Distributional Treatment Effect with Latent Rank Invariance*

Myungkou Shin†

June 6, 2024

Abstract

Treatment effect heterogeneity is of a great concern when evaluating the treatment. However, even with a simple case of a binary random treatment, the distribution of treatment effect is difficult to identify due to the fundamental limitation that we cannot observe both treated potential outcome and untreated potential outcome for a given individual. This paper assumes a conditional independence assumption that the two potential outcomes are independent of each other given a scalar latent variable. Using two proxy variables, we identify conditional distribution of the potential outcomes given the latent variable. To pin down the location of the latent variable, we assume strict monotonicty on some functional of the conditional distribution; with specific example of strictly increasing conditional expectation, we label the latent variable as 'latent rank' and motivate the identifying assumption as 'latent rank invariance.'

Keywords: distributional treatment effect, proximal inference, latent heterogeneity, nonnegative matrix factorization

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.

†Department of Economics, University of Oxford. email: myungkou.shin@economics.ox.ac.uk

1 Introduction

The fundamental limitation that we in general cannot simultaneously observe the two potential outcomes—treated potential outcome and untreated potential outcome—for a given individual 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 of the treatment effect distribution provide insights into the treatment effect distribution and thus help researchers with policy recommendations. However, there still remain a lot of research questions that can only be answered from the treatment effect distributions; 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 distributional treatment effect.

When we believe that there is no underlying individual-level heterogeneity and thus the two potential outcomes are independence of each other, 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. In light of this observation, we propose an econometric framework where there exists a scalar latent variable that captures the individual-level heterogeneity in terms of the dependence between the two potential outcomes. Assuming that the two potential outcomes are independent of each other conditioning on the latent variable, the task of identifying the joint distribution of the two potential outcomes becomes that of identifying the two conditional distributions of a potential outcomes given the latent variable and the marginal distribution of the latent variable. Let $Y(1)$ denote the treated potential outcome, $Y(0)$ denote the untreated potential outcome and $U \in \mathcal{U} \subset \mathbb{R}$ denote the individual-level latent variable. When

$$Y(1) \perp\!\!\!\perp Y(0) \;\Big|\; U, \tag{1}$$

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

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

The assumption that there exists a latent variable U_i that captures the dependence between the two potential outcomes is not new to the literature. A condition in the same spirit as (1) 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) and more.) The rank invariance condition assumes that the rank of one potential outcome is retained as the rank of the other potential outcome. Let $F_{Y(1)}$ denote the marginal distribution of the treated potential outcome and let $F_{Y(0)}$ denote the marginal distribution of the treated potential outcome. The rank invariance assumes that $U = F_{Y(1)}(Y(1)) = F_{Y(0)}(Y(0))$ and the one scalar variable U captures all the dependence between the two potential outcomes. The conditional independence of $Y(1)$ and $Y(0)$ given U directly follows.

However, the usage of the rank invariance condition is mostly limited to the quantile treatment effect and not applied to the distributional treatment effect, due to the fact that it imposes excessive restriction on the joint distribution of the two potential outcomes.[1] When we assume that the ranks of the two potential outcomes are the same, the conditional distribution of one potential outcome given the other potential outcome becomes degenerate. Thus, while the rank invariance is a sensible condition to utilize when the goal is to identify the quantile treatment effect, since we are not interested in the conditional distribution of one potential outcome given the other, it may be imposing too much restriction on the dependence between the two potential outcomes when we are interested in the joint distribution of the two potential outcomes.

As a byproduct of relaxing the rank invariance assumption and not assuming that there is some known function, such as $F_{Y(1)}$, to retrieve U directly from Y, it becomes nontrivial to identify the conditional distributions and the marginal distributions involving the latent variable. For that, we assume that there are two additional proxy variables X, Z, observable to the econometrician: Hu and Schennach (2008); Miao et al. (2018); Deaner (2023); Kedagni (2023) and more. With the two proxy variables that are conditionally independent of each other and the potential outcomes given the latent variable, we characterize a pair of conditional densities—the conditional density of $Y(1)$ given U and the conditional density of $Y(0)$ given U—for each u in the support of U.

For the last step of the identification argument, we assume that there exists a functional of

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

the joint distribution of the potential outcomes given the latent variable, strictly increasing in the latent variable U. An example of such a functional is conditional expectation. Suppose that the two conditional expectations $\mathbf{E}[Y(1)|U=u]$ and $\mathbf{E}[Y(0)|U=u]$ are strictly increasing in u. Using the strict monotonicity, we can identify the marginal distribution of (a transformation of) U and integrate the product of the conditional densities across the support of U. Within this example, the latent variable U can be thought of as the rank of the expected mean, which is invariant across the two treatment regimes.

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. 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 we do: Carneiro et al. (2003); Gautier and Hoderlein (2015); Noh (2023). Gautier and Hoderlein (2015) is similar to our paper in the sense that they also consider a conditional independence assumption given a latent variable; the difference is that the distribution of the latent variable is identified through treatment assignment model with instruments in their paper while it is identified through conditionally independent proxy variables in this paper. Carneiro et al. (2003) is closest to this paper in that they also assume conditional independence of potential outcomes given a latent variable and use repeated measurements on the latent variable, similar to our proxies. However, by assuming linear structure with a factor model, Carneiro et al. (2003) allows for multi-dimensional latent vector while our result allows for more flexibility in the conditional distribution of the potential outcomes given the latent variable, at the cost of assuming that U is a scalar.

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 procedure based on a nonnegative matrix factorization problem and proves the uniform consistency for the joint distribution of the potential outcomes. Section 4 extends the main identification result to account for treatment endogneity. Section 5 applies the estimation procedure to 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.

$$Y_i = D_i \cdot Y_i(1) + (1 - D_i) \cdot Y_i(0). \tag{2}$$

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 some restrictions on the joint distribution of $Y_i(1)$ and $Y_i(0)$ and bypassing the fundamental limitation that we observe only one potential outcome for a given individual. 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, we assume that the treatment D_i is random, with regard to the potential outcomes, one of the proxy variables and the latent variable U_i.

Assumption 1. *(random treatment)* $\big(Y_i(1), Y_i(0), X_i, U_i\big) \perp\!\!\!\perp D_i$.

Assumption 1 is already more than enough to identify numerous treatment effect parameters such as average treatment effect (ATE), quantile treatment effect (QTE) and more. However, the random treatment does not give us sufficient identifying power to identify the distribution of treatment effect, since it still suffers from the limitation that we cannot simultaneously observe $Y_i(1)$ and $Y_i(0)$ for a given individual.

Thus, we impose restrictions on the joint distribution of $Y_i(1)$ and $Y_i(0)$ by assuming that the latent variable U_i captures all of the dependence between the two potential outcomes and the two proxy variables.

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

The conditional independence assumption is a nontrivial restriction on the joint distribution of $Y_i(1)$ and $Y_i(0)$. 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 $U_i = \big(Y_i(1), Y_i(0)\big)$.

Example 1. Consider a short panel with nonseparable model: for $t = 1, 2, 3$ and $d = 0, 1$,

$$Y_{it}(d) = g_d\big(V_{it}, \varepsilon_{it}(d)\big), \tag{3}$$
$$Y_{it} = D_{it} \cdot Y_{it}(1) + (1 - D_{it}) \cdot Y_{it}(0).$$

A potential outcome $Y_{it}(d)$ is a nonlinear function of two latent variables V_{it} and $\varepsilon_{it}(d)$. Note that V_{it} appears in the model twice; both 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. For simplicity, assume $D_{i1} = 0$ and $D_{i3} = 1$ with probability one. Then by assuming that $\{V_{it}\}_{t=1}^{3}$ is Markovian and that $\{V_{it}\}_{t=1}^{3}, \varepsilon_{i1}(0), \varepsilon_{i2}(1), \varepsilon_{i2}(0)$ and $\varepsilon_{i3}(1)$ are mutually independent, Assumption 2 is satisfied by letting $Y_i = Y_{i2}, X_i = Y_{i1}, Z_i = Y_{i3}$ and $U_i = V_{i2}$.

Example 1 constructs a nonseparable panel model where a past outcome and a future outcome can be used as the two proxy variables satisfying the conditional independence assumption. The key element of the nonseparable panel model in Example 1 is that there is a common shock V_{it} that is applied to both treated and untreated potential outcomes and that V_{it} is independent of all the treatment-status-specific shocks $\{\varepsilon_{is}(d)\}_{s,d}$ while dependence within $\{V_{is}\}_s$ themselves is restricted to be first-order Markovian. When the treatment D_{i2} is assigned randomly, both Assumption 1 and Assumption 2 hold, under the nonseparable panel model.

An intuitive way to motivate the role of V_{it} in this example is to label V_{it} as a 'latent rank' and assume rank invariance. Assume that $\mathbf{E}\left[Y_{it}(1)|V_{it} = v\right]$ and $\mathbf{E}\left[V_{it}(0)|V_{it} = v\right]$ are monotone increasing in v. In addition, WLOG suppose that in the data generating process, 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. Then, at the interim stage where only V_{it} has been drawn, the two potential outcome variables have the same 'latent rank' in the sense that their expected values $\mathbf{E}\left[Y_{it}(1)|V_{it}\right]$ and $\mathbf{E}\left[Y_{it}(0)|V_{it}\right]$ have the same rank in their respective distributions. A similar motivational assumption can be made with other average measures such as median or mode. With this interpretation, the sufficient condition for Assumption 2 in Example 1 is that the latent rank process is independent of all the other random elements of the model and the latent rank process itself is first-order Markovian; any intertemporal and intratemporal dependence among potential outcomes is captured with the first-order Markovian latent rank.

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 sample and the untreated sample. 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 the conditional independence, we obtain the following integral representation: for $x, z \in \mathbb{R}$,

$$f_{Y=y,X|D=d,Z}(x|z) = \int f_{Y(d)|U}(y|u) \cdot f_{X|U}(x|u) \cdot f_{U|D=d,Z}(u|z)du, \tag{4}$$
$$f_{X|D=d,Z}(x|z) = \int f_{X|U}(x|u) \cdot f_{U|D=d,Z}(u|z)du.$$

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 $\mathcal{L}^1(\mathbb{R})$:

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

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:

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

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:

$$Mf = \int_{\mathbb{R}} \int_{\mathbb{R}} (y_1 + y_0) f(y_1, y_0) dy_1 dy_0.$$

Alternatively, if two functionals such as

$$\begin{aligned} M_1 f &= \int_{\mathbb{R}} y_1 f_{Y(1)|U}(y_1|u) dy_1, \\ M_2 f &= \int_{\mathbb{R}} y_0 f_{Y(0)|U}(y_0|u) dy_0 \end{aligned}$$

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.

Assumption 3 formally states the completeness condition on the conditional density $f_{X|D=d,Z}$ and the latent rank assumption on $f_{Y(1),Y(0)|U}$ and U_i. Theorem 1 contains the identification result.

Assumption 3. *Assume*

- ***a.*** *(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.*

- ***b.*** *(completeness) The integral operators $L_{X|U}, L_{X|D=1,Z}$ and $L_{X|D=0,Z}$ are injective on $\mathcal{L}^1(\mathbb{R})$.*

c. *(no repeated eigenvalue) For any* $u \neq u'$*,*

$$\Pr\left\{f_{Y(1)|U}(Y_i|u) \neq f_{Y(1)|U}(Y_i|u')|D_i = 1\right\} \cdot \Pr\left\{f_{Y(0)|U}(Y_i|u) \neq f_{Y(0)|U}(Y_i|u')|D_i = 0\right\} > 0.$$

d. *(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*

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

Theorem 1. *Under Assumptions 1-3, the joint density of* $Y_i(1), Y_i(0)$ *and* X_i *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., $\Pr\{Y_i(1) \geq Y_i(0)\}, \Pr\{Y_i(1) \geq Y_i(0)|Y_i(0)\}$ and etc.

3 Implementation

In this section, we propose an estimation method, based on the identification result from the previous section.

3.1 Nonnegative matrix factorization

For this subsection, let us assume that U_i is finite.

Assumption 4. $\mathcal{U} = \{u^1, \cdots, u^K\}$.

While Assumption 3.d does not hold anymore, the identification result discussed in the previous section still holds, thanks to the finiteness of $\mathcal{U}$; marginal probability of U_i can be retrieved, without any ordering on $\left\{f_{X|U}(\cdot|u)\right\}_{u=u^1}^{u^K}$. Now that U_i is finite, the estimation of the joint distribution of $Y_i(1)$ and $Y_i(0)$ reduces down to the estimation of the $2K$ conditional densities $\left\{f_{Y(1)|U}(\cdot|u), f_{Y(0)|U}(\cdot|u)\right\}_{u=u^1}^{u^K}$. As with the identification strategy, we rely on the proxy variable Z_i to create variation in the conditional density of X_i given U_i: the mixture component distribution, in the terminology of a finite mixture model.

For each of the three variables Y_i, X_i, Z_i, construct a partition on $\mathbb{R}$:

$$\left\{\mathcal{Y}_m = \left(y^{m-1}, y^m\right]\right\}_{m=1}^{M_y}, \quad \left\{\mathcal{X}_m = \left(x^{m-1}, x^m\right]\right\}_{m=1}^{M_x}, \quad \left\{\mathcal{Z}_m = \left(z^{m-1}, z^m\right]\right\}_{m=1}^{M_z}$$

Let $M = M_y \cdot M_x$ and 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, we construct two $M \times M_z$ matrices whose elements are conditional probability (Y_i, X_i) given (D_i, Z_i): let

$$\mathbf{H}_d = \begin{pmatrix} \Pr\{(Y_i, X_i) \in \mathcal{W}_1 | D_i = d, Z_i \in \mathcal{Z}_1\} & \cdots & \Pr\{(Y_i, X_i) \in \mathcal{W}_1 | D_i = d, Z_i \in \mathcal{Z}_{M_z}\} \\ \vdots & \ddots & \vdots \\ \Pr\{(Y_i, X_i) \in \mathcal{W}_M | D_i = d, Z_i \in \mathcal{Z}_1\} & \cdots & \Pr\{(Y_i, X_i) \in \mathcal{W}_M | D_i = d, Z_i \in \mathcal{Z}_{M_z}\} \end{pmatrix}$$

for each $d = 0, 1$. From Assumption 2, both $\mathbf{H}_0$ and $\mathbf{H}_1$ decompose into a multiplication of two matrices: for each $d = 0, 1$, $\mathbf{H}_d = \Gamma_d \cdot \Lambda_d$ where

$$\Gamma_d = \begin{pmatrix} \Pr\left\{(Y_i(d), X_i) \in \mathcal{W}_1 | U_i = u^1\right\} & \cdots & \Pr\left\{(Y_i(d), X_i) \in \mathcal{W}_1 | U_i = u^K\right\} \\ \vdots & \ddots & \vdots \\ \Pr\left\{(Y_i(d), X_i) \in \mathcal{W}_M | U_i = u^1\right\} & \cdots & \Pr\left\{(Y_i(d), X_i) \in \mathcal{W}_M | U_i = u^K\right\} \end{pmatrix}$$

$$\Lambda_d = \begin{pmatrix} \Pr\left\{U_i = u^1 | D_i = d, Z_i \in \mathcal{Z}_1\right\} & \cdots & \Pr\left\{U_i = u^1 | D_i = d, Z_i \in \mathcal{Z}_{M_z}\right\} \\ \vdots & \ddots & \vdots \\ \Pr\left\{U_i = u^K | D_i = d, Z_i \in \mathcal{Z}_1\right\} & \cdots & \Pr\left\{U_i = u^K | D_i = d, Z_i \in \mathcal{Z}_{M_z}\right\} \end{pmatrix}$$

The equation $\mathbf{H}_d = \Gamma_d \cdot \Lambda_d$ shows us that under the finiteness of U_i, the conditional density model in (4) is indeed a finite mixture model. For each subpopulation $\{i : D_i = d, Z_i \in C_m\}$, there is a column in the matrix Λ_d which denotes the (conditional) probability of U_i in that subpopulation. Then, the conditional density of (Y_i, X_i) in that subpopulation admits a finite 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 $\mathbf{H}_d = \Gamma_d \cdot \Lambda_d$ discretizes the relationship and collects them across the partition $\{\mathcal{Z}_m\}_{m=1}^{M_z}$ on the support of Z_i.

Recall that Assumption 3.b assumes that the conditional density of X_i given $D_i = d, Z_i$ is complete. In the discretized version of the model, the completeness condition translates to Λ_0 and Λ_1 having full rank, with a properly chosen partition $\{\mathcal{Z}_m\}_{m=1}^{M_z}$. When Λ_0 and Λ_1 are full rank matrices, we can define pseudo-inverses and identify the conditional density of $(Y_i(d), X_i)$ given U_i from the following linear equations:

$$\begin{aligned} &\left(f_{Y,X|D=d,Z}(y, x|\mathcal{Z}_1) \quad \cdots \quad f_{Y,X|D=d,Z}(y, x|\mathcal{Z}_{M_z})\right) \\ &\quad = \left(f_{Y(d),X|U}(y, x|u^1) \quad \cdots \quad f_{Y(d),X|U}(y, x|u^K)\right) \Lambda_d \qquad \forall (y, x) \in \mathbb{R}^2 \end{aligned}$$

for each $d = 0, 1$.[2] Thus, under the finiteness assumption, the estimation of the $2K$ conditional densities $\left\{f_{Y(1)|U}(\cdot|u), f_{Y(0)|U}(\cdot|u)\right\}_{u=u^1}^{u^K}$ further reduces down to the estimation of the two matrices Λ_0 and Λ_1; the rest of the estimation can be done nonparametrically.

To estimate the mixture weight matrices Λ_0 and Λ_1, construct sample analogues of the conditional probability matrices $\mathbf{H}_0$ and $\mathbf{H}_1$: let

$$\mathbb{H}_d = \begin{pmatrix} \frac{\sum_{i=1}^n \mathbf{1}\{(Y_i,X_i)\in\mathcal{W}_1, D_i=d, Z_i\in\mathcal{Z}_1\}}{\sum_{i=1}^n \mathbf{1}\{D_i=d, Z_i\in\mathcal{Z}_1\}} & \cdots & \frac{\sum_{i=1}^n \mathbf{1}\{(Y_i,X_i)\in\mathcal{W}_1, D_i=d, Z_i\in\mathcal{Z}_{M_z}\}}{\sum_{i=1}^n \mathbf{1}\{D_i=d, Z_i\in\mathcal{Z}_{M_z}\}} \\ \vdots & \ddots & \vdots \\ \frac{\sum_{i=1}^n \mathbf{1}\{(Y_i,X_i)\in\mathcal{W}_M, D_i=d, Z_i\in\mathcal{Z}_1\}}{\sum_{i=1}^n \mathbf{1}\{D_i=d, Z_i\in\mathcal{Z}_1\}} & \cdots & \frac{\sum_{i=1}^n \mathbf{1}\{(Y_i,X_i)\in\mathcal{W}_M, D_i=d, Z_i\in\mathcal{Z}_{M_z}\}}{\sum_{i=1}^n \mathbf{1}\{D_i=d, Z_i\in\mathcal{Z}_{M_z}\}} \end{pmatrix}.$$

Each column of $\mathbb{H}_0$ is a (discretized) conditional distribution function of (Y_i, X_i) given $\big(D_i = 0, Z_i\big)$ and each column of $\mathbb{H}_1$ is a (discretized) conditional distribution function of (Y_i, X_i) given $\big(D_i = 1, Z_i\big)$. To estimate Λ_0 and Λ_1, construct a nonnegative matrix factorization problem:

$$\min_{\Lambda_0,\Lambda_1,\Gamma_0,\Gamma_1} \left\|\mathbb{H}_0 - \Gamma_0\Lambda_0\right\|_F^{\,2} + \left\|\mathbb{H}_1 - \Gamma_1\Lambda_1\right\|_F^{\,2} \tag{5}$$

subject to linear constraints that

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

where $P = I_{M_x}\otimes\iota_{M_y}{}^{\intercal}$ and quadratic constraints that Γ_0 and Γ_1 satisfy the conditional independence

$$\begin{aligned}&\Pr\left\{(Y_i(d), X_i) \in \mathcal{Y}_m \times \mathcal{X}_{m'} | U_i = u^k\right\} \\ &\quad = \left(\sum_{l=1}^{M_x} \Pr\left\{(Y_i(d), X_i) \in \mathcal{Y}_m \times \mathcal{X}_l | U_i = u^k\right\}\right) \cdot \left(\sum_{l=1}^{M_y} \Pr\left\{(Y_i(d), X_i) \in \mathcal{Y}_l \times \mathcal{X}_{m'} | U_i = u^k\right\}\right)\end{aligned}$$

for each (m, m'). ι_x is a x-dimensional column vector of ones. Let $\widehat{\Lambda}_0$, $\widehat{\Lambda}_1$, $\widehat{\Gamma}_0$ and $\widehat{\Gamma}_1$ denote the solution to the minimization problem.

Note that the objective function in (5) is quadratic when either (Λ_0, Λ_1) or (Γ_0, Γ_1) is fixed. Thus, we suggest an iterative algorithm to solve the minimization problem.

[2] A slight abuse of notation is used here to denote the conditional density of (Y_i, X_i) given $(D_i = d, Z_i \in C_m)$.

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:

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

subject to $\Lambda_0 \in \mathbb{R}_+{}^{K \times J_0}, \Lambda_1 \in \mathbb{R}_+{}^{K \times J_1}, \iota_K{}^\intercal \Lambda_0 = \iota_{J_0}{}^\intercal$ and $\iota_K{}^\intercal \Lambda_1 = \iota_{J_1}{}^\intercal$.

3. *(Update* Γ*)* Given $\left(\Lambda_0^{(s)}, \Lambda_1^{(s)}\right)$, solve the following quadratic program:

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

subject to $\Gamma_0 \in \mathbb{R}_+{}^{M \times K}, \Gamma_1 \in \mathbb{R}_+{}^{M \times K}, \iota_M{}^\intercal \Gamma_0 = \iota_K{}^\intercal, \iota_M{}^\intercal \Gamma_1 = \iota_K{}^\intercal$ and $P\Gamma_0 = P\Gamma_1$. Compute marginal probabilities from $\left(\tilde{\Gamma}_0^{(s+1)}, \tilde{\Gamma}_1^{(s+1)}\right)$ and expand to construct $\left(\Gamma_0^{(s+1)}, \Gamma_1^{(s+1)}\right)$.

4. Repeat **2-3** until convergence.

As discussed above, the number of columns in $\mathbf{H}_0$ and $\mathbf{H}_1$ have to at least K to estimate full rank Λ_0 and Λ_1; $M_z \geq K$. Thus, to initialize $\Gamma_0^{(0)}, \Gamma_1^{(1)}$, we can consider every K-combination of columns from $\mathbb{H}_d$ for small enough M_z and choose the combination that minimizes the objective: $\binom{M_z}{K}$ initial values. Likewise, we can consider randomly drawn K set of weights that sum to one and use the weighted sums of columns of $\mathbb{H}_d$ as an initial value. 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.

The following Assumption replaces Assumption 3.b-c in the context of Assumption 4.

Assumption 5.

a. *Λ_0 and Λ_1 have rank K.*

b. *For each $k = 1, \cdots, K$ and $d = 0, 1$, let*

$$p_k = \left(\Pr\left\{X_i \in \mathcal{X}_1 | U_i = u^k\right\} \quad \cdots \quad \Pr\left\{X_i \in \mathcal{X}_{M_x} | U_i = u^k\right\}\right)^\intercal,$$
$$q_{dk} = \left(\Pr\left\{Y_i(d) \in \mathcal{Y}_1 | U_i = u^k\right\} \quad \cdots \quad \Pr\left\{Y_i(d) \in \mathcal{Y}_{M_y} | U_i = u^k\right\}\right)^\intercal.$$

For any $k \neq k'$, $q_{0k} \neq q_{0k'}$ and $q_{1k} \neq q_{1k'}$. In addition, $p_1, \cdots p_K$ are linearly independent.

Assumption 5 implicitly assumes that $M_x \geq K$. The restriction that $M_x, M_z \geq K$ is intuitive in the sense that we use the variation in the conditional density of one proxy variable given another 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.

Theorem 2. *Assumptions 1-2, 4-5 hold. Up to some permutation on* $\{u^1, \cdots, u^K\}$,

$$\widehat{\Lambda}_0 \xrightarrow{p} \Lambda_0 \quad \text{and} \quad \widehat{\Lambda}_1 \xrightarrow{p} \Lambda_1$$

as $n \to \infty$.

Proof. See Appendix. □

Given the estimates of the two mixture weights matrices Λ_0 and Λ_1, we can construct an estimator for the joint distribution of $Y_i(1)$ and $Y_i(0)$. Firstly, estimate the conditional distribution of the potential outcomes given the latent variable U_i. Let

$$\mathbb{F}_{Y|D=d,Z}(y|\mathcal{Z}_m) = \frac{1}{\sum_{i=1}^n \mathbf{1}\{D_i = d, Z_i \in \mathcal{Z}_m\}} \sum_{i=1}^n \mathbf{1}\{Y_i \leq y, D_i = d, Z_i \in \mathcal{Z}_m\} \quad \forall y \in \mathcal{Y}$$

for each $m = 1, \cdots, M_z$ and $d = 0, 1$. Let

$$\left(\hat{F}_{Y(d)|U}\left(y|u^1\right) \quad \cdots \quad \hat{F}_{Y(d)|U}\left(y|u^K\right)\right) = \left(\mathbb{F}_{Y|D=d,Z}(y|\mathcal{Z}_1) \quad \cdots \quad \mathbb{F}_{Y|D=d,Z}(y|\mathcal{Z}_{M_z})\right)\left(\widehat{\Lambda}_d\right)^{-1}.$$

Secondly, estimate the marginal distribution of the latent variable U_i. Let

$$\mathbb{F}_{Z|D=d}(\mathcal{Z}_m) = \frac{\sum_{i=1}^n \mathbf{1}\{D_i = d, Z_i \in \mathcal{Z}_m\}}{\sum_{i=1}^n \mathbf{1}\{D_i = d\}}$$

for each $m = 1, \cdots, M_z$ and $d = 0, 1$. Let

$$\left(\hat{f}_U\left(u^1\right) \quad \cdots \quad \hat{f}_U\left(u^K\right)\right) = \iota_{M_z}{}^\intercal \widehat{\Lambda}_0 \cdot \operatorname{diag}\left(\mathbb{F}_{Z|D=0}(\mathcal{Z}_1), \cdots, \mathbb{F}_{Z|D=0}(\mathcal{Z}_{M_z})\right)$$

By summing over $\mathcal{U}$, we get

$$\hat{F}_{Y(1),Y(0)}(y_1, y_0) = \sum_{k=1}^K \hat{F}_{Y(1)|U}(y_1|u^k) \cdot \hat{F}_{Y(0)|U}(y_0|u^k) \cdot \hat{f}_U\left(u^k\right).$$

Corollary 1 shows that $\hat{F}_{Y(1),Y(0)}$ is a uniformly consistent estimator.

Corollary 1. *Assumptions 1-2, 4-5 hold. Then,*

$$\sup_{(y_1,y_0)\in\mathbb{R}^2} \left| \hat{F}_{Y(1),Y(0)}(y_1,y_0) - F_{Y(1),Y(0)}(y_1,y_0) \right| \xrightarrow{p} 0$$

as $n \to \infty$.

Proof. See Appendix. ☐

The fact that M_z and $|\mathcal{U}|$ are finite plays a crucial role in deriving the uniform consistency result. Likewise, relying on the finiteness and the result of Devroye (1983), we can prove L^1 consistency on the joint density estimator; the estimator for the treatment effect distribution, i.e., $\Pr\{Y_i(1) - Y_i(0) \leq \delta\}$, is uniformly consistent across $\delta \in \mathbb{R}$.

4 Extension

4.1 Nonrandom treatment

In the identification argument, Assumption 1 is used to connect the spectral decomposition result from the treated subsample and that from the untreated subsample, by establishing the conditional independence of $(Y_i(1), Y_i(0), X_i)$ and D_i given U_i. The random treatment condition frmo Assumption 1 can be relaxed as long as we get this conditional independence of the treatment given U_i.

Assumption 6. $(Y_i(1), Y_i(0), X_i) \perp\!\!\!\perp D_i \mid (Z_i, U_i)$.

Assumption 6 assumes that the potential outcomes and one of the proxy variables are conditionally independent of the treatment given the other proxy variable and the latent rank variable U_i. Under Assumption 6, the decomposition in (4) still holds, for each of the two subsamples:

$$\begin{aligned} f_{Y=y,X|D=d,Z}(x|z) &= \int_{\mathcal{U}} f_{Y(d),X|D=d,Z,U}(y,x|z,u) \cdot f_{U|D=d,Z}(u|z)du \\ &= \int_{\mathcal{U}} f_{Y(d),X|Z,U}(y,x|z,u) \cdot f_{U|D=d,Z}(u|z)du \quad \because \text{ Assumption 6} \\ &= \int_{\mathcal{U}} f_{Y(d)|U}(y|u) \cdot f_{X|U}(x|u) \cdot f_{U|D=d,Z}(u|z)du \quad \because \text{ Assumption 2} \end{aligned}$$

Thus, the identification result discussed in Section 2 holds.

Assumption 6 assumes that the latent rank variable U_i and the proxy variable Z_i contains sufficient information in the treatment assignment process. Recall that the key assumption of the econometric framework of this paper is that the one-dimensional latent variable U_i is rich enough to contain sufficient information regarding the dependence between the two potential outcomes. Assumption 6 builds onto this and assumes that the richness of U_i also captures the treatment endogeneity.

That being said, the additional assumption on the role of the latent rank U_i may not be as restrictive as it seems, in the sense that Assumption 2 is already an assumption about heterogeneity in the potential outcomes, which is believed to be the key determinant in the treatment assignment process in many empirical contexts. In the short panel example from Example 1, Assumption 2 was motivated with a nonseparable panel model with a common shock V_{it} that affects both treated potential outcome and untreated potential outcome. Suppose that at the beginning of the time period $t = 2$, individuals select into treatment by comparing their expected gain from being under treatment with their costs η_i:

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

The assignment model above assumes that the timing of selection is when the individuals are only aware of their 'latent rank' V_{i2} and thus the (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 6 is satisfied.

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 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 fall semester of 2016, until 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.

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. Note that the distribution of the additional random shocks is allowed to depend on the underlying health status, but not on each other. For example, one's susceptibility to the wellness program may depend on their underlying health status, but not on, for example, their disappointment from not being offered the wellness program.

Figure 1 contains the estimated joint distribution of the two potential outcomes. For visibility, the joint distribution is plotted for their quantiles in the marginal distribution of the outcome; Figure 1 plots the joint distribution of $F_Y(Y_i(1)) \in [0,1]$ and $F_Y(Y_i(0)) \in [0,1]$ where F_Y is the marginal distribution of Y_i. From the estimates, we see strong positive dependence between the two potential outcomes 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 if someone has a low underlying health status, their medical spending is likely to be large, regardless of participating in the wellness program.

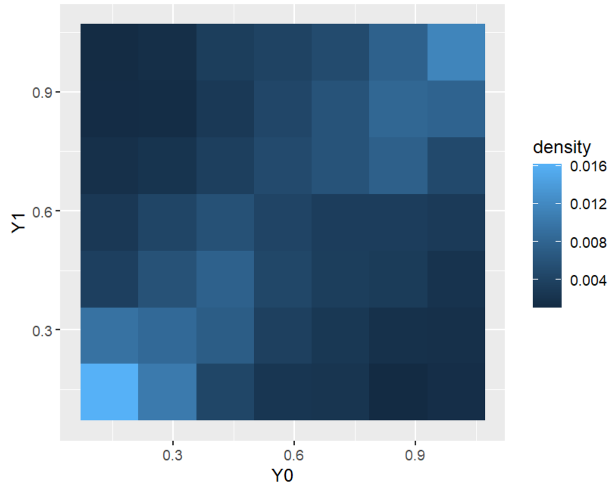

Figure 1: Joint density of $F_Y\big(Y_i(1)\big)$ and $F_Y\big(Y_i(0)\big)$, $K = 5$.

Figure 2 contains the marginal distribution of the $Y_i(1) - Y_i(0)$. As predicted by the ATE estimate not being significantly away from zero, the marginal distribution of the treatment effect seems to be symmetric around zero. In addition, we see high probability mass around zero, meaning that for a majority of the participants, the intent-to-treat effect from the wellness program eligibility was near zero.

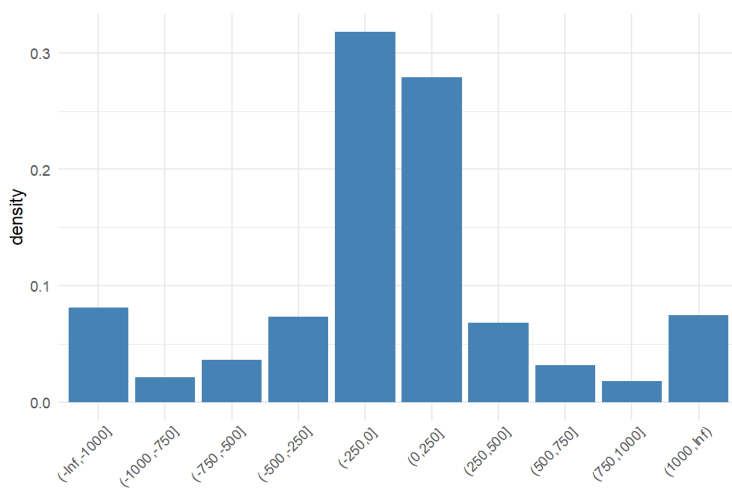

Figure 2: Marginal density of $Y_i(1) - Y_i(0)$, $K = 5$.

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 assumption in the identification is that there exists a latent variable that captures the dependence between the two potential outcomes. By assuming strict monotonicity for some functional of the conditional distribution of potential outcomes given the latent variable, we interpret the latent variable as 'latent rank' and strict monotonicty as 'latent rank invariance.' Provided two conditionally independent proxy variables, the conditional distribution of each potential outcome given the latent variable and thus the joint distribution of the two potential outcomes are identified. In implementation, we discretize the support of the latent rank variable and use a nonnegative matrix factorization algorithm. Lastly, we 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.

References

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

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.

Devroye, Luc, "The equivalence of weak, strong and complete convergence in L1 for kernel density estimates," *The Annals of Statistics*, 1983, pp. 896–904.

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.

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.

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.

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.

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 Proofs

A.1 Proof for Theorem 1

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 $\tilde{U}_i = h(U_i) := Mf_{Y(1),Y(0)|U}(\cdot,\cdot|U_i)$ 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

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

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

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

Since the conditional density of Z_i given $D_i = d$ is directly observed, the marginal density of $\tilde{U}_i$ is also identified.

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

A.2 Proof for Theorem 2

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

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

Then,

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

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

Construct a matrix $\tilde{P} \in \mathbb{R}^{K \times M}$ that satisfies the following: an element of $\tilde{P}$ is one or zero; product of any two rows of $\tilde{P}$ is zero; $\|\tilde{P}\|_F = M$ and $\tilde{P}\Gamma_0$ is invertible. When $M_x = K$, we can directly use P as $\tilde{P}$; when $M_x > K$, the existence of such $\tilde{P}$ is guaranteed from $rank(P) = M_x$. Given $\tilde{P}$,

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

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

Let $E = \widehat{\Gamma}_0\widehat{\Lambda}_0 - \Gamma_0\Lambda_0$. Then,

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

whenever $\left(\widehat{\Lambda}_0\widehat{\Lambda}_0^\intercal\right)^{-1}$ exists. In addition, when both $\left(\tilde{P}\widehat{\Gamma}_0\right)^{-1}$ and $\left(\widehat{\Lambda}_0\widehat{\Lambda}_0^\intercal\right)^{-1}$ exist,

$$\left\|\left(\widehat{\Lambda}_0\widehat{\Lambda}_0^\intercal\right)^{-1}\right\|_F \leq \left\|\widehat{\Gamma}_0^\intercal\tilde{P}^\intercal\right\|_F \left\|\left(\tilde{P}\widehat{\Gamma}_0\widehat{\Lambda}_0\widehat{\Lambda}_0^\intercal\widehat{\Gamma}_0^\intercal\tilde{P}^\intercal\right)^{-1}\right\|_F \left\|\tilde{P}\widehat{\Gamma}_0\right\|_F.$$

$\left\|\tilde{P}\widehat{\Gamma}_0\right\|_F$ is bounded by K^2 and $\left\|\left(\tilde{P}\widehat{\Gamma}_0\widehat{\Lambda}_0\widehat{\Lambda}_0^\intercal\widehat{\Gamma}_0^\intercal\tilde{P}^\intercal\right)^{-1}\right\|_F$ converges to $\left\|\left(\tilde{P}\Gamma_0\Lambda_0{\Lambda_0}^\intercal{\Gamma_0}^\intercal\tilde{P}^\intercal\right)^{-1}\right\|_F$. Both $\|\widehat{\Lambda}_0\|_F$ and $\left\|\left(\widehat{\Lambda}_0\widehat{\Lambda}_0^\intercal\right)^{-1}\right\|_F$ are bounded. $\left\|E\widehat{\Lambda}_0^\intercal\left(\widehat{\Lambda}_0\widehat{\Lambda}_0^\intercal\right)^{-1}\right\|_F = O_p\left(\frac{1}{\sqrt{n}}\right)$. By letting

$$A = \Lambda_0\widehat{\Lambda}_0^\intercal\left(\widehat{\Lambda}_0\widehat{\Lambda}_0^\intercal\right)^{-1},$$

we have that $\left\|\widehat{\Gamma}_0 - \Gamma_0 A\right\|_F = \left\|E\widehat{\Lambda}_0^\intercal\left(\widehat{\Lambda}_0\widehat{\Lambda}_0^\intercal\right)^{-1}\right\|_F = O_p\left(\frac{1}{\sqrt{n}}\right)$.

Step 3. *A* converges to a permutation matrix.

Firstly, find that

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

The columns of A sum to one.

WLOG choose the k-th column of Γ_0 and construct a $M_x \times M_y$ matrix $\tilde{\Gamma}_k$ where the m-th row and m'-th column element of $\tilde{\Gamma}_k$ is

$$
\Pr\left\{Y_i(0) \in \mathcal{Y}_{m'}, X_i \in \mathcal{X}_m | U_i = u^k\right\}.
$$

Note that $\tilde{\Gamma}_k = p_k {q_{0k}}^{\intercal}$, with p_k, q_{0k} as defined from Assumption 5. Likewise, construct $\widehat{\tilde{\Gamma}}_k$ from $\widehat{\Gamma}_0$. 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. Find that

$$
\begin{aligned}
\min_{p\in\mathbb{R}^{M_x}, q\in\mathbb{R}^{M_y}} \left\|\sum_{j=1}^{K}\tilde{\Gamma}_j a_{jk} - pq^{\intercal}\right\|_F &\le \left\|\sum_{j=1}^{K}\tilde{\Gamma}_j a_{jk} - \widehat{\tilde{\Gamma}}_k\right\|_F + \min_{p,q}\left\|\widehat{\tilde{\Gamma}}_k - pq^{\intercal}\right\|_F \\
&= \left\|\sum_{j=1}^{K}\tilde{\Gamma}_j a_{jk} - \widehat{\tilde{\Gamma}}_k\right\|_F \le \left\|\widehat{\Gamma}_0 - \Gamma_0 A\right\|_F.
\end{aligned}
$$

The equality holds from the construction of the estimator $\widehat{\Gamma}_0$; the estimated mixture component distribution satisfies the conditional independence of $Y_i(0)$ and X_i given U_i. The second inequality holds by expanding from the k-th column to the entire matrix.

Let e_k denote the K-dimensional elementary vector whose k-th element is one and the rest are zeros. Suppose $\min_{k'}\|a_{\cdot k} - e_{k'}\|_\infty \ge \varepsilon$. Then, there exist some $j \ne j'$ such that $|a_{jk} - 1| \le |a_{lk} - 1|$ for all l and $|a_{j'k}| \ge \frac{\varepsilon}{K-1}$ since each column has to sum to one. Find that

$$
\min_{p\in\mathbb{R}^{M_x}, q\in\mathbb{R}^{M_y}} \left\|\sum_{j=1}^{K}\tilde{\Gamma}_j a_{jk} - pq^{\intercal}\right\|_F = \min_{p\in\mathbb{R}^{M_x}, q\in\mathbb{R}^{M_y}} \left\|\sum_{j=1}^{K} a_{jk} p_j {q_{0j}}^{\intercal} - pq^{\intercal}\right\|_F = O_p\left(\frac{1}{\sqrt{n}}\right).
$$

WTS

$$\min_{p\in\mathbb{R}^{M_x},q\in\mathbb{R}^{M_y}} \left\| \sum_{j=1}^K a_{jk} p_j q_{0j}{}^\intercal - pq^\intercal \right\|_F$$

is bounded from below when $\min_{k'} \|a_{\cdot k} - e_{k'}\|_\infty \geq \varepsilon$, to create a contradiction. Consider a singular value decomposition of $\sum_{j=1}^K a_{jk} p_j q_{0j}{}^\intercal$ for the given $a_{\cdot k}$. There are at least two nonzero singular values: $|a_{j'k}| \geq \frac{\varepsilon}{K-1}$. Since the Frobenius norm is the l_2 norm on the vector of singular values and from the known results on the principal component analysis, the minimized objective above is the l_2 norm on the vector of singular values minus the largest singular value, which is positive. Let us expand this argument: with some arbitrary small $\eta > 0$, let

$$\rho(\varepsilon) = \min\left\{ \min_{p,q} \left\| \sum_{j=1}^K a_j p_j q_{0j}{}^\intercal - pq^\intercal \right\|_F : \min_k \|a - e_k\|_\infty \geq \varepsilon, \iota_K{}^\intercal a = 1, \sum_{j=1}^K a_j p_j q_{0j}{}^\intercal \geq -\eta \right\}.$$

Since the set is bounded and closed, the minimum is defined for some a that satisfies the constraints. Thus, $\rho(\varepsilon) > 0$ and $\rho(\varepsilon)$ is decreasing in ε. For each k, $a_{\cdot k}$ satisfies that $\sum_{j=1}^K a_{jk} p_j q_{0j}{}^\intercal \geq -\eta$ with probability going to one from the fact that $\widehat{\Gamma}_0 \geq 0$ and $\|\widehat{\Gamma}_0 - \Gamma_0 A\|_F$ converges to zero in probability.

Fix $\varepsilon > 0$. Then, $\rho(\varepsilon)$ is also fixed to be a positive constant and

$$\begin{aligned}
&\Pr\left\{\min_{k'} \|a_{\cdot k} - e_{k'}\|_\infty \geq \varepsilon\right\} \\
&= \Pr\left\{\min_{k'} \|a_{\cdot k} - e_{k'}\|_\infty \geq \varepsilon, \sum_{j=1}^K a_{jk} p_j q_{0j}{}^\intercal \geq -\eta\right\} \\
&\quad + \Pr\left\{\text{at least one element of } \sum_{j=1}^K a_{jk} p_j q_{0j}{}^\intercal \text{ is smaller than } -\eta\right\} \\
&\leq \Pr\left\{\min_{p\in\mathbb{R}^{M_x},q\in\mathbb{R}^{M_y}} \left\| \sum_{j=1}^K a_{jk} p_j q_{0j}{}^\intercal - pq^\intercal \right\|_F \geq \rho(\varepsilon)\right\} + o(1) = o(1).
\end{aligned}$$

Each column of A converges to an elementary vector.

It remains to show that the rotation 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. Let $\tilde{\varepsilon} \leq \frac{1}{2K}$. Assume to the contrary that there exist some k such that for every k', $\|a_{\cdot k'} - e_j\|_\infty \leq \tilde{\varepsilon}$ with some $j \neq k$.

Then, the k-th row of $A\widehat{\Lambda}_0$ lies in $[0,\tilde{\varepsilon}]^K$. Find that

$$\begin{aligned}\left\|A\widehat{\Lambda}_0-\Lambda_0\right\|_F &\leq \left\|\left(\tilde{P}\Gamma_0\right)^{-1}\tilde{P}\right\|_F\left\|\Gamma_0\left(\Lambda_0-A\widehat{\Lambda}_0\right)\right\|_F\\ &\leq \left\|\left(\tilde{P}\Gamma_0\right)^{-1}\tilde{P}\right\|_F\left(\left\|\Gamma_0\Lambda_0-\widehat{\Gamma}_0\widehat{\Lambda}_0\right\|_F+\left\|\widehat{\Gamma}_0\widehat{\Lambda}_0-\Gamma_0A\widehat{\Lambda}_0\right\|_F\right)=O_p\left(\frac{1}{\sqrt{n}}\right).\end{aligned}$$

Thus, the k-th row of $A\widehat{\Lambda}_0$ lying in $[0,\tilde{\varepsilon}]^K$ leads to a contradiction. Let

$$c=\min_k\max_m \Pr\left\{U_i=u^k|Z_i\in\mathcal{Z}_m\right\}$$

and let P_π denote a (right) permutation matrix induced by a permutation π on $\{1,\cdots,K\}$. c searches for the biggest element from each row of Λ_0 and then takes the minimum across rows. Since the row-specific search should give us a positive probability, $c>0$; otherwise Λ_0 does not have full rank. Then, for small enough $\varepsilon<c$,

$$\begin{aligned}&\Pr\left\{\min_\pi\|A-P_\pi\|_\infty\geq\varepsilon\right\}\\ &\leq\Pr\left\{\text{there exists some } k \text{ s.t. for every } k', \|a_{\cdot k'}-e_j\|_\infty\leq\varepsilon \text{ with some } j\neq k\right\}\\ &\qquad+\Pr\left\{\text{there exists some } k \text{ s.t.} \min_{k'}\|a_{\cdot k}-e_{k'}\|_\infty\geq\varepsilon\right\}\\ &\leq\Pr\left\{\left\|A\widehat{\Lambda}_0-\Lambda_0\right\|_F\geq|c-\varepsilon|\right\}+o(1)=o(1).\end{aligned}$$

A converges to a permutation matrix.

Step 4. For an arbitrary $\varepsilon>0$, $\Pr\left\{\left\|\widehat{\Lambda}_0-\Lambda_0\right\|_F>\varepsilon\right\}\to0$ as $n\to\infty$.

Find that

$$\begin{aligned}\left\|\Lambda_0-\widehat{\Lambda}_0\right\|_F &\leq\left\|\Lambda_0-\left(\tilde{P}\Gamma_0\right)^{-1}\tilde{P}\widehat{\Gamma}_0\widehat{\Lambda}_0\right\|_F+\left\|\left(\tilde{P}\Gamma_0\right)^{-1}\tilde{P}\widehat{\Gamma}_0\widehat{\Lambda}_0-\widehat{\Lambda}_0\right\|_F\\ &\leq\left\|\left(\tilde{P}\Gamma_0\right)^{-1}\tilde{P}\right\|_F\cdot\left\|\Gamma_0\Lambda_0-\widehat{\Gamma}_0\widehat{\Lambda}_0\right\|_F+\left\|\left(\tilde{P}\Gamma_0\right)^{-1}\tilde{P}\widehat{\Gamma}_0-I_K\right\|_F\cdot\left\|\widehat{\Lambda}_0\right\|_F\\ &\leq O_p\left(\frac{1}{\sqrt{n}}\right)+\left\|\widehat{\Lambda}_0\right\|_F\cdot\left(\left\|\left(\tilde{P}\Gamma_0\right)^{-1}\tilde{P}\left(\widehat{\Gamma}_0-\Gamma_0A\right)\right\|_F+\left\|\left(\tilde{P}\Gamma_0\right)^{-1}\tilde{P}\Gamma_0\left(A-I_K\right)\right\|_F\right)\\ &=O_p\left(\frac{1}{\sqrt{n}}\right)+\left\|\widehat{\Lambda}_0\right\|_F\cdot\left\|A-I_K\right\|_F.\end{aligned}$$

From the result of Step 3, we can relabel $\mathcal{U}$ so that A converges to I_K.

A.3 Proof for Corollary 1

From the Glivenko-Cantelli Theorem and Theorem 2, we have

$$\begin{aligned}
\sup_y \left| \hat{F}_{Y(1)|U}\left(y|u^k\right) - F_{Y(1)|U}\left(y|u^k\right)\right| &\xrightarrow{p} 0, \\
\sup_y \left| \hat{F}_{Y(0)|U}\left(y|u^k\right) - F_{Y(0)|U}\left(y|u^k\right)\right| &\xrightarrow{p} 0, \\
\hat{f}_U\left(u^k\right) - f_U\left(u^k\right) &\xrightarrow{p} 0
\end{aligned}$$

for each $k = 1, \cdots, K$, up to some permutation. Note that

$$\begin{aligned}
\left|\hat{a}\hat{b}\hat{c} - abc\right| &= \left|\hat{a}\hat{b}\hat{c} - a\hat{b}\hat{c} + a\hat{b}\hat{c} - ab\hat{c} + ab\hat{c} - abc\right| \\
&\leq |\hat{a} - a| \cdot \left|\hat{b}\hat{c}\right| + \left|\hat{b} - b\right| \cdot |a\hat{c}| + |\hat{c} - c| \cdot |ab| \\
&\leq |\hat{a} - a| \cdot \left(\left|\hat{b}\hat{c} - bc\right| + |bc|\right) + \left|\hat{b} - b\right| \cdot |a\hat{c}| + |\hat{c} - c| \cdot |ab| \\
&\leq |\hat{a} - a| \cdot \left(\left|\hat{b} - b\right| \cdot |\hat{c}| + |\hat{c} - c| \cdot |b| + |bc|\right) + \left|\hat{b} - b\right| \cdot |a\hat{c}| + |\hat{c} - c| \cdot |ab| .
\end{aligned}$$

Note that $\hat{a}$ and $\hat{b}$ only appear as differences $\hat{a} - a$ and $\hat{b} - b$ on the RHS. By taking a and b to be the estimates of the conditional distribution functions $\hat{F}_{Y(d)|U}$, the uniform consistency of $\hat{a} - a$ and $\hat{b} - b$ over $\mathbb{R}$ implies the uniform consistency of $\hat{a}\hat{b}\hat{c} - abc$ over $\mathbb{R}^2$.