independent *across the treatment status.* While the conditional independence assumption 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 1 tests if the outcomes are intertemporally independent, conditioning on some latent variable.

In the case of discrete U_i, Assumption 5 was not used in the identification. In fact, without introducing any further assumptions, we have a testable implication:

$$\sum_{k=1}^{K} \min_{k'} \sum_{j=1}^{M_X} \left(\Pr\left\{ X_i = x^j \,\middle|\, (D_i, U_i) = \left(1, u^k\right) \right\} - \Pr\left\{ X_i = x^j \,\middle|\, (D_i, U_i) = \left(0, u^k\right) \right\} \right)^2 = 0. \tag{9}$$

I develop an asymptotic theory in the next section under the finite support assumption on U_i, formally proposing a falsification test.

3 Implementation

Based on the identification result for discrete U_i, I estimate the conditional density of $Y_i(1)$ and $Y_i(0)$ given U_i, by assuming a finite support for U_i and solving a nonnegative matrix factorization (NMF) problem. The focus on the case of discrete U_i has several reasons. Firstly, a dicretization is often used in econometric models with latent heterogeneity as an approximation to a continuous latent heterogeneity space: see Bonhomme et al. (2022) for more. Secondly, with parametrization, the estimation of infinite-dimensional objects such as conditional densities $f_{U|D=0,Z}$ and $f_{U|D=1,Z}$ becomes an estimation of finite-dimensional objects Λ_0 and Λ_1, giving us $\sqrt{n}$ rate. The $\sqrt{n}$ rate becomes helpful in deriving an asymptotic distribution for the distributional treatment effect estimators. Lastly, the linearity induced from discretization reduces the computational burden substantially.[4] This does not mean that we do not have a feasible estimation method that utilizes the identifcation result for continuous U_i. For a continuous latent variable case, we can construct a sieve maximum likelihood estimator, as suggested in the nonclassical measurement error literature. The specifics are discussed in the appendix subsection A.2.

The parameters of interest in this paper are the joint distribution of the potential outcomes $Y_i(1)$ and $Y_i(0)$ and the marginal distribution of the treatment effect $Y_i(1) - Y_i(0)$. To estimate these distributional treatment effect (DTE) parameters, I first estimate the conditional probabilites of U_i given Z_i, namely the mixture weight matrices Λ_0 and Λ_1 in the finite mixture interpretation, by solving a nonnegative matrix factorization problem. Given the first step estimators on Λ_0 and Λ_1, I characterize the the joint distribution of $Y_i(1)$ and $Y_i(0)$ and the marginal distribution of $Y_i(1) - Y_i(0)$ as quadratic moments and estimate the distributions by plugging in the first step estimates to the induced U-statistics. In doing so, to account for the estimation error from the first step, I orthogonalize the score function. The Neyman orthogonality makes our estimator robust to the first step estimation error and help derive a limiting distribution for the estimator.

3.1 Nonnegative matrix factorization

To estimate the mixture weight matrices Λ_0 and Λ_1 from (7), I first let $M_Z = K$ by using a partition on $\mathbb{R}$ when the support of Z_i has more than K points and construct sample analogues of the conditional probability matrices $\mathbf{H}_0$ and $\mathbf{H}_1$ defined in the previous section: 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 $\mathbb{H}_0$ is a conditional empirical distribution function of (Y_i, X_i) given $\left(D_i = 0, Z_i\right)$ and each column of $\mathbb{H}_1$ is a conditional empirical distribution function of (Y_i, X_i) given $\left(D_i = 1, Z_i\right)$. As discussed in Section 2, I use partitioning on $\mathbb{R}$ in constructing $\mathbb{H}_0$ and $\mathbb{H}_1$ when any of Y_i, X_i and Z_i is continuous.

To estimate Λ_0 and Λ_1, I formulate a nonnegative matrix factorization problem. Let ι_x be a x-dimensional column vector of ones. Then, the nonnegative matrix factorization problem is

[4] For the empirical dataset used in Section 5, the nonnegative matrix factorization problem with a moderate choice of $K = 5$ can be solved within several minutes on a personal laptop.

constructed as follows:

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

subject to linear constraints that

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

and quadratic constraints that

$$\begin{aligned}&\Pr\left\{(Y_i(d), X_i) = (y, x)|U_i = u^k\right\}\\ &= \left(\sum_{k=1}^{M_X} \Pr\left\{\left(Y_i(d), X_i\right) = \left(y, x^k\right)|U_i = u^k\right\}\right) \cdot \left(\sum_{j=1}^{M_Y} \Pr\left\{(Y_i(d), X_i) = \left(y^j, x\right)|U_i = u^k\right\}\right)\end{aligned} \tag{11}$$

for each (y, x). The linear constraints are probabilities being nonnegative and summing to one. The quadratic constraints are X_i satisfying the exclusion restriction from Assumption 2. When $\mathbb{H}_0$ and $\mathbb{H}_1$ are sufficiently close to $\mathbf{H}_0$ and $\mathbf{H}_1$, the identification result discussed in the previous section says that there is a unique decomposition of $\mathbb{H}_0$ and $\mathbb{H}_1$ which satisfies the linear and the quadratic constraints.

Note that the objective function in (10) is quadratic when we fix either (Λ_0, Λ_1) or (Γ_0, Γ_1). Moreover, Γ_0 and Γ_1 can be further decomposed into three matrices $\Gamma_X, \Gamma_{Y(0)}, \Gamma_{Y(1)}$, each of which correponds to the conditional probability of $X_i, Y_i(0)$ and $Y_i(1)$ given U_i, respectively. Let $\Gamma_d(\cdot,\cdot)$ denote how Γ_X and $\Gamma_{Y(d)}$ recover Γ_d: $\Gamma_d = \Gamma_d\left(\Gamma_X, \Gamma_{Y(d)}\right)$. The quadratic constraints are trivially imposed by optimizing over $\Gamma_X, \Gamma_{Y(0)}$ and $\Gamma_{Y(1)}$. Using these, I propose an iterative algorithm to solve the minimization problem.

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

2. *(Update 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:

$$\left(\Gamma_X^{(s+1)}\right) = \arg\min_{\Gamma_X} \left\|\mathbb{H}_0 - \Gamma_0\left(\Gamma_X, \Gamma_{Y(0)}^{(s)}\right)\Lambda_0^{(s+1)}\right\|_F^2 + \left\|\mathbb{H}_1 - \Gamma_1\left(\Gamma_X, \Gamma_{Y(1)}^{(s)}\right)\Lambda_1^{(s+1)}\right\|_F^2$$

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:

$$\begin{aligned}&\left(\Gamma_{Y(0)}^{(s+1)}, \Gamma_{Y(0)}^{(s+1)}\right)\\ &= \arg\min_{\Gamma_{Y(0)}, \Gamma_{Y(1)}} \left\|\mathbb{H}_0 - \Gamma_0\left(\Gamma_X^{(s+1)}, \Gamma_{Y(0)}\right)\Lambda_0^{(s+1)}\right\|_F^2 + \left\|\mathbb{H}_1 - \Gamma_1\left(\Gamma_X^{(s+1)}, \Gamma_{Y(1)}\right)\Lambda_1^{(s+1)}\right\|_F^2\end{aligned}$$

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 linear constraints, which can be solved with a built-in optimization tool in most statistical softwares. The stepwise optimization assures a convergence to a local minimum. To find the global minimum, I consider various initial values $\left(\Gamma_0^{(0)}, \Gamma_1^{(0)}\right)$.[5]

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 and X_i are discrete, the estimates $\widehat{\Gamma}_0$ and $\widehat{\Gamma}_1$ directly estimate the conditional distribution of $Y_i(1)$ and $Y_i(0)$ given U_i. When Y_i are X_i are continuous and therefore partitioning was used in constructing $\mathbf{H}_0, \mathbf{H}_1$, we use $\widehat{\Lambda}_0$ and $\widehat{\Lambda}_1$ to estimate the distribution of $Y_i(1)$ and $Y_i(0)$ given U_i.

3.2 Distributional treatment effect estimators

Given the estimates of the two mixture weights matrices Λ_0 and Λ_1, I construct an estimator for the joint distribution of $Y_i(1)$ and $Y_i(0)$ and the marginal distribution of $Y_i(1) - Y_i(0)$. Firstly, find that for any $y \in \mathbb{R}$,

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

[5]To initialize $\Gamma_0^{(0)}, \Gamma_1^{(1)}$, I consider columns from $\mathbb{H}_d$ 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.

Since Λ_d is full rank, we have

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

The conditional distribution of $F_{Y(d)|U}(\cdot|u)$ is identified as a linear combination of the observed distributions $\{F_{Y|D=d,Z}(\cdot|z)\}_{z=1}^K$. Building on this, let

$$\tilde{\Lambda}_d = (\Lambda_d)^{-1}$$

for $d = 0, 1$. Let $\tilde{\lambda}_{jk,d}$ denote the j-th row and k-th column component of $\tilde{\Lambda}_d$. $\left(\tilde{\lambda}_{1k,d}, \cdots, \tilde{\lambda}_{Kk,d}\right)^\intercal$, the k-th column of $\tilde{\Lambda}_d$, is a set of linear coefficients on $\{F_{Y|D=d,Z}(\cdot|z)\}_{z=1}^K$ to retrieve the conditional distribution of $Y_i(d)$ given $U_i = u^k$. Using the estimators on Λ_0, Λ_1 from the nonnegative matrix factorization, we estimate the linear coefficients as follows:

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

for $d = 0, 1$.

Secondly, the distribution of U_i is also identified from Λ_0 and Λ_1:

$$\begin{pmatrix} \Pr\{U_i = u^1\} \\ \vdots \\ \Pr\{U_i = u^K\} \end{pmatrix} = \Lambda_0 \begin{pmatrix} \Pr\{D_i = 0, Z_i = z^1\} \\ \vdots \\ \Pr\{D_i = 0, Z_i = z^K\} \end{pmatrix} + \Lambda_1 \begin{pmatrix} \Pr\{D_i = 1, Z_i = z^1\} \\ \vdots \\ \Pr\{D_i = 1, Z_i = z^K\} \end{pmatrix}. \tag{12}$$

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

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

and

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

By combining the two results, we get

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

Using this characterization, I estimate the joint distribution of $Y_i(1)$ and $Y_i(0)$ as a linear combination of $\{F_{Y|D=0,Z}(y|z^j) \cdot F_{Y|D=1,Z}(y'|z^{j'})\}_{j,j'}$ where the weights are computed with $\widehat{\Lambda}_0, \widehat{\Lambda}_1$ and $\{\hat{p}_{D,Z}(d,j)\}_{d,j}$. We can derive a similar result for the marginal distribution of $Y_i(1) - Y_i(0)$: for any $\delta \in \mathbb{R}$,

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

Both parameters of interest are identified as a weighted sum of quantities that are indexed by pairs of subpopulations $\{i : D_i = 0, Z_i = z^j\}$ and $\{i : D_i = 1, Z_i = z^{j'}\}$. As shown above, weights are estimated from the first step nonnegative matrix factorization and empirical measures of the subpopulations. It remains to estimate the quantities associated with each pair of subpopulations. I will discuss this for the marginal distribution of $Y_i(1) - Y_i(0)$; the case for the joint distribution of $Y_i(1)$ and $Y_i(0)$ follows naturally. For some δ, let

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

Firstly, find that θ is a summation over K treated subpopulations and K untreated subpopulations. Fix j, j' and find

$$
\begin{aligned}
\theta_{jj'} &:= \left(\sum_{k=1}^{K} p_U(k) \tilde{\lambda}_{jk,0} \tilde{\lambda}_{j'k,1} \right) \int_{\mathbb{R}} F_{Y|D=1,U}(y+\delta|z^j) \cdot f_{Y|D=0,U}(y|z^{j'}) dy \\
&= \left(\sum_{k=1}^{K} p_U(k) \tilde{\lambda}_{jk,0} \tilde{\lambda}_{j'k,1} \right) \frac{\mathbf{E}\left[\mathbf{1}\{Y_{i'} \le Y_i + \delta, D_i = 0, Z_i = z^j, D_i = 1, Z_{i'} = z^{j'}\} \right]}{\mathbf{E}\left[\mathbf{1}\{D_i = 0, Z_i = z^j, D_{i'} = 1, Z_{i'} = z^{j'}\} \right]}
\end{aligned}
$$

with $(Y_i, D_i, Z_i) \perp\!\!\!\perp (Y_{i'}, D_{i'}, Z_{i'})$. Thus, $\theta_{jj'}$ is identfied from a quadratic moment

$$\mathbf{E}\left[m_{jj'}\left(W_i, W_{i'}; \theta_{jj'}, \tilde{\Lambda}_0, \tilde{\Lambda}_1, \{p_U(k)\}_k, \{p_{D,Z}(d,j)\}_{d,j}\right)\right] = 0$$

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

$$\begin{aligned}
& m_{jj'}\left(W_i, W_{i'}; \theta_{jj'}, \tilde{\Lambda}_0, \tilde{\Lambda}_1, \{p_U(k)\}_k, \{p_{D,Z}(d,j)\}_{d,j}\right) \\
&= \frac{\sum_{k=1}^K p_U(k)\tilde{\lambda}_{jk,0}\tilde{\lambda}_{j'k,1}}{p_{D,Z}(0,j)\cdot p_{D,Z}(1,j')} \cdot \left(\frac{1}{2}\mathbf{1}\{Y_{i'} \le Y_i + \delta, D_i = 0, Z_i = z^j, D_i = 1, Z_{i'} = z^{j'}\}\right. \\
&\qquad\qquad\qquad\qquad \left. + \frac{1}{2}\mathbf{1}\{Y_i \le Y_{i'} + \delta, D_i = 1, Z_i = z^{j'}, D_i = 0, Z_{i'} = z^j\}\right) - \theta_{jj'}.
\end{aligned}$$

By summing over j and j', we can construct a moment function $m = \sum_{j=1}^K \sum_{j'=1}^K m_{jj'}$ such that

$$\mathbf{E}\left[m\left(W_i, W_{i'}; \theta, \tilde{\Lambda}_0, \tilde{\Lambda}_1, \{p_U(k)\}_k, \{p_{D,Z}(d,j)\}_{d,j}\right)\right] = 0$$

identifies θ.

If the nuisance parameters $\tilde{\Lambda}_0, \tilde{\Lambda}_1, p_U, p_{D,Z}$ were known, the standard asymptotic theory of U statistic would apply to the GMM estimator of θ using $\mathbf{E}[m(W_i, W_{i'}; \theta)] = 0$ as the moment condition. However, in practice, we use first step estimates for the nuisance parameters. Thus, to account for the first step estimation error, we 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 matrix $\mathbb{H}_0$ and $\mathbb{H}_1$, $\left(\tilde{\Lambda}_0, \tilde{\Lambda}_1\right)$ satisfy the following equations at their true values:

$$\begin{aligned}
\sum_{j=1}^K \tilde{\lambda}_{jk,d} \Pr\left\{Y_i = y, X_i = x | Z_i = z^j\right\} = & \left(\sum_{j=1}^K \tilde{\lambda}_{jk,d} \Pr\left\{Y_i = y | Z_i = z^j\right\}\right) \\
& \cdot \left(\sum_{j=1}^K \tilde{\lambda}_{jk,d} \Pr\left\{X_i = x | Z_i = z^j\right\}\right) \quad \forall y, d, x, k
\end{aligned} \tag{13}$$

$$\Pr\{X_i = x\} = \sum_{k=1}^K p_U(k) \sum_{j=1}^K \tilde{\lambda}_{jk,d} \Pr\left\{X_i = x | D_i = d, Z_i = z^j\right\} \quad \forall d, x. \tag{14}$$

Equation (13) corresponds to the conditional independence assumption that

$$\Pr\{Y_i(d) = y, X_i = x | U_i = u\} = \Pr\{Y_i(d) = y | U_i = u\} \cdot \Pr\{X_i = x | U_i = u\}.$$

and Equation (14) corresponds to the law of iterated expectation that

$$\Pr\{X_i = x\} = \sum_{k=1}^{K} p_U(k) \Pr\{X_i = x | U_i = u^k\}.$$

Given $\{p_{D,Z}(d,j)\}_{d,j}$, Equation (13) can be written as a quadratic moment condition and Equation (14) as a linear moment condition. I use these additional moments in orthogonalizing the moment m so that the Neyman orthogonality holds.

Let $\tilde{\lambda}$ and p denote vectorizations of $\left(\tilde{\Lambda}_0, \tilde{\Lambda}_1\right)$ and $(\{p_U(k)\}_k, \{p_{D,Z}(d,j)\}_{d,j})$. The orthogonalized score is constructed with the additional moment function

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

To complete the orthogonalization argument, I show that the Jacobian matrix of ϕ has full rank.

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

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

has a full rank.

Proof. See Appendix. □

Then, we can construct an additional nuisance parameter

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

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

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

and

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

satisfies the Neyman orthogonality. μ is estimated by taking a sample analogue of the expression above. Given estimators $\left(\hat{\tilde{\lambda}}, \hat{p}, \hat{\mu}\right)$, I estimate θ with

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

$\widehat{F}_{Y(0),Y(1)}$ and $\widehat{F}_{Y(1)-Y(0)}$ denote the distributional treatment effect estimators we obtain from this two-step procedure.

3.3 Asymptotic properties

Theorem 2 establishes the consistency of the mixture weight estimators $\hat{\Lambda}_0$ and $\hat{\Lambda}_1$.

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

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

as $n \to \infty$.

Proof. See Appendix. $\square$

A direct corollary of Theorem 2 is that $\widehat{\tilde{\Lambda}}_0, \widehat{\tilde{\Lambda}}_1$ are consistent for $\tilde{\Lambda}_0$ and $\tilde{\Lambda}_1$ at the rate of $\frac{1}{\sqrt{n}}$. Theorem 3 establishes the asymptotic normality of the distributional treatment effect estimators.

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

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

as $n \to \infty$.

Proof. See Appendix. □

The asymptotic variance is computed from a projection of the orthogonal scores. Fix some distributional treatment effect parameter: $F_{Y(0),Y(1)}(y,y')$ or $F_{Y(1)-Y(0)}(\delta)$. Then, the asymptotic variance, $\sigma(y,y')^2$ or $\sigma(\delta)^2$, is computed from

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

4 Simulation

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

[6] The specifics of the DGPs are as follows: $p_U = (0.286, 0.286, 0.438)$,

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

and Λs in the order of decreasing smallest singular value are

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

Table 1 contains the bias and the root mean squared error (rMSE) of the distributional treatment effect estimators $\widehat{F}_{Y(1)-Y(0)}$. As Λ becomes less informative about the distribution of U_i, i.e. the smallest singular value $\sigma_{\min}(\Lambda)$ decreases, the rMSE goes up. This suggests that the first step nonnegative matrix factorization estimation quality depends on how informative the proxy variables X_i and Z_i are for the latent variable U_i. Additionally, Table 2 contains the coverage probability of the confidence interval constructed with the asymptotic standard error and the type I error of the falsification test proposed in Subsection 2.2. The 95% confidence interval shows mostly correct coverage, sometimes slightly too conservative, and the falsification test is valid.

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

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

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

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

5 Empirical illustration

In this section, we revisit Jones et al. (2019) and estimate the distributional treatment effect of workplace wellness program on medical spending. Jointly with the Campus Well-being Services at the University of Illinois Urbana-Champaign, the authors of Jones et al. (2019) conducted a large-scale randomized control trials. The experiment started in July 2016, by inviting 12,459 eligible university employees to participate in an online survey. Of 4,834 employees who completed the survey, 3,300 employees were randomly selected into treatment, being offered to participate in a workplace wellness program names iThrive. The participation itself was not enforced; the treated individuals were merely financially incentivized to participate by being offered monetary reward for completing each step of the wellness program. Thus, the main treatment effect parameter of Jones et al. (2019) is the 'intent-to-treat' effect. The workplace wellness program consisted of various activities such as chronic disease management, weight management, and etc. The treated individuals were offered to participate in the wellness program starting the fall semester of 2016, until the spring semester of 2018.

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

We build onto this ATE result from Jones et al. (2019) and estimate the distributional treatment effect of the randomly assigned eligibility for the wellness program. The dataset built by the authors of Jones et al. (2019) fits the context of the short panel model in Example 1. For each individual, the dataset contains monthly medical spending records for the following three time durations: July 2015-July 2016, August 2016-July 2017 and August 2017-January 2019. Since the experiment started in the summer of 2016 and the treated individuals were offered to participate in the wellness program starting the fall semester of 2016, the monthly medical spending record for July 2015-July 2016 could be thought of as a 'pretreatment' outcome variable. Thus, we could use the information from the distribution of the pretreatment outcome variable to connect the treated subsample and the untreated subsample. The followings are the variables taken from the dataset.

Y_i : monthly medical spending for August 2016-July 2017

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

X_i : monthly medical spending for July 2015-July 2016

Z_i : monthly medical spending for August 2017-January 2019

In this specific empirical context, the common shock V_{it} could be thought of as underlying health status and the treatment-status-specific shocks $(\varepsilon_{it}(1), \varepsilon_{it}(0))$ could be thought of as additional random shocks such as susceptibility to the workplace wellness program or transient health shock which does not persist over time.

Before applying the DTE estimators to the dataset, I implemented the falsification test with $K = 5$.[7] The test statistic is computed with a 25×1 vector

$$W_n = \begin{pmatrix} \Pr\{X_i \leq \widehat{F_X^{-1}(0.2)|D_i} = 1, U_i = u^1\} - \Pr\{X_i \leq \widehat{F_X^{-1}(0.2)|D_i} = 0, U_i = u^1\} \\ \vdots \\ \Pr\{X_i > \widehat{F_X^{-1}(0.8)|D_i} = 1, U_i = u^5\} - \Pr\{X_i > \widehat{F_X^{-1}(0.8)|D_i} = 0, U_i = u^5\} \end{pmatrix}.$$

Theorem 3 can be easily extended to the marginal distribution of X_i as well and therefore we test the null (9) with

$$T_n = nW_n{}^{\intercal} Avar(W)^{-1} W_n,$$

from $\sqrt{n}W_n$ being asymptotically normal. In the dataset, T_n was 16.435 and its p-value was 0.901.

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

[7]When constructing $\mathbb{H}_0$ and $\mathbb{H}_1$ to be used in the first step nonnegative matrix factorization, we used the quintiles of the marginal distributions: $\left(-\infty, F_Y^{-1}(0.2), F_Y^{-1}(0.4), F_Y^{-1}(0.6), F_Y^{-1}(0.8), \infty\right)$ and so on. Thus, the matrices $\mathbb{H}_0$ and $\mathbb{H}_1$ were 25×5 matrices.

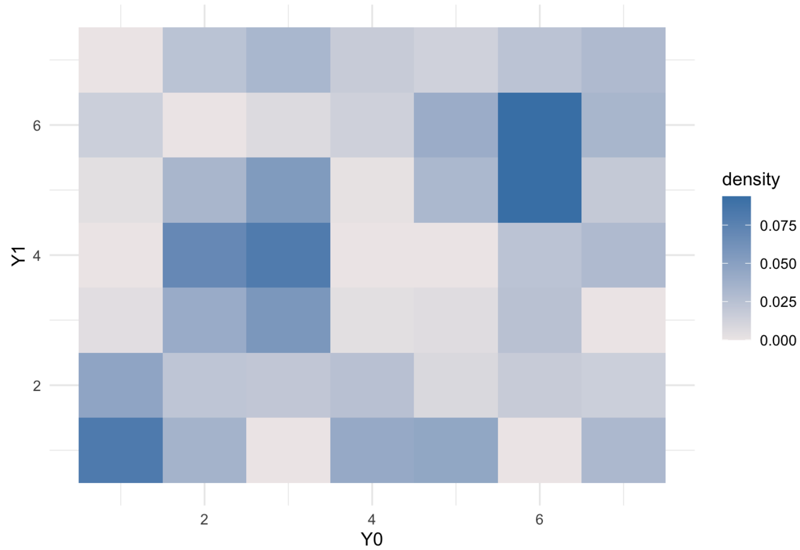

Figure 1: Joint density of $F_Y\big(Y_i(1)\big)$ and $F_Y\big(Y_i(0)\big)$, $K = 5$.

workplace wellness program must be limited.

Figure 2 contains the estimated marginal distribution of the treatment effect and its 95% pointwise confidence interval. Note that the point estimates are mostly upward-sloping and lie between zero and one. Though the quadratic moment representation used in the DTE estimators does not impose any monotoncity or nonnegativity restrictions, the estimated marginal distribution violates these constraints only on a small subset of the range $[-1000, 1000]$. Overall, it is unclear if

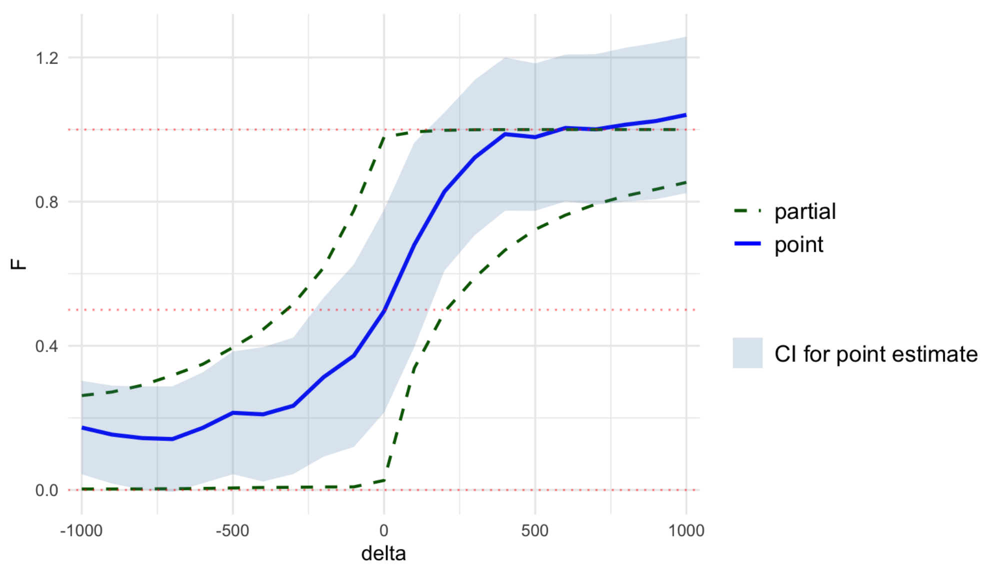

Figure 2: Marginal distribution of $Y_i(1) - Y_i(0)$, $K = 5$.

more than half of the people would be better off from the treatment; the confidence interval for $\Pr\{Y_i(1) - Y_i(0) \geq 0\}$ contains 0.5, not being able to reject the null $\Pr\{Y_i(1) - Y_i(0) \geq 0\} \leq 0.5$.

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

Lastly, the point identification helps us analyze the pattern of the treatment heterogeneity. Recall that the ATE estimate was inconclusive about the effectiveness of the treatment. However, the DTE estimates on $\Pr\{Y_i(1) - Y_i(0) \leq \delta\}$ for $\delta \leq -600$ and the DTE estimates on $\Pr\{Y_i(1) - Y_i(0) \leq \delta\}$ for $\delta \geq 400$ shows us interesting treatment effect heterogeneity patterns, in favor of implementing the treatment. The negative impact of the treatment, i.e. how much more money you spend under the treatment, is capped at \$400: $\widehat{F}_{Y(1)-Y(0)}(400) \approx 1$. On the other hand, the left tail of the treatment effect distribution is thicker, implying that some people are greatly benefitted from participating in the program: $\widehat{F}_{Y(1)-Y(0)}(-600) \approx 0.15$.

6 Conclusion

This paper presents an identification result for the joint distribution of treated potential outcome and untreated potential outcome, given conditionally random binary treatment. The key assumptions in the identification are that there exists a latent variable that captures the dependence between the two potential outcomes and that there exist two proxy variables for the latent variable. By assuming strict monotonicity for some functional of the conditional distribution of potential outcomes given the latent variable, I interpret the latent variable as 'latent rank' and strict monotonicty as 'latent rank invariance.' In implementation, I propose a first step nonnegative matrix factorization and a second step plug-in GMM. $\sqrt{n}$-consistency of the first-step estimator and the asymptotic normality of the second step GMM estimator are proven. Lastly, I apply the estimation method to revisit Jones et al. (2019) and find that the potential medical spendings are positively correlated at the two ends of the support and the marginal distribution of the treatment effect has thicker left tail.

References

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

Bonhomme, Stéphane, Thibaut Lamadon, and Elena Manresa, "Discretizing unobserved heterogeneity," *Econometrica*, 2022, *90* (2), 625–643.

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

Carneiro, Pedro, 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*, 2003, *44* (2), 361–422.

Chernozhukov, Victor and Christian Hansen, "An IV model of quantile treatment effects," *Econometrica*, 2005, *73* (1), 245–261.

Chernozhukov, Victor and Christian Hansen, "Instrumental quantile regression inference for structural and treatment effect models," *Journal of Econometrics*, 2006, *132* (2), 491–525.

Deaner, Ben, "Proxy controls and panel data," 2023.

Fan, Yanqin and Sang Soo Park, "Sharp bounds on the distribution of treatment effects and their statistical inference," *Econometric Theory*, 2010, *26* (3), 931–951.

Fan, Yanqin, Robert Sherman, and Matthew Shum, "Identifying treatment effects under data combination," *Econometrica*, 2014, *82* (2), 811–822.

Firpo, Sergio and Cristine Pinto, "Identification and estimation of distributional impacts of interventions using changes in inequality measures," *Journal of Applied Econometrics*, 2016, *31* (3), 457–486.

Firpo, Sergio and Geert Ridder, "Partial identification of the treatment effect distribution and its functionals," *Journal of Econometrics*, 2019, *213* (1), 210–234.

Frandsen, Brigham R and Lars J Lefgren, "Partial identification of the distribution of treatment effects with an application to the Knowledge is Power Program (KIPP)," *Quantitative Economics*, 2021, *12* (1), 143–171.

Gautier, Eric and Stefan Hoderlein, "A triangular treatment effect model with random coefficients in the selection equation," 2015.

Han, Sukjin and Haiqing Xu, "On Quantile Treatment Effects, Rank Similarity, and Variation of Instrumental Variables," *arXiv preprint arXiv:2311.15871*, 2023.

Heckman, James J, 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*, 1997, *64* (4), 487–535.

Hong, YP and C-T Pan, "A lower bound for the smallest singular value," *Linear Algebra and its Applications*, 1992, *172*, 27–32.

Hu, Yingyao, "Identification and estimation of nonlinear models with misclassification error using instrumental variables: A general solution," *Journal of Econometrics*, 2008, *144* (1), 27–61.

Hu, Yingyao and Susanne M Schennach, "Instrumental variable treatment of nonclassical measurement error models," *Econometrica*, 2008, *76* (1), 195–216.

Jones, Damon, David Molitor, and Julian Reif, "What do workplace wellness programs do? Evidence from the Illinois workplace wellness study," *The Quarterly Journal of Economics*, 2019, *134* (4), 1747–1791.

Kedagni, Desire, "Identifying treatment effects in the presence of confounded types," *Journal of Econometrics*, 2023, *234* (2), 479–511.

Makarov, GD, "Estimates for the distribution function of a sum of two random variables when the marginal distributions are fixed," *Theory of Probability & its Applications*, 1982, *26* (4), 803–806.

Miao, Wang, Zhi Geng, and Eric J Tchetgen Tchetgen, "Identifying causal effects with proxy variables of an unmeasured confounder," *Biometrika*, 2018, *105* (4), 987–993.

Nagasawa, Kenichi, "Treatment effect estimation with noisy conditioning variables," *arXiv preprint arXiv:1811.00667*, 2022.

Noh, Sungho, "Nonparametric identification and estimation of heterogeneous causal effects under conditional independence," *Econometric Reviews*, 2023, *42* (3), 307–341.

Vuong, Quang and Haiqing Xu, "Counterfactual mapping and individual treatment effects in nonseparable models with binary endogeneity," *Quantitative Economics*, 2017, *8* (2), 589–610.

APPENDIX

A Discussion on a continuous latent variable

A.1 Identification

Assumptions 1-2 are powerful enough for us to apply the known spectral decomposition results with proxy variables (see Hu (2008); Hu and Schennach (2008) and more) to each of the treated subsample and the untreated subsample. Let $f_{Y=y,X|D=d,Z}(x|z)$ denote the conditional density of (Y_i, X_i) given (D_i, Z_i) evaluated at $Y_i = y$ and $D_i = d$; the density has only two arguments x and z. Likewise, let $f_{U|D=d,Z}$ denote the conditional density of U_i given (D_i, Z_i) evaluated at $D_i = d$. From Assumptions 1-2, we obtain the following integral representation: for $x, z \in \mathbb{R}$,

$$\begin{aligned} f_{Y=y,X|D=d,Z}(x|z) &= \int_{\mathcal{U}} f_{Y(d),X|D=d,Z,U}(y,x|z,u) \cdot f_{U|D=d,Z}(u|z)du \\ &= \int_{\mathcal{U}} f_{Y(d),X|Z,U}(y,x|z,u) \cdot f_{U|D=d,Z}(u|z)du \quad \because \text{ Assumption 1} \\ &= \int_{\mathcal{U}} f_{Y(d)|U}(y|u) \cdot f_{X|U}(x|u) \cdot f_{U|D=d,Z}(u|z)du \quad \because \text{ Assumption 2} \qquad (15) \\ f_{X|D=d,Z}(x|z) &= \int f_{X|U}(x|u) \cdot f_{U|D=d,Z}(u|z)du. \end{aligned}$$

To discuss the spectral decomposition result of Hu and Schennach (2008), let us construct integral operators $L_{X|U}$, $L_{U|D=d,Z}$ and a diagonal operator $\Delta_{Y(d)=y|U}$ which map a function in $\mathcal{L}^1(\mathbb{R})$ to a function in $\mathcal{L}^1(\mathbb{R})$:

$$\begin{aligned} \left[L_{X|U}g\right](x) &= \int_{\mathbb{R}} f_{X|U}(x|u)g(u)du, \\ \left[L_{U|D=d,Z}g\right](u) &= \int_{\mathbb{R}} f_{U|D=d,Z}(u|z)g(z)dz, \\ \left[\Delta_{Y(1)=y|U}g\right](u) &= f_{Y(1)|U}(y|u)g(u). \end{aligned}$$

For example, when g is a density, $L_{X|U}$ takes the density g as a marginal density of U_i and maps it to a marginal density of X_i, implied by $f_{X|U}$ and g. Define $L_{Y=y,X|D=d,Z}$ and $L_{X|D=d,Z}$ similarly, with the conditional density $f_{Y=y,X|D=d,X}$ and $f_{X|D=d,Z}$. Then,

$$L_{Y=y,X|D=d,Z} = L_{X|U} \cdot \Delta_{Y(d)|U} \cdot L_{U|D=d,Z},$$
$$L_{X|D=d,Z} = L_{X|U} \cdot L_{U|D=d,Z}.$$

To get to a spectral decomposition result, we additionally assume that the conditional density $f_{X|D=d,Z}$ is complete. The completeness assumption imposes restriction on the proxy variables X_i and Z_i; the conditional density of U_i given Z_i, within each subsample, should preserve the variation in the conditional density of X_i given U_i. With completeness condition on the conditional density $f_{X|D=d,Z}$, we can define an inverse of the integral operator $L_{X|D=d,Z}$ and therefore obtain a spectral decomposition:

$$L_{Y=y,X|D=d,Z} \cdot \left(L_{X|D=d,Z}\right)^{-1} = L_{X|U} \cdot \Delta_{Y(d)=y|U} \cdot \left(L_{X|U}\right)^{-1}.$$

The RHS of the equation above admits a spectral decomposition with $\left\{f_{X|U}(\cdot|u)\right\}_u$ as eigenfunctions and $\left\{f_{Y(d)|U}(y|u)\right\}_u$ as eigenvalues.

However, the individual spectral decomposition results on the two subsamples by themselves are not enough to identify the joint distribution of the potential outcomes. To connect the two spectral decomposition results, we resort to Assumption 1. Under Assumption 1, the conditional density of X_i given U_i is identical across the two subsamples. Thus, the two decomposition results should admit the same density functions $\left\{f_{X|U}(\cdot|u)\right\}_u$ as eigenfunctions. Using this, we connect the eigenvalues of the two decompositions; we identify $\left\{f_{Y(1)|U}(\cdot|u) \cdot f_{Y(0)|U}(\cdot|u)\right\}_u$.

Lastly, to find the marginal distribution of U_i, we fully invoke the latent rank interpretation and assume that there is some functional M defined on $\mathcal{L}^1(\mathbb{R}^2)$ such that $Mf_{Y(1),Y(0)|U}(\cdot,\cdot|u)$ is strictly increasing in u. An example of such a functional is the sum of the means:

$$Mf = \int_{\mathbb{R}}\int_{\mathbb{R}}(y_1 + y_0)f(y_1, y_0)dy_1dy_0.$$

Alternatively, if two functionals such as

$$M_1f = \int_{\mathbb{R}} y_1 f_{Y(1)|U}(y_1|u)dy_1,$$
$$M_2f = \int_{\mathbb{R}} y_0 f_{Y(0)|U}(y_0|u)dy_0$$

satisfy that $M_1 f_{Y(1)|U}(\cdot|u)$ and $M_2 f_{Y(0)|U}(\cdot|u)$ are both strictly increasing in u, the latent rank invariance holds in a truer sense that U_i determines the rank of $\mathbf{E}\left[Y_i(1)|U_i\right]$ and the rank of $\mathbf{E}\left[Y_i(0)|U_i\right]$ and that the two ranks coincide. The latent rank assumption finds on ordering on the eigenfunctions $\left\{f_{X|U}(\cdot|u)\right\}_u$ using information from $\left\{f_{Y(1),Y(0)|U}(\cdot,\cdot|u)\right\}_u$ and allows us to use a transformation on U_i without precisely locating U_i.

A.2 Sieve maximum likelihood

To estimate the conditional densities of interest, i.e. $f_{Y(1)|U}, f_{Y(0)|U}, f_{X|U}, f_{U|D=1,Z}, f_{U|D=0,Z}$, we again utilize the decomposition given in (6). Especially, with U_i being a continuous random variable, the decomposition can be rewritten as an integration:

$$f_{Y,X|D,Z}(y,x|d,z) = \int_{\mathcal{U}} f_{Y(d)|U}(y|u) \cdot f_{X|U}(x|u) \cdot f_{U|D=d,Z}(u|z)du.$$

Given some sieves to approximate the conditional densities, characterized with finite-dimensional parameters $\theta = \left(\theta_1, \theta_0, \theta_X, \theta_{1Z}, \theta_{0Z}\right)$, the sieve ML estimator is:

$$\begin{aligned}\hat{\theta} &= \arg\max_{\theta\in\Theta_n} \sum_{i=1}^{n} \log f_{Y,X|D,Z,n}(Y_i, X_i|D_i, Z_i; \theta) \qquad (16)\\ &= \arg\max_{\theta\in\Theta_n} \sum_{i=1}^{n} \Bigg(D_i \log \int_{\mathcal{U}} f_{Y(1)|U,n}(Y_i|u;\theta_1) \cdot f_{X|U,n}(X_i|u;\theta_X) \cdot f_{U|D=1,Z,n}(u|Z_i;\theta_{1Z})du \\ &\qquad (1-D_i) \log \int_{\mathcal{U}} f_{Y(0)|U,n}(Y_i|u;\theta_0) \cdot f_{X|U,n}(X_i|u;\theta_X) \cdot f_{U|D=0,Z,n}(u|Z_i;\theta_{0Z})du \Bigg).\end{aligned}$$

In particular, we propose tensor product spaces of Bernstein polynomials as sieves $\{\Theta_n\}_{n=1}^{\infty}$. For example, the conditional density $f_{Y(1)|U}$ approximated to a tensor product space with a given dimension of $\left(p^y+1, p^u+1\right)$ is as follows: with y normalized to be on $[0,1]$,

$$f_{Y(1)|U,n}(y|u;\theta_1) = \sum_{j=0}^{p^y} \sum_{k=0}^{p^u} \theta_{jk,1} \binom{p^y}{j} y^j (1-y)^{p^y-j} \cdot \binom{p^u}{k} u^k (1-u)^{p^u-k}$$

and $\theta_1 = \{\theta_{jk,1}\}_{0\le j\le p^j, 0\le k\le p^u}$.[8] The tensor product construction and the properties of Bernstein polynomials make it remarkably straightforward to impose that the approximated functions are densities. Using properties of Berstein polynomials, we can impose that $f_{Y(1)|U,n}(y|u;\theta_1)$ is non-

[8] The degree of Bernstein polynomial does not need to be uniform across different conditional densities; for example p^y for $f_{Y(1)|U,n}$ may differ from p^y for $f_{Y(0)|U,n}$. However, p^u being uniform across all five conditional densities facilitates computation.

negative and integrate to one, by imposing that

$$\theta_{jk,1} \geq 0 \quad \forall j, k \qquad \textit{(nonnegative)}$$

$$\sum_{j=0}^{p^y} \frac{\theta_{j0,1}}{p^y + 1} = 1 \qquad \textit{(sum-to-one)}$$

$$\sum_{l=0}^{k} \sum_{j=0}^{p^y} \frac{1}{p^y + 1} (-1)^{k-l} \binom{p^u}{k} \binom{k}{l} \theta_{jl,1} = 0 \quad \forall k = 1, \cdots, p^u \qquad \textit{(sum-to-one)}$$

Moreover, when the latent rank interpretation from Assumption 5 is assumed with average, the monotonicity condition can be easily imposed as linear constraints. For example, $\mathbf{E}\left[Y_i(1)|U_i = u\right]$ being monotone increasing in u translates to

$$\sum_{j=0}^{p^y} w_j \theta_{jk,1} \leq \sum_{j=0}^{p^y} w_j \theta_{jk+1,1} \quad \forall k = 0, \cdots, p^u - 1 \qquad \textit{(monotonicity)}$$

Below are the details on the linear constraints that correspond to nonnegativity, sum-to-one and monotonicity. Use the same example from before—$f_{Y(1)|U,n}$—and find that we can rearrange the approximated function as a univariate Bernstein polynomial of degree p^u by fixing u:

$$f_{Y(1)|U,n}(y|u; \theta_1) = \sum_{j=0}^{p^y} \left(\sum_{k=0}^{p^u} \theta_{jk,1} \binom{p^u}{k} u^k (1-u)^{p^u - k} \right) \binom{p^y}{j} y^j (1-y)^{p^y - j}.$$

$f_{Y(1)|U,n}(y|u; \theta_1)$ is nonnegative if and only if

$$\sum_{k=0}^{p^u} \theta_{jk,1} \binom{p^u}{k} u^k (1-u)^{p^u - k} \geq 0$$

for every $j = 0, \cdots, p^y$ at the fixed u. Since $f_{Y(1)|U,n}(y|u; \theta_1)$ needs to be a nonnegative function at any value of u, this translates to $\sum_{k=0}^{p^u} \theta_{jk,1} \binom{p^u}{k} u^k (1-u)^{p^u-k}$, which is a Bernstein polynomial itself, being a nonnegative function. Thus, the nonnegativity constraints become

$$\theta_{jk,1} \geq 0 \quad \forall j, k.$$

Also, find that

$$\int_0^1 f_{Y(1)|U,n}(y|u;\theta_1)dy = \sum_{k=0}^{p^u}\left(\sum_{j=0}^{p^y}\theta_{jk,1}\int_0^1\sum_{j=0}^{p^y}\binom{p^y}{j}y^j(1-y)^{p^y-j}dy\right)\binom{p^u}{k}u^k(1-u)^{p^u-k}$$
$$= \sum_{k=0}^{p^u}\sum_{j=0}^{p^y}\frac{\theta_{jk,1}}{p^y+1}\binom{p^u}{k}u^k(1-u)^{p^u-k}.$$

For $\int_0^1 f_{Y(1)|U,n}(y|u;\theta_1)dy = 1$ to hold uniformly over u, $\sum_{k=0}^{p^u}\sum_{j=0}^{p^y}\frac{\theta_{jk,1}}{p^y+1}\binom{p^u}{k}u^k(1-u)^{p^u-k}$ must be constant in u and equal to one. Again, $\sum_{k=0}^{p^u}\sum_{j=0}^{p^y}\frac{\theta_{jk,1}}{p^y+1}\binom{p^u}{k}u^k(1-u)^{p^u-k}$ is a Bernstein polynomial itself and can be transformed to a sum of monomials:

$$\binom{p^u}{l}u^l(1-u)^{p^u-l} = \sum_{k=l}^{p^u}(-1)^{k-l}\binom{p^u}{k}\binom{k}{l}u^k$$
$$\sum_{l=0}^{p^u}\sum_{j=0}^{p^y}\frac{\theta_{jl,1}}{p^y+1}\binom{p^u}{l}u^l(1-u)^{p^u-l} = \sum_{l=0}^{p^u}\sum_{j=0}^{p^y}\frac{\theta_{jl,1}}{p^y+1}\sum_{k=l}^{p^u}(-1)^{k-l}\binom{p^u}{k}\binom{k}{l}u^k$$
$$= \sum_{k=0}^{p^u}\left(\sum_{l=0}^{k}\sum_{j=0}^{p^y}\frac{\theta_{jl,1}}{p^y+1}(-1)^{k-l}\binom{p^u}{k}\binom{k}{l}\right)u^k$$

Thus, the sum-to-one constraints are

$$\sum_{j=0}^{p^y}\frac{\theta_{j0,1}}{p^y+1} = 1,$$
$$\sum_{l=0}^{k}\sum_{j=0}^{p^y}\frac{1}{p^y+1}(-1)^{k-l}\binom{p^u}{k}\binom{k}{l}\theta_{jl,1} = 0 \quad \forall k = 1,\cdots,p^u.$$

Lastly, for the monotonicity constraint, find that

$$\int_0^1 y f_{Y(1)|U,n}(y|u;\theta_1)dy = \sum_{k=0}^{p^u}\underbrace{\left(\sum_{j=0}^{p^y}\theta_{jk,1}\int_0^1\binom{p^y}{j}y^{j+1}(1-y)^{p^y-j}dy\right)}_{=:\theta_{\cdot k,1}}\binom{p^u}{k}u^k(1-u)^{p^u-k}$$

Again, the conditional expectation is also a Berstein polynomial and it is monotone increasing if and only if $\theta_{\cdot k,1} \leq \theta_{\cdot k+1,1}$ for $k = 0,\cdots,p^u-1$. By applying the monomial transformation again, we get

$$\binom{p^y}{j} y^{j+1}(1-y)^{p^y-j} = \binom{p^y}{j}\binom{p^y+1}{j+1}^{-1} \sum_{l=j+1}^{p^y+1} (-1)^{l-j-l} \binom{p^y+1}{j+1}\binom{j+1}{l} u^l,$$

$$\int_0^1 \binom{p^y}{j} y^{j+1}(1-y)^{p^y-j} dy = \frac{j+1}{p^y+1} \sum_{l=j+1}^{p^y+1} (-1)^{l-j-l} \binom{p^y+1}{j+1}\binom{j+1}{l} \frac{1}{l+1} =: w_j.$$

The monotonicity constraints are

$$\sum_{j=0}^{p^y} w_j \theta_{jk,1} \leq \sum_{j=0}^{p^y} w_j \theta_{jk+1,1} \quad \forall k = 0, \cdots, p^u - 1.$$

Now, we discuss how to estimate the distributional treatment effect parameters. Unlike the nonnegative matrix factorization estimator, the sieve ML estimator fully estimates the five conditional densities. Thus, an estimator on the joint distribution of the potential outcomes and the marginal distribution of treatment effect can be directly constructed from $\hat{\theta}$. For example, the joint density estimator can be constructed as follows: for any $\left(y, y'\right)$,

$$\begin{aligned}\widehat{F}_{Y(1),Y(0)}\left(y, y'\right) = \frac{1}{n}\sum_{i=1}^{n} \int_{\mathcal{U}} \int_{-\infty}^{y} \int_{-\infty}^{y'} & f_{Y(1)|U,n}\left(w|u;\hat{\theta}_1\right) \cdot f_{Y(0)|U}\left(w'|u;\hat{\theta}_0\right) dw dw' \\ & \cdot \left(D_i f_{U|D=1,Z,n}\left(u|Z_i;\hat{\theta}_{1Z}\right) + (1-D_i) f_{U|D=0,Z,n}\left(u|Z_i;\hat{\theta}_{0Z}\right)\right) du.\end{aligned}$$

Likewise, the marginal treatment effect distribution estimator can be constructed as follows: for any δ,

$$\begin{aligned}\widehat{F}_{Y(1)-Y(0)}(\delta) = \frac{1}{n}\sum_{i=1}^{n} \int_{\mathcal{U}} \int_{\mathbb{R}} \int_{-\infty}^{y+\delta} & f_{Y(1)|U}(y'|u;\hat{\theta}_1) \cdot f_{Y(0)|U}(y|u;\hat{\theta}_0) dy' dy \\ & \cdot \left(D_i f_{U|D=1,Z,n}\left(u|Z_i;\hat{\theta}_{1Z}\right) + (1-D_i) f_{U|D=0,Z,n}\left(u|Z_i;\hat{\theta}_{0Z}\right)\right) du.\end{aligned}$$

In constructing induced estimators, the conditional densities $f_{U|D=1,Z}$ and $f_{U|D=0,Z}$ are used to obtain the marginal density of U_i, taking advantage of the following equivalence:

$$\mathbf{E}\left[g(U_i)\right] = \mathbf{E}\left[\mathbf{E}\left[g(U_i)|D_i, Z_i\right]\right].$$

B Proofs

B.1 Proof for Theorem 1

This subsection completes the proof for Theorem 1 under Assumptions 1-2, 4-5, by extending the spectral decomposition result of Hu and Schennach (2008).[9] For the proof of the spectral decomposition results, refer to Hu and Schennach (2008). By applying assumptions of Hu and Schennach (2008), except their Assumption 5, 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 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. To find an ordering on the infinite number of triads, let $\tilde{U}_i = h(U_i) := M f_{Y(1),Y(0)|U}(\cdot,\cdot|U_i)$ and $\tilde{\mathcal{U}} = h(\mathcal{U})$. Now, we have labeled each triad with $\tilde{u} = h(u)$. The remainder of the proof constructs conditional densities and a marginal density in terms of the new latent variable $\tilde{U}_i$ as ingredients in identifying the joint density of $Y_i(1)$ and $Y_i(0)$ and shows that the strict monotonicity of h allows us to identify the joint distribution of $Y_i(1)$ and $Y_i(0)$ using $\tilde{U}_i$ instead of U_i.

Firstly, let us establish the injectivity of the integral operator based on the conditional density of X_i given $\tilde{U}_i$. Find that

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

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 also identified.

Secondly, it remains to show that the arbitrary choice of $\tilde{U}_i$ does not matter. Under the

[9] The identification under Assumptions 1-3 is straightforward from the discussion in the main text.

conditional independence of $Y_i(1)$ and $Y_i(0)$ given U_i, the joint distribution of $Y_i(1)$ and $Y_i(0)$ is a function of three distributions: 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. For each $(y_1, y_0) \in \mathbb{R}^2$,

$$\begin{aligned}
f_{Y(1),Y(0)}(y_1, y_0) &= \int_{\mathcal{U}} f_{Y(1)|U}(y_1|u) f_{Y(0)|U}(y_0|u) f_U(u) du \\
&= \int_{\mathcal{U}} f_{Y(1)|\tilde{U}}(y_1|h(u)) f_{Y(0)|\tilde{U}}(y_0|h(u)) f_U(u) du \\
&= \int_{\tilde{\mathcal{U}}} f_{Y(1)|\tilde{U}}(y_1|\tilde{u}) f_{Y(0)|\tilde{U}}(y_0|\tilde{u}) \frac{f_U\left(h^{-1}(\tilde{u})\right)}{h'\left(h^{-1}(\tilde{u})\right)} d\tilde{u}, \quad \text{by letting } u = h^{-1}(\tilde{u}) \\
&= \int_{\tilde{\mathcal{U}}} f_{Y(1)|\tilde{U}}(y_1|\tilde{u}) f_{Y(0)|\tilde{U}}(y_0|\tilde{u}) f_{\tilde{U}}(\tilde{u}) d\tilde{u}, \quad \text{since } F_U\left(h^{-1}(\tilde{u})\right) = F_{\tilde{U}}(\tilde{u}).
\end{aligned}$$

The last two equalities are from the inverse function theorem: $\left(h^{-1}(\tilde{u})\right)' = 1/h'\left(h^{-1}(\tilde{u})\right)$. The joint distribution of $Y_i(1)$ and $Y_i(0)$ is identified. The expansion to include (D_i, X_i, Z_i) follows the same argument.

B.2 Proof for Lemma 1

Let us consider three different parts of ϕ: ϕ_A, ϕ_B, ϕ_C. Firstly, ϕ_A is the part of ϕ that corresponds to the quadratic constraints (13). Fix some (y, d, x, k) and let

$$\begin{aligned}
&\phi_A(W_i, W_{i'}; \tilde{\lambda}, p) \\
&= \sum_j \frac{\tilde{\lambda}_{jk,d}}{p_{D,Z}(d,j)} \cdot \frac{\mathbf{1}\{Y_i = y, D_i = d, X_i = x, Z_i = z^j\} + \mathbf{1}\{Y_{i'} = y, D_{i'} = d, X_{i'} = x, Z_{i'} = z^j\}}{2} \\
&\quad - \sum_{j,j'} \frac{\tilde{\lambda}_{jk,d}\tilde{\lambda}_{j'k,d}}{p_{D,Z}(d,j) \cdot p_{D,Z}(d,j')} \cdot \frac{1}{2} \Big(\mathbf{1}\{Y_i = y, D_i = d, Z_i = z^j, X_{i'} = x, D_{i'} = d, Z_{i'} = z^{j'}\} \\
&\qquad\qquad + \mathbf{1}\{X_i = x, D_i = d, Z_i = z^{j'}, Y_{i'} = y, D_{i'} = d, Z_{i'} = z^j\} \Big).
\end{aligned}$$

Then,

$$\begin{aligned}
&\mathbf{E}\left[\frac{\partial}{\partial \tilde{\lambda}_{jk,d}} \phi_A\big(W_i, W_{i'}; \tilde{\lambda}, p\big)\right] \\
&= \Pr\{Y_i = y, X_i = x | D_i = d, Z_i = z^j\} - \Pr\{Y_i = y | D_i = d, Z = z^j\} \cdot \Pr\{X_i = x | U_i = u^k\} \\
&\qquad - \Pr\{X_i = x | D_i = d, Z = z^j\} \cdot \Pr\{Y_i(d) = y | U_i = u^k\}
\end{aligned}$$

and $\mathbf{E}\Big[\frac{\partial}{\partial \tilde{\lambda}_{jk',d'}}\phi_A\big(W_i, W_{i'}; \tilde{\lambda}, p\big)\Big]$ is zero when $k' \neq k$ or $d' \neq d$. $\mathbf{E}\Big[\frac{\partial}{\partial p_U(k)}\phi_A\big(W_i, W_{i'}; \tilde{\lambda}, p\big)\Big] = 0$ for every k. Lastly,

$$
\begin{aligned}
&\mathbf{E}\left[\frac{\partial}{\partial p_{D,Z}(d,j)}\phi_A\big(W_i, W_{i'}; \tilde{\lambda}, p\big)\right] \\
&\quad = -\frac{\tilde{\lambda}_{jk,d}}{p_{D,Z}(d,j)} \cdot \Pr\{Y_i = y, X_i = x | D_i = d, Z_i = z^j\} \\
&\qquad + \frac{\tilde{\lambda}_{jk,d}}{p_{D,Z}(d,j)} \cdot \Pr\{Y_i = y | D_i = d, Z_i = z^j\} \cdot \Pr\{X_i = x | U_i = u^k\} \\
&\qquad + \frac{\tilde{\lambda}_{jk,d}}{p_{D,Z}(d,j)} \cdot \Pr\{X_i = x | D_i = d, Z_i = z^j\} \cdot \Pr\{Y_i(d) = y | U_i = u^k\}
\end{aligned}
$$

and $\mathbf{E}\Big[\frac{\partial}{\partial p_{D,Z}(d',j)}\phi_A\big(W_i, W_{i'}; \tilde{\lambda}, p\big)\Big]$ is zero when $d' \neq d$.

Secondly, ϕ_B is the part of ϕ that corresponds to the linear constraints (14). Fix some (d,x) and let

$$
\begin{aligned}
&\phi_B(W_i, W_{i'}; \tilde{\lambda}, p) \\
&= \frac{\mathbf{1}\{X_i = x\} + \mathbf{1}\{X_{i'} = x\}}{2} \\
&\quad - \sum_k p_U(k) \sum_j \frac{\tilde{\lambda}_{jk,d}}{p_{D,Z}(d,j)} \cdot \frac{\mathbf{1}\{D_i = d, X_i = x, Z_i = z^j\} + \mathbf{1}\{D_{i'} = d, X_{i'} = x, Z_{i'} = z^j\}}{2}.
\end{aligned}
$$

Then,

$$
\mathbf{E}\left[\frac{\partial}{\partial \tilde{\lambda}_{jk,d}}\phi_B\big(W_i, W_{i'}; \tilde{\lambda}, p\big)\right] = -p_U(k) \cdot \Pr\{X_i = x | D_i = d, Z_i = z^j\}
$$

and $\mathbf{E}\Big[\frac{\partial}{\partial \tilde{\lambda}_{jk,d'}}\phi_B\big(W_i, W_{i'}; \tilde{\lambda}, p\big)\Big]$ is zero when $d' \neq d$. Also,

$$
\begin{aligned}
\mathbf{E}\left[\frac{\partial}{\partial p_U(k)}\phi_B\big(W_i, W_{i'}; \tilde{\lambda}, p\big)\right] &= -\Pr\{X_i = x | U_i = u^k\} \\
\mathbf{E}\left[\frac{\partial}{\partial p_{D,Z}(d,j)}\phi_B\big(W_i, W_{i'}; \tilde{\lambda}, p\big)\right] &= \sum_{k=1}^{K} \frac{p_U(k)\tilde{\lambda}_{jk,d}}{p_{D,Z}(d,j)} \cdot \Pr\{X_i = x | D_i = d, Z_i = z^j\}
\end{aligned}
$$

and $\mathbf{E}\Big[\frac{\partial}{\partial p_{D,Z}(d',j)}\phi_B\big(W_i, W_{i'}; \tilde{\lambda}, p\big)\Big]$ is zero when $d' \neq d$.

Thirdly, ϕ_C is the moment condition for $p_{D,Z}$. Fix some (d, j) and let

$$\phi_C(W_i, W_{i'}; \tilde{\lambda}, p) = \frac{\mathbf{1}\{D_i = d, Z_i = z^j\} + \mathbf{1}\{D_{i'} = d, Z_{i'} = z^j\}}{2} - p_{D,Z}(d, j).$$

Then, $\mathbf{E}\Big[\frac{\partial}{\partial \tilde{\lambda}_{jk,d'}} \phi_C\big(W_i, W_{i'}; \tilde{\lambda}, p\big)\Big]$ and $\mathbf{E}\Big[\frac{\partial}{\partial p_U(k)} \phi_C\big(W_i, W_{i'}; \tilde{\lambda}, p\big)\Big]$ are zero for every (d', j, k). Also,

$$\mathbf{E}\Big[\frac{\partial}{\partial p_{D,Z}(d, j)} \phi_C\big(W_i, W_{i'}; \tilde{\lambda}, p\big)\Big] = -1$$

and $\mathbf{E}\Big[\frac{\partial}{\partial p_{D,Z}(d', j')} \phi_C\big(W_i, W_{i'}; \tilde{\lambda}, p\big)\Big]$ is zero when $d' \neq d$ or $j' \neq j$.

The order of ϕ_A, ϕ_B and ϕ_C across different values of (y, x, d, j, k) in ϕ is as follows. Firstly, stack ϕ_A across every value of (y, x) for $(d = 0, k = 1)$ and then for $(d = 1, k = 1)$. Then, repeat this for $k = 2, \cdots, K$. These will be the first $2MK$ components of ϕ. Secondly, stack ϕ_B across every value of x for $d = 0$ and then for $d = 1$. These will be the second $2M_X$ components of ϕ. Then, stack ϕ_C across every value of j for $d = 0$ and then for $d = 1$. These will be the last $2K$ components of ϕ.

Also, we need to decide on the order of $\tilde{\lambda}_{jk,d}$ in vectorized $\tilde{\lambda}$ and similarly for p. In a similar manner to ϕ, collect $\tilde{\lambda}_{jk,d}$ across j for $(d = 0, k = 1)$ and then for $(d = 1, k = 1)$. Then, repeat this for $k = 2, \cdots, K$. These will be the $2K^2$-dimensional vector $\tilde{\lambda}$. For p, first collect $p_U(k)$ across k, collect $p_{D,Z}(0, j)$ across j, and then collect $p_{D,Z}(1, j)$ across j.

With this order of stacking/vectorization, the Jacobian matrix becomes

$$\begin{aligned}
&\begin{pmatrix} \mathbf{E}\left[\frac{\partial}{\partial \tilde{\lambda}} \phi\left(W_i, W_{i'}; \tilde{\lambda}, p\right)\right] \\ \mathbf{E}\left[\frac{\partial}{\partial p} \phi\left(W_i, W_{i'}; \tilde{\lambda}, p\right)\right] \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{E}\left[\frac{\partial}{\partial \tilde{\lambda}} \phi_A\left(W_i, W_{i'}; \tilde{\lambda}, p\right)\right] & \mathbf{E}\left[\frac{\partial}{\partial \tilde{\lambda}} \phi_B\left(W_i, W_{i'}; \tilde{\lambda}, p\right)\right] & \mathbf{O}_{2K^2 \times 2K} \\ \mathbf{O}_{K \times 2MK} & \mathbf{E}\left[\frac{\partial}{\partial p_U} \phi_B\left(W_i, W_{i'}; \tilde{\lambda}, p\right)\right] & \mathbf{O}_{K \times 2K} \\ \mathbf{E}\left[\frac{\partial}{\partial p_{D,Z}} \phi_A\left(W_i, W_{i'}; \tilde{\lambda}, p\right)\right] & \mathbf{E}\left[\frac{\partial}{\partial p_{D,Z}} \phi_B\left(W_i, W_{i'}; \tilde{\lambda}, p\right)\right] & -\mathbf{I}_{2K \times 2K} \end{pmatrix}.
\end{aligned}$$

It suffices to show that the submatrix

$$\begin{pmatrix} \mathbf{E}\left[\frac{\partial}{\partial \tilde{\lambda}} \phi_A\left(W_i, W_{i'}; \tilde{\lambda}, p\right)\right] & \mathbf{E}\left[\frac{\partial}{\partial \tilde{\lambda}} \phi_B\left(W_i, W_{i'}; \tilde{\lambda}, p\right)\right] \\ \mathbf{O}_{K \times 2MK} & \mathbf{E}\left[\frac{\partial}{\partial p_U} \phi_B\left(W_i, W_{i'}; \tilde{\lambda}, p\right)\right] \end{pmatrix}. \tag{17}$$

is full rank. Assume to the contrary that the rows of the submatrix from (17) are linearly dependent:

with linear coefficients $\alpha = \left(\alpha_{A,1}, \cdots, \alpha_{A,2K^2}, \alpha_{B,1}, \cdots, \alpha_{B,2K}\right)^\intercal$,

$$\alpha^\intercal \begin{pmatrix} \mathbf{E}\left[\frac{\partial}{\partial \tilde{\lambda}}\phi_A\left(W_i, W_{i'}; \tilde{\lambda}, p\right)\right] & \mathbf{E}\left[\frac{\partial}{\partial \tilde{\lambda}}\phi_B\left(W_i, W_{i'}; \tilde{\lambda}, p\right)\right] \\ \mathbf{O}_{K\times 2MK} & \mathbf{E}\left[\frac{\partial}{\partial p_U}\phi_B\left(W_i, W_{i'}; \tilde{\lambda}, p\right)\right] \end{pmatrix} = \mathbf{0}.$$

Note that $\mathbf{E}\left[\frac{\partial}{\partial \tilde{\lambda}}\phi_A\left(W_i, W_{i'}; \tilde{\lambda}, p\right)\right]$ is a diagonal block matrix, consisting of $2K$ block matrices, each of which is a $K \times M$ matrix. For example, the first block matrix is

$$\begin{aligned} &{\Lambda_0}^\intercal{\Gamma_0}^\intercal - ({\Lambda_0}^\intercal{\Gamma_X}^\intercal) \otimes \left(\Pr\left\{Y_i(0) = y^1 | U_i = u^1\right\} \quad \cdots \quad \Pr\left\{Y_i(0) = y^{M_Y} | U_i = u^1\right\}\right) \\ &\quad - \left(\Pr\left\{X_i = x^1 | U_i = u^1\right\} \quad \cdots \quad \Pr\left\{X_i = x^{M_X} | U_i = u^1\right\}\right) \otimes {\Lambda_0}^\intercal{\Gamma_{Y(0)}}^\intercal \end{aligned}$$

where $\otimes$ is the Kronecker product. From Assumption 3.b-c, the rows of the block matrices are linearly independent. Thus, the first $2K^2$ components of α are zeroes. Then, it must satisfy that

$${\alpha_B}^\intercal\mathbf{E}\left[\frac{\partial}{\partial p_U}\phi_B\left(W_i, W_{i'}; \tilde{\lambda}, p\right)\right] = {\alpha_B}^\intercal{\Gamma_X}^\intercal = \mathbf{0}.$$

From Assumption 3.b, α_B must be a zero vector. The Jacobian matrix has full rank.

B.3 Proof for Theorem 2

All of the following proof is for $K \geq 2$.

Step 1. $\left\|\Gamma_0\Lambda_0 - \widehat{\Gamma}_0\widehat{\Lambda}_0\right\|_F^2 = O_p\left(\frac{1}{\sqrt{n}}\right)$ and $\left\|\Gamma_1\Lambda_1 - \widehat{\Gamma}_1\widehat{\Lambda}_1\right\|_F^2 = O_p\left(\frac{1}{\sqrt{n}}\right)$.

From iid-ness of observations, we have

$$\|\mathbb{H}_0 - \mathbf{H}_0\|_F = O_p\left(\frac{1}{\sqrt{n}}\right) \quad \text{and} \quad \|\mathbb{H}_1 - \mathbf{H}_1\|_F = O_p\left(\frac{1}{\sqrt{n}}\right).$$

From the definition of $\widehat{\Lambda}_0$ and $\widehat{\Lambda}_1$, we have

$$\begin{aligned} \left\|\mathbb{H}_0 - \widehat{\Gamma}_0\widehat{\Lambda}_0\right\|_F^2 + \left\|\mathbb{H}_1 - \widehat{\Gamma}_1\widehat{\Lambda}_1\right\|_F^2 &\leq \|\mathbb{H}_0 - \Gamma_0\Lambda_0\|_F^2 + \|\mathbb{H}_1 - \Gamma_1\Lambda_1\|_F^2 \\ &= \|\mathbb{H}_0 - \mathbf{H}_0\|_F^2 + \|\mathbb{H}_1 - \mathbf{H}_1\|_F^2 = O_p\left(\frac{1}{n}\right). \end{aligned}$$

Then,

$$\left\|\Gamma_0\Lambda_0 - \widehat{\Gamma}_0\widehat{\Lambda}_0\right\|_F^2 = \left\|\mathbf{H}_0 - \widehat{\Gamma}_0\widehat{\Lambda}_0\right\|_F^2 \leq \left(\|\mathbf{H}_0 - \mathbb{H}_0\|_F + \left\|\mathbb{H}_0 - \widehat{\Gamma}_0\widehat{\Lambda}_1\right\|_F\right)^2 = O_p\left(\frac{1}{n}\right)$$

and likewise for $\left\|\Gamma_1\Lambda_1 - \widehat{\Gamma}_1\widehat{\Lambda}_1\right\|_F = \left\|\mathbf{H}_1 - \widehat{\Gamma}_1\widehat{\Lambda}_1\right\|_F$. From the submultiplicavity of $\|\cdot\|_F$, we also get $\left\|P\Gamma_1\Lambda_1 - P\widehat{\Gamma}_1\widehat{\Lambda}_1\right\|_F = \left\|P\Gamma_0\Lambda_1 - P\widehat{\Gamma}_0\widehat{\Lambda}_1\right\|_F = O_p\left(\frac{1}{\sqrt{n}}\right)$.

To avoid repetition, we will only prove the consistency of $\widehat{\Lambda}_0$; the same argument applies to $\widehat{\Lambda}_1$.

Step 2. $\left\|\widehat{\Gamma}_0 - \Gamma_0 A\right\|_F = O_p\left(\frac{1}{\sqrt{n}}\right)$ with some $K \times K$ matrix A.

Firstly, I show that $\widehat{\Lambda}_0^{-1}$ exists with probability going to one. Find that

$$\left\|{\Gamma_0}^\intercal\widehat{\Gamma}_0\widehat{\Lambda}_0 - {\Gamma_0}^\intercal\Gamma_0\Lambda_0\right\|_F \leq \|\Gamma_0\|_F \cdot \left\|\Gamma_0\Lambda_0 - \widehat{\Gamma}_0\widehat{\Lambda}_0\right\|_F = O_p\left(\frac{1}{\sqrt{n}}\right).$$

The determinant of ${\Gamma_0}^\intercal\widehat{\Gamma}_0\widehat{\Lambda}_0$ converges in probability to the determinant of ${\Gamma_0}^\intercal\Gamma_0\Lambda_0$, which is nonzero. Thus, with probability converging to one, both ${\Gamma_0}^\intercal\widehat{\Gamma}_0$ and $\widehat{\Lambda}_0$ have full rank and $\left({\Gamma_0}^\intercal\widehat{\Gamma}_0\right)^{-1}$ and $\widehat{\Lambda}_0^{-1}$ exist.

Let

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

with $\mathbf{I}_K$ being the $K \times K$ identity matrix. When ${\Gamma_0}^\intercal\widehat{\Gamma}_0\widehat{\Lambda}_0$ is invertible,

$$\begin{aligned} \left\|\widehat{\Gamma}_0 - \Gamma_0 A\right\|_F &= \left\|\left(\widehat{\Gamma}_0\widehat{\Lambda}_0 - \Gamma_0\Lambda_0\right)\widehat{\Lambda}_0^{-1}\right\|_F \\ &\leq \left\|\widehat{\Gamma}_0\widehat{\Lambda}_0 - \Gamma_0\Lambda_0\right\|_F \left\|\left({\Gamma_0}^\intercal\widehat{\Gamma}_0\widehat{\Lambda}_0\right)^{-1}\right\|_F \left\|{\Gamma_0}^\intercal\widehat{\Gamma}_0\right\|_F. \end{aligned}$$

There is some $\delta > 0$ such that $\left\|{\Gamma_0}^\intercal\widehat{\Gamma}_0\widehat{\Lambda}_0 - {\Gamma_0}^\intercal\Gamma_0\Lambda_0\right\|_F \leq \delta$ implies the invertibility of ${\Gamma_0}^\intercal\widehat{\Gamma}_0\widehat{\Lambda}_0$ and

$$C = \left\{\left\|\left({\Gamma_0}^\intercal\widehat{\Gamma}_0\widehat{\Lambda}_0\right)^{-1}\right\|_F : \left\|{\Gamma_0}^\intercal\widehat{\Gamma}_0\widehat{\Lambda}_0 - {\Gamma_0}^\intercal\Gamma_0\Lambda_0\right\|_F \leq \delta\right\} < \infty$$

since $\left\|\left({\Gamma_0}^\intercal\widehat{\Gamma}_0\widehat{\Lambda}_0\right)^{-1}\right\|_F$ is a continuous function of ${\Gamma_0}^\intercal\widehat{\Gamma}_0\widehat{\Lambda}_0$ and

$$\left\{{\Gamma_0}^\intercal\widehat{\Gamma}_0\widehat{\Lambda}_0 : \left\|{\Gamma_0}^\intercal\widehat{\Gamma}_0\widehat{\Lambda}_0 - {\Gamma_0}^\intercal\Gamma_0\Lambda_0\right\|_F \leq \delta\right\}$$

is closed and bounded. Then,

$$\Pr\left\{\left(\left\|{\Gamma_0}^\intercal\widehat{\Gamma}_0\widehat{\Lambda}_0\right)^{-1}\right\|_F \geq C, {\Gamma_0}^\intercal\widehat{\Gamma}_0\widehat{\Lambda}_0 \text{ is invertible}\right\} = o(1)$$

Also, $\left\|{\Gamma_0}^\intercal\widehat{\Gamma}_0\right\|_F$ is bounded by K^2. Thus,

$$
\begin{aligned}
&\Pr\left\{\sqrt{n}\left\|\widehat{\Gamma}_0-\Gamma_0 A\right\|_F\geq\varepsilon\right\}\\
&\leq\Pr\left\{\sqrt{n}\left\|\widehat{\Gamma}_0\widehat{\Lambda}_0-\Gamma_0\Lambda_0\right\|_F\left\|\left({\Gamma_0}^\intercal\widehat{\Gamma}_0\widehat{\Lambda}_0\right)^{-1}\right\|_F\left\|{\Gamma_0}^\intercal\widehat{\Gamma}_0\right\|_F\geq\varepsilon,{\Gamma_0}^\intercal\widehat{\Gamma}_0\widehat{\Lambda}_0\text{ is invertible}\right\}+o(1)\\
&\leq\Pr\left\{\sqrt{n}\left\|\widehat{\Gamma}_0\widehat{\Lambda}_0-\Gamma_0\Lambda_0\right\|_F\geq\frac{\varepsilon}{CK^2}\right\}+o(1)
\end{aligned}
$$

Therefore, we have

$$
\left\|\widehat{\Gamma}_0-\Gamma_0 A\right\|_F=O_p\left(\frac{1}{\sqrt{n}}\right).
$$

A is a $K\times K$ matrix that reorders the columns of Γ_0 so that it resembles $\widehat{\Gamma}_0$. Let a_{jk} denote the j-th row and k-th column element of A and $a_{\cdot k}$ denote the k-th column of A. In this sense, $a_{\cdot k}$ is a set of weights on the columns of $\widehat{\Gamma}_0$ so that we get the k-th column in Γ_0.

Step 3. Each column of A converges to an elementary vector at the rate of $n^{-\frac{1}{2}}$.

Firstly, the columns of A sum to one. To see this, compute column-wise sums of

$$
\widehat{\Gamma}_0=\Gamma_0 A+\left(\widehat{\Gamma}_0\widehat{\Lambda}_0-\Gamma_0\Lambda_0\right)\widehat{\Lambda}_0^{-1}
$$

when ${\Gamma_0}^\intercal\widehat{\Gamma}_0\widehat{\Lambda}_0$ is invertible:

$$
\begin{aligned}
{\iota_M}^\intercal\widehat{\Gamma}_0&={\iota_M}^\intercal\Gamma_0 A+{\iota_M}^\intercal\left(\widehat{\Gamma}_0\widehat{\Lambda}_0-\Gamma_0\Lambda_0\right)\widehat{\Lambda}_0^{-1}\\
{\iota_K}^\intercal&={\iota_K}^\intercal A+\left({\iota_K}^\intercal\widehat{\Lambda}_0-{\iota_K}^\intercal\Lambda_0\right)\widehat{\Lambda}_0^{-1}\\
{\iota_K}^\intercal&={\iota_K}^\intercal A+\left({\iota_K}^\intercal-{\iota_K}^\intercal\right)\widehat{\Lambda}_0^{-1}\\
{\iota_K}^\intercal&={\iota_K}^\intercal A.
\end{aligned}
$$

Secondly, with probability going to one, the columns of A are bounded with $\|\cdot\|_\infty$. To see this, let $\Gamma_{0,k}$ be the k-the column of Γ_0 and let $\Gamma_{0,-k}$ be the rest of the $K-1$ columns formed into a $M\times(K-1)$ matrix. Let

$$
\delta^*:=\min_k\left\|\Gamma_{0,k}-\Gamma_{0,-k}\left({\Gamma_{0,-k}}^\intercal\Gamma_{0,-k}\right)^{-1}{\Gamma_{0,-k}}^\intercal\Gamma_{0,k}\right\|.
$$

$\delta^* > 0$ from Assumption 3.b. Then, for any linear combination of $\Gamma_{0,-k}$,

$$\|\Gamma_{0,k} - \Gamma_{0,-k}\alpha\|_\infty \geq \frac{\delta^*}{2\sqrt{M}}.$$

Since each column of A sum to one, a k-th column element of $\Gamma_0 A$ can be written as follows:

$$\begin{aligned}
&\sum_{j=1}^{K} \Pr\{Y_i(0) = y, X_i = x | U_i = u^j\} a_{jk} \\
&= \Pr\{Y_i(0) = y, X_i = x | U_i = u^1\} \\
&\quad + (1 - a_{1k}) \left(\sum_{j=2}^{K} \Pr\{Y_i(0) = y, X_i = x | U_i = u^j\} \cdot \frac{a_{jk}}{\sum_{j=2}^{K} a_{jk}} - \Pr\{Y_i(0) = y, X_i = x | U_i = u^1\} \right)
\end{aligned}$$

For any given $\{a_{jk}\}_{j=2}^{K}$, we know from the construction of δ^* that there must be a row in $\Gamma_0 A$ such that

$$\left| \Pr\{Y_i(0) = y, X_i = x | U_i = u^1\} - \sum_{j=2}^{K} \Pr\{Y_i(0) = y, X_i = x | U_i = u^j\} \cdot \frac{a_{jk}}{\sum_{j=2}^{K} a_{jk}} \right| \geq \frac{\delta^*}{2\sqrt{M}}.$$

Thus, $\sum_{j=1}^{K} \Pr\{Y_i(0) = y, X_i = x | U_i = u^j\} a_{jk}$ lies outside of

$$\Pr\{Y_i(0) = y, X_i = x | U_i = u^1\} + \left[-\frac{|1 - a_{1k}|\delta^*}{2\sqrt{M}}, \frac{|1 - a_{1k}|\delta^*}{2\sqrt{M}} \right]$$

and

$$\Pr\left\{ |1 - a_{1k}| \geq \frac{4\sqrt{M}}{\delta^*} \right\} \leq \Pr\left\{ \left\| \widehat{\Gamma}_0 - \Gamma_0 A \right\|_F \geq 1 \right\} = o(1).$$

The inequality holds since $\widehat{\Gamma}_0$ is a well-defined probability matrix and therefore its elements all lie between 0 and 1. We can repeat this for every a_{jk} and we have $\Pr\left\{ \|a_{\cdot k}\|_\infty \geq \frac{4\sqrt{M}}{\delta^*} + 1 \right\} = o(1)$ for every k.

Using these two observations, now I show that each column of A converges to an elementary vector at the rate of $\frac{1}{\sqrt{n}}$: with e_k being the k-th elementary vector whose k-th element is one and the rest are zeros and some $\varepsilon > 0$,

$$\Pr\left\{ \sqrt{n} \cdot \min_k \|a_{\cdot 1} - e_k\| \geq \varepsilon \right\} = o(1).$$

To put a bound on the probability, I first show that $\sqrt{n} \cdot \min_k \|a_{\cdot 1} - e_k\| \geq \varepsilon$ implies that there is at least one j such that $|a_{j1}| \geq \frac{1}{K}$ and another $j' \neq j$ such that $|a_{j'1}| \geq \frac{\varepsilon}{2\sqrt{n}K}$. The existence of such j is trivial from $\sum_{k=1}^{K} a_{k1} = 1$. Assume to the contrary that there exists only one j such that $|a_{j1}| \geq \frac{\varepsilon}{2\sqrt{n}K}$. Then, for the rest of $K-1$ elements, it must be that $|a_{k1}| \leq \frac{\varepsilon}{2\sqrt{n}K}$, which leads to $a_{j1} \in [1 - \frac{\varepsilon}{2\sqrt{n}}, 1 + \frac{\varepsilon}{2\sqrt{n}}]$. Then,

$$\|a_{\cdot 1} - e_j\| \leq \left(\frac{\varepsilon^2}{4n} \cdot \frac{K-1}{K^2} + \frac{\varepsilon^2}{4n} \right)^{\frac{1}{2}} \leq \frac{\varepsilon}{\sqrt{2n}} < \min_k \|a_{\cdot 1} - e_k\|,$$

which leads to a contradiction. Thus, we have

$$\Pr\left\{ \sqrt{n} \cdot \min_k \|a_{\cdot 1} - e_k\| \geq \varepsilon \right\} \leq \Pr\left\{ \exists\ j, j' \text{ such that } j \neq j', |a_{j1}| \geq \frac{1}{K}, |a_{j'1}| \geq \frac{\varepsilon}{2\sqrt{n}K} \right\}.$$

Two elements of $a_{\cdot 1}$ being away from zero creates a contradiction to $\left\|\widehat{\Gamma}_0 - \Gamma_0 A\right\|_F = O_p\left(\frac{1}{\sqrt{n}}\right)$ since the convergence says that each column of $\Gamma_0 A$ can be well-approximated by a column in $\widehat{\Gamma}_0$, which satisfies the quadratic constraints (11). To see this, let $\tilde{\Gamma}_{0,k}$ be a $M_X \times M_Y$ matrix whose m-th row and m'-th column element is

$$\Pr\left\{ Y_i(0) = y^{m'}, X_i = x^m | U_i = u^k \right\}.$$

$\tilde{\Gamma}_{0,k}$ takes the k-th column of Γ_0 and makes it into a $M_X \times M_Y$ matrix. Note that $\tilde{\Gamma}_{0,k} = p_k q_{0k}{}^\intercal$, with

$$\begin{aligned} p_k &= \left(\Pr\left\{ X_i = x^1 | U_i = u^k \right\} \quad \cdots \quad \Pr\left\{ X_i = x^{M_X} | U_i = u^k \right\} \right)^\intercal, \\ q_{dk} &= \left(\Pr\left\{ Y_i(d) = y^1 | U_i = u^k \right\} \quad \cdots \quad \Pr\left\{ Y_i(d) = y^{M_Y} | U_i = u^k \right\} \right)^\intercal \qquad \forall k = 1, \cdots, K. \end{aligned}$$

Then, $\min_{p,q} \left\| \sum_{k=1}^{K} \tilde{\Gamma}_{0,k} a_{k1} - pq^\intercal \right\|_F = O_p\left(\frac{1}{\sqrt{n}}\right)$ since

$$\min_{p \in \mathbb{R}^{M_X}, q \in \mathbb{R}^{M_Y}} \left\| \sum_{k=1}^{K} \tilde{\Gamma}_{0,k} a_{k1} - pq^\intercal \right\|_F \leq \left\| \sum_{k=1}^{K} \tilde{\Gamma}_{0,k} a_{k1} - \widehat{\tilde{\Gamma}}_{0,1} \right\|_F \leq \left\| \widehat{\Gamma}_0 - \Gamma_0 A \right\|_F$$

with $\widehat{\tilde{\Gamma}}_{0,k}$ constructed from $\widehat{\Gamma}_0$ in the same manner as $\tilde{\Gamma}_{0,k}$. The first inequality holds from the construction of the estimator $\widehat{\Gamma}_0$; the estimated mixture component distribution satisfies the exclusion restriction of $Y_i(0)$ and X_i given U_i and thus $\widehat{\tilde{\Gamma}}_{0,1}$ is a rank one matrix. The second inequality holds since $\sum_{k=1}^{K} \tilde{\Gamma}_{0,k} a_{k1}$ corresponds to the first column of $\Gamma_0 A$ and $\widehat{\widehat{\Gamma}}_{0,1}$ corresponds to the first column of $\widehat{\Gamma}_0$. However, since two elements of $a_{\cdot 1}$ are away from zero, the matrix $\sum_{k=1}^{K} \tilde{\Gamma}_{0,k} a_{k1}$ cannot be well-approximated by a rank one matrix as implied by $\left\|\widehat{\Gamma}_0 - \Gamma_0 A\right\|_F = O_p\left(\frac{1}{\sqrt{n}}\right)$, giving us a contradiction.

The rest of the step completes the argument. Assume that there exist some j, j' such that $j \neq j', |a_{j1}| \geq \frac{1}{K}, |a_{j'1}| \geq \frac{\varepsilon}{2\sqrt{n}K}$. Let $p_k(x) = \Pr\{X_i = x | U_i = u^k\}$, $q_{dk}(y) = \Pr\{Y_i(d) = y | U_i = u^k\}$ for $k = 1, \cdots, K$ and let

$$w(y) = \begin{pmatrix} a_{11} q_{01}(y) & \cdots & a_{K1} q_{0K}(y) \end{pmatrix}^{\intercal}.$$

Then,

$$\sum_{k=1}^{K} \tilde{\Gamma}_{0,k} a_{k1} = \sum_{k=1}^{K} a_{k1} p_k q_{0k}^{\intercal} = \Gamma_X \begin{pmatrix} w(y^1) & \cdots & w(y^{M_Y}) \end{pmatrix}.$$

From Assumption 3.c,

$$c^* := \min_{k \neq k'} \left\{ \max_y \left(q_{0k}(y) - q_{0k'}(y) \right) \right\} > 0.$$

WLOG let y^1 and y^2 satisfy that

$$q_{0j}(y^1) - q_{0j'}(y^1) \geq c^* \quad \text{and} \quad q_{0j'}(y^2) - q_{0j}(y^2) \geq c^*.$$

Then, since $\left(q_{0j}(y^1) q_{0j'}(y^2) - q_{0j'}(y^1) q_{0j}(y^2)\right) \geq c^{*2}$,

$$\left|w_j(y^1) w_{j'}(y^2) - w_{j'}(y^1) w_j(y^2)\right| = \left|a_{j1} a_{j'1}\right| \left(q_{0j}(y^1) q_{0j'}(y^2) - q_{0j'}(y^1) q_{0j}(y^2)\right) \geq \frac{\varepsilon c^{*2}}{2\sqrt{n} K^2}.$$

With the columns corresponding to $\left(y^1, y^2\right)$, the submatrix of $\sum_{k=1}^{K} \tilde{\Gamma}_{0,k} a_{k1}$ is

$$\tilde{A} = \Gamma_X \begin{pmatrix} w(y^1) & w(y^2) \end{pmatrix}.$$

Then,

$$\min_{p,q} \left\| \sum_{k=1}^{K} \tilde{\Gamma}_{0,k} a_{k1} - p q^{\intercal} \right\|_F \geq \min_{p \in \mathbb{R}^{M_X}, q \in \mathbb{R}^2} \left\| \tilde{A} - p q^{\intercal} \right\|_F = \text{ the smallest singular value of } \tilde{A}.$$

The equality is from the Echkart-Young theorem. The smallest singular value of Γ_X is bounded away from zero from Assumption 3.b. To show that the smallest singular value of $\begin{pmatrix} w(y^1) & w(y^2) \end{pmatrix}$ is bounded away from zero with a lower bound proportional to $\frac{1}{\sqrt{n}}$, I use the following result:

Theorem 1 Hong and Pan (1992) *Let $A \in \mathbb{R}^{\rho\times\rho}$. Then, singular values of A are bounded from below by*

$$\left(\frac{\rho-1}{\rho}\right)^{\frac{\rho-1}{2}} |det(A)| \max\left\{\frac{\min_r \|A_{r\cdot}\|_2}{\prod_{r=1}^{\rho}\|A_{r\cdot}\|_2}, \frac{\min_s \|A_{\cdot s}\|_2}{\prod_{s=1}^{\rho}\|A_{\cdot s}\|_2}\right\}$$

where $A_{r\cdot}$ is the r-th row of A and $A_{\cdot s}$ is the s-th column of A.

Find that

$$\begin{aligned}
\text{the smallest eigenvalue of } \begin{pmatrix} w(y^1) & w(y^2)\end{pmatrix} &= \min_{p\in\mathbb{R}^{M_X}, q\in\mathbb{R}^2} \left\| \begin{pmatrix} w(y^1) & w(y^2)\end{pmatrix} - pq^\intercal \right\|_F \\
&\geq \min_{p,q\in\mathbb{R}^2} \left\| \begin{pmatrix} w_j(y^1) & w_j(y^2) \\ w_{j'}(y^1) & w_{j'}(y^2)\end{pmatrix} - pq^\intercal \right\|_F \\
&= \text{the smallest eigenvalue of } \begin{pmatrix} w_j(y^1) & w_j(y^2) \\ w_{j'}(y^1) & w_{j'}(y^2)\end{pmatrix}.
\end{aligned}$$

We have shown that

$$det \begin{pmatrix} w_j(y^1) & w_j(y^2) \\ w_{j'}(y^1) & w_{j'}(y^2)\end{pmatrix} \geq \frac{\varepsilon c^{*2}}{2\sqrt{n}K^2}.$$

With probability going to one, $w(y^1)$ and $w(y^2)$ is bounded by $\frac{4\sqrt{M}}{\delta^*}+1$ and therefore

$$\left(w_j(y^1)^2 + w_j(y^1)^2\right)^{-\frac{1}{2}} \leq \left(\frac{4\sqrt{2M}}{\delta^*} + \sqrt{2}\right)^{-1} > 0.$$

Thus, with probability going to one,

$$\text{the smallest eigenvalue of } \begin{pmatrix} w(y^1) & w(y^2)\end{pmatrix} \geq \frac{1}{\sqrt{n}} \cdot \frac{\varepsilon c^{*2}}{2K^2} \cdot \left(\frac{4\sqrt{2M}}{\delta^*} + \sqrt{2}\right)^{-1}$$

Consequently, with some constant $C^* > 0$ which does not depend on ε,

$$\begin{aligned}
&\Pr\left\{\sqrt{n}\cdot \min_k \|a_{\cdot 1} - e_k\| \geq \varepsilon\right\} \\
&\quad\leq \Pr\left\{\exists\ j, j' \text{ such that } j\neq j', |a_{j1}| \geq \frac{1}{K}, |a_{j'1}| \geq \frac{\varepsilon}{2\sqrt{n}K}\right\} \\
&\quad\leq \Pr\left\{\left\|\widehat{\Gamma}_0 - \Gamma_0 A\right\|_F \geq \frac{C^*\varepsilon}{\sqrt{n}}\right\} + \Pr\left\{\exists\ y \text{ s.t. } \|w(y)\|_\infty \geq \frac{4\sqrt{M}}{\delta^*} + 1\right\} = o(1).
\end{aligned}$$

We repeat this for every column of A: $a_{\cdot 2}, \cdots, a_{\cdot K}$.

Step 4. No two columns of A converge to the same elementary vector.

It remains to show that A is indeed a permutation; each of the elementary vector $e_1, \cdots, e_K$ has to show up once and only once, across the columns of A. To see this, let

$$\delta^{**} = \min_{1 \le k \le K} \max_{1 \le j \le K} \Pr\{U_i = u^k | D_i = 0, Z_i = z^j\} > 0.$$

δ^{**} finds row-wise maximums of Λ_0 and then finds the minimum among the maximum values. $\delta^{**} > 0$ since there cannot be a zero row in Λ_0, due to Assumption 3.b. From the result of Step 3, we have

$$\sum_{k=1}^{K} \Pr\left\{\min_{k'} \|a_{\cdot k} - e_{k'}\| \ge \frac{\delta^{**}}{K}\right\} = o(1).$$

If $\min_{k'} \|a_{\cdot k} - e_{k'}\| \le \frac{\delta^{**}}{K}$ for every k, there is a bijection between the columns of A and $\{e_1, \cdots, e_K\}$. Firstly, see that $\|a_{\cdot 1} - e_k\| \le \frac{\delta^{**}}{K}$ means that

$$\|a_{\cdot 1} - e_{k'}\| \ge 1 - \frac{\delta^{**}}{K} > \frac{\delta^{**}}{K} \quad \forall k' \ne k$$

since $\delta^{*}* < 1$ and $K \ge 2$. Thus, $\pi(k) = \arg\min_{k'} \|a_{\cdot k} - e_{k'}\|$ is a well-defined function when $\min_{k'} \|a_{\cdot k} - e_{k'}\| \le \frac{\delta^{**}}{K}$ for every k. Secondly, assume to the contrary that there is some j such that $j \ne \pi(k)$ for every k. Then, the j-th row of A lies in $\left[-\frac{\delta^{**}}{K}, \frac{\delta^{**}}{K}\right]$. Since the columns of $\tilde{\Lambda}_0$ sum to one, the j-th row of $\Lambda_0 = A\widehat{\Lambda}_0$ lies in $\left[-\frac{\delta^{**}}{K}, \frac{\delta^{**}}{K}\right]$, leading to a contradiction. Thus, π is a bijection.

Thus, with some permutation on the rows of $\widehat{\Lambda}_0$,

$$\begin{aligned}
&\Pr\left\{\sqrt{n}\,\|A - \mathbf{I}_K\|_F \ge \varepsilon\right\} \\
&\quad \le \Pr\left\{\sqrt{n}\,\|A - \mathbf{I}_K\|_F \ge \varepsilon, \min_{k'} \|a_{\cdot k} - e_{k'}\| \le \frac{\delta^{**}}{K} \text{ for every } k\right\} + o(1) \\
&\quad \le \sum_{k=1}^{K} \Pr\left\{\sqrt{n} \cdot \min_{k'} \|a_{\cdot k} - e_{k'}\| \ge \frac{\varepsilon}{\sqrt{K}}\right\} + o(1) = o(1).
\end{aligned}$$

Step 5. Lastly, $\left\|\widehat{\Lambda}_0 - \Lambda_0\right\|_F = O_p\left(\frac{1}{\sqrt{n}}\right)$.

Find that

$$
\begin{aligned}
&\left\|\Lambda_0-\widehat{\Lambda}_0\right\|_F \\
&\le \left\|\Lambda_0-\left(\Gamma_0{}^\intercal\Gamma_0\right)^{-1}\Gamma_0{}^\intercal\widehat{\Gamma}_0\widehat{\Lambda}_0\right\|_F+\left\|\left(\Gamma_0{}^\intercal\Gamma_0\right)^{-1}\Gamma_0{}^\intercal\widehat{\Gamma}_0\widehat{\Lambda}_0-\widehat{\Lambda}_0\right\|_F \\
&\le \left\|\left(\Gamma_0{}^\intercal\Gamma_0\right)^{-1}\Gamma_0{}^\intercal\right\|_F\cdot\left\|\Gamma_0\Lambda_0-\widehat{\Gamma}_0\widehat{\Lambda}_0\right\|_F+\left\|\left(\Gamma_0{}^\intercal\Gamma_0\right)^{-1}\Gamma_0{}^\intercal\widehat{\Gamma}_0-\mathbf{I}_K\right\|_F\cdot\left\|\widehat{\Lambda}_0\right\|_F \\
&\le \left\|\left(\Gamma_0{}^\intercal\Gamma_0\right)^{-1}\Gamma_0{}^\intercal\right\|_F\cdot\left\|\Gamma_0\Lambda_0-\widehat{\Gamma}_0\widehat{\Lambda}_0\right\|_F \\
&\qquad+\left\|\widehat{\Lambda}_0\right\|_F\cdot\left(\left\|\left(\Gamma_0{}^\intercal\Gamma_0\right)^{-1}\Gamma_0{}^\intercal\left(\widehat{\Gamma}_0-\Gamma_0A\right)\right\|_F+\left\|\left(\Gamma_0{}^\intercal\Gamma_0\right)^{-1}\Gamma_0{}^\intercal\Gamma_0\left(A-\mathbf{I}_K\right)\right\|_F\right) \\
&=\left(\left\|\left(\Gamma_0{}^\intercal\Gamma_0\right)^{-1}\Gamma_0{}^\intercal\right\|_F+\left\|\widehat{\Lambda}_0\right\|_F\cdot\left\|\left(\Gamma_0{}^\intercal\Gamma_0\right)^{-1}\Gamma_0{}^\intercal\right\|_F+\left\|\widehat{\Lambda}_0\right\|_F\right)\cdot O_p\left(\frac{1}{\sqrt{n}}\right).
\end{aligned}
$$

B.4 Proof for Theorem 3

Step 1. $\left\|\widehat{\tilde{\Lambda}}_d-\tilde{\Lambda}_d\right\|_F=O_p\left(\frac{1}{\sqrt{n}}\right)$.

Find that

$$
\begin{aligned}
\left\|\widehat{\tilde{\Lambda}}_0-\tilde{\Lambda}_0\right\|_F &= \left\|\widehat{\Lambda}_0^{-1}\left(\Lambda_0-\widehat{\Lambda}_0\right)\Lambda_0{}^{-1}\right\|_F \\
&\le \left\|\widehat{\Lambda}_0^{-1}\right\|_F\cdot\left\|\Lambda_0-\widehat{\Lambda}_0\right\|_F\cdot\left\|\Lambda_0{}^{-1}\right\|_F
\end{aligned}
$$

and $\left\|\widehat{\Lambda}_0^{-1}\right\|_F=O_p(1)$.

Step 2. $\|\hat{p}-p\|=O_p\left(\frac{1}{\sqrt{n}}\right)$ and $\|\hat{\mu}-\mu\|=O_p\left(\frac{1}{\sqrt{n}}\right)$ as $n\to\infty$.

Firstly,

$$
\hat{p}_{D,Z}=\begin{pmatrix}\frac{1}{n}\sum_{i=1}^n\mathbf{1}\{D_i=0,Z_i=z^1\}\\ \vdots\\ \frac{1}{n}\sum_{i=1}^n\mathbf{1}\{D_i=1,Z_i=z^K\}\end{pmatrix}
$$

is $O_p\left(\frac{1}{\sqrt{n}}\right)$ from the central limit theorem. Thus,

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

is also $O_p\left(\frac{1}{\sqrt{n}}\right)$.

Secondly, let

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

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

and let $\widehat{\partial\phi}$ and $\widehat{\partial m}$ be the estimators of $\partial\phi$ and ∂m by taking their sample analogues, plugging in $\hat{p}$ and $\hat{\tilde{\lambda}}$. μ is estimated with

$$\hat{\mu} = \widehat{\partial\phi}^{\intercal}\left(\widehat{\partial\phi}\widehat{\partial\phi}^{\intercal}\right)^{-1}\widehat{\partial m}.$$

$\widehat{\partial\phi}$ and $\widehat{\partial m}$ converge to $\partial\phi$ and ∂m at the rate of $\frac{1}{\sqrt{n}}$ in $\|\cdot\|_F$ since each element of $\widehat{\partial\phi}$ and $\widehat{\partial m}$ is a ratio of a product of $\sqrt{n}$-consistent estimators over a product of $\sqrt{n}$-consistent estimators which converge to a nonzero constant. For example,

$$\begin{aligned} &\mathbf{E}\left[\frac{\partial}{\partial\tilde{\lambda}_{jk,d}}\phi_A\big(W_i, W_{i'}; \tilde{\lambda}, p\big)\right] \\ &\quad = \Pr\{Y_i = y, X_i = x | D_i = d, Z_i = z^j\} - \Pr\{Y_i = y | D_i = d, Z = z^j\} \cdot \Pr\{X_i = x | U_i = u^k\} \\ &\qquad\qquad - \Pr\{X_i = x | D_i = d, Z = z^j\} \cdot \Pr\{Y_i(d) = y | U_i = u^k\} \end{aligned}$$

is estimated with

$$\begin{aligned} &\binom{n}{2}^{-1}\sum_{i\neq i'}\frac{\frac{1}{2}\mathbf{1}\{Y_i = y, D_i = d, X_i = x, Z_i = z^j\}}{\hat{p}_{D,Z}(d,j)} \\ &- \binom{n}{2}^{-1}\sum_{i\neq i'}\frac{\frac{1}{2}\mathbf{1}\{Y_i = y, D_i = d, Z_i = z^j\}\cdot\sum_{j'=1}^{K}\hat{\tilde{\lambda}}_{j'k,d}\mathbf{1}\{D_{i'} = d, X_{i'} = x, Z_{i'} = z^{j'}\}}{\hat{p}_{D,Z}(d,j)\cdot\hat{p}_U(k)} \\ &- \binom{n}{2}^{-1}\sum_{i\neq i'}\frac{\frac{1}{2}\mathbf{1}\{D_i = d, X_i = x, Z_i = z^j\}\cdot\sum_{j'=1}^{K}\hat{\tilde{\lambda}}_{j'k,d}\mathbf{1}\{Y_{i'} = y, D_{i'} = d, Z_{i'} = z^{j'}\}}{\hat{p}_{D,Z}(d,j)\cdot\hat{p}_U(k)}. \end{aligned}$$

The exact expression of $\partial\phi$ and ∂m is given in the proof for Lemma 1. Then,

$$
\begin{aligned}
\hat{\mu} - \mu &= \widehat{\partial\phi}^\intercal \left(\widehat{\partial\phi}\widehat{\partial\phi}^\intercal\right)^{-1} \widehat{\partial m} - \partial\phi^\intercal \left(\partial\phi\partial\phi^\intercal\right)^{-1} \partial m \\
&\leq \left\| \widehat{\partial\phi}^\intercal \left(\widehat{\partial\phi}\widehat{\partial\phi}^\intercal\right)^{-1} \right\|_F \cdot \left\| \widehat{\partial m} - \partial m \right\|_F + \left\| \widehat{\partial\phi}^\intercal - \partial\phi^\intercal \right\|_F \cdot \left\| \left(\partial\phi\partial\phi^\intercal\right)^{-1} \partial m \right\|_F \\
&\quad + \left\| \widehat{\partial\phi}^\intercal \left(\partial\phi\partial\phi^\intercal\right)^{-1} \right\|_F \cdot \left\| \partial\phi\partial\phi^\intercal - \widehat{\partial\phi}\widehat{\partial\phi}^\intercal \right\|_F \cdot \left\| \left(\widehat{\partial\phi}\widehat{\partial\phi}^\intercal\right)^{-1} \partial m \right\|_F \\
&= O_p\left(\frac{1}{\sqrt{n}}\right).
\end{aligned}
$$

Step 3. Find that

$$
\begin{aligned}
&\psi\left(W_i, W_{i'}; \hat{\theta}, \hat{\tilde{\lambda}}, \hat{p}, \hat{\mu}\right) \\
&= \psi\left(W_i, W_{i'}; \theta, \tilde{\lambda}, p, \mu\right) \\
&\qquad + \frac{\partial}{\partial\theta}\psi(W_i, W_{i'}; \bar{\theta}, \bar{\lambda}, \bar{p}, \bar{\mu}) \cdot \left(\hat{\theta} - \theta\right) + \frac{\partial}{\partial\tilde{\lambda}}\psi(W_i, W_{i'}; \bar{\theta}, \bar{\lambda}, \bar{p}, \bar{\mu})^\intercal \cdot \left(\hat{\tilde{\lambda}} - \tilde{\lambda}\right) \\
&\qquad + \frac{\partial}{\partial p}\psi(W_i, W_{i'}; \bar{\theta}, \bar{\lambda}, \bar{p}, \bar{\mu})^\intercal \cdot (\hat{p} - p) + \frac{\partial}{\partial\mu}\psi(W_i, W_{i'}; \bar{\theta}, \bar{\lambda}, \bar{p}, \bar{\mu})^\intercal \cdot (\hat{\mu} - \mu) \\
&= \psi\left(W_i, W_{i'}; \theta, \tilde{\lambda}, p, \mu\right) \\
&\qquad - \left(\hat{\theta} - \theta\right) + \frac{\partial}{\partial\tilde{\lambda}}m(W_i, W_{i'}; \bar{\theta}, \bar{\lambda}, \bar{p})^\intercal \cdot \left(\hat{\tilde{\lambda}} - \tilde{\lambda}\right) - \mu^\intercal \frac{\partial}{\partial\tilde{\lambda}}\phi(W_i, W_{i'}; \bar{\lambda}, \bar{p})^\intercal \cdot \left(\hat{\tilde{\lambda}} - \tilde{\lambda}\right) \\
&\qquad + \frac{\partial}{\partial p}m(W_i, W_{i'}; \bar{\theta}, \bar{\lambda}, \bar{p})^\intercal \cdot (\hat{p} - p) - \mu^\intercal \frac{\partial}{\partial p}\phi(W_i, W_{i'}; \bar{\lambda}, \bar{p})^\intercal \cdot (\hat{p} - p) \\
&\qquad + \phi(W_i, W_{i'}; \bar{\lambda}, \bar{p})^\intercal (\hat{\mu} - \mu)
\end{aligned}
$$

with $(\bar{\theta}, \bar{\lambda}, \bar{p}, \bar{\mu})$ being the intermediate values between $\left(\theta, \tilde{\lambda}, p, \mu\right)$ and $\left(\hat{\theta}, \hat{\tilde{\lambda}}, \hat{p}, \hat{\mu}\right)$. Therefore,

$$
\begin{aligned}
&\sqrt{n}\left(\hat{\theta} - \theta\right) \\
&\quad = \sqrt{n}\binom{n}{2}^{-1} \sum_{i<i'} \psi\left(W_i, W_{i'}; \theta, \tilde{\lambda}, p, \mu\right) \\
&\qquad + \binom{n}{2}^{-1} \sum_{i<i'} \left(\frac{\partial}{\partial\tilde{\lambda}}m(W_i, W_{i'}; \bar{\theta}, \bar{\lambda}, \bar{p})^\intercal - \mu^\intercal \frac{\partial}{\partial\tilde{\lambda}}\phi(W_i, W_{i'}; \bar{\lambda}, \bar{p})^\intercal\right) \cdot \sqrt{n}\left(\hat{\tilde{\lambda}} - \tilde{\lambda}\right) \\
&\qquad + \binom{n}{2}^{-1} \sum_{i<i'} \left(\frac{\partial}{\partial p}m(W_i, W_{i'}; \bar{\theta}, \bar{\lambda}, \bar{p})^\intercal - \mu^\intercal \frac{\partial}{\partial p}\phi(W_i, W_{i'}; \bar{\lambda}, \bar{p})^\intercal\right) \cdot \sqrt{n}\,(\hat{p} - p) \\
&\qquad + \binom{n}{2}^{-1} \sum_{i<i'} \phi(W_i, W_{i'}; \bar{\lambda}, \bar{p})^\intercal \cdot \sqrt{n}\,(\hat{\mu} - \mu).
\end{aligned}
$$

The intermediate values $\left(\bar{\theta}, \bar{\lambda}, \bar{p}, \bar{\mu}\right)$ depend on $(W_i, W_{i'})$. From the construction of the Neyman orthogonal score and the consistency of the nuisance parameter estimators,

$$\begin{aligned}
&\binom{n}{2}^{-1} \sum_{i<i'} \left(\frac{\partial}{\partial \tilde{\lambda}} m(W_i, W_{i'}; \bar{\theta}, \bar{\lambda}, \bar{p}) - \frac{\partial}{\partial \tilde{\lambda}} \phi(W_i, W_{i'}; \bar{\lambda}, \bar{p}) \mu \right) \\
&\qquad \xrightarrow{p} \mathbf{E}\left[\frac{\partial}{\partial \tilde{\lambda}} m(W_i, W_{i'}; \bar{\theta}, \bar{\lambda}, \bar{p}) \right] - \mathbf{E}\left[\frac{\partial}{\partial \tilde{\lambda}} \phi(W_i, W_{i'}; \bar{\lambda}, \bar{p}) \right] \mu = \mathbf{0}_{2K^2}
\end{aligned}$$

and similarly for $\binom{n}{2}^{-1} \sum_{i<i'} \left(\frac{\partial}{\partial p} m(W_i, W_{i'}; \bar{\theta}, \bar{\lambda}, \bar{p}) - \frac{\partial}{\partial p} \phi(W_i, W_{i'}; \bar{\lambda}, \bar{p}) \mu \right)$. From $\sqrt{n}\left(\hat{\tilde{\lambda}} - \tilde{\lambda}\right) = O_p(1)$, $\sqrt{n}\left(\hat{p} - p\right) = O_p(1)$, $\sqrt{n}\left(\hat{\mu} - \mu\right) = O_p(1)$ and the asymptotic theory for U statistics, we get

$$\sqrt{n}\left(\hat{\theta} - \theta\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{\psi}\left(W_i; \theta, \tilde{\lambda}, p, \mu\right) + o_p(1)$$

where

$$\tilde{\psi}\left(w; \theta, \tilde{\lambda}, p, \mu\right) = \mathbf{E}\left[\psi(w, W_i; \theta, \tilde{\lambda}, p, \mu)\right].$$