Supplemental Appendix to "Distributional treatment effect with latent rank invariance"

Myungkou Shin*

September 13, 2026

1 Multidimensional U_i

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

Assumption 1. *Assume*

- ***a.*** *(multidimensional U_i) $\mathcal{U} \subset \mathbb{R}^p$.*
- ***b.*** *(bounded density) The conditional densities $f_{Y(1)|U}, f_{Y(0)|U}, f_{X|U}, f_{U|D=1,Z}$ and $f_{U|D=0,Z}$ and the marginal densities $f_U, f_{Z|D=1}$ and $f_{Z|D=0}$ are bounded.*

*School of Social Sciences, University of Surrey. email: m.shin@surrey.ac.uk

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

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

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

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

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

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

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

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

Importantly, Assumption 1 does not impose any restriction on the dependence structure within U_i, X_i and Z_i. Different dimensions of the latent variable U_i may be correlated. In addition, Assumption 1 does not impose any relationship between different dimensions of U_i and those of X_i and Z_i. We do not need each dimension of X_i and Z_i to correspond to a specific dimension of U_i.

2 Sieve maximum likelihood estimation

When U_i is continuous, I propose using a nonparametric estimation method such as sieve maximum likelihood. Under the conditional independence framework,

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

Given some sieves to approximate the conditional densities $f_{Y(1)|U}$, $f_{Y(0)|U}$, $f_{X|U}$, $f_{U|D=1,Z}$, $f_{U|D=0,Z}$ with finite-dimensional parameters $\theta = \big(\theta_1, \theta_0, \theta_X, \theta_{1Z}, \theta_{0Z}\big)$, the sieve ML estimator of the nonparametric finite mixture is:

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

In particular, I use 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 $\big(p^y+1, p^u+1\big)$ is as follows: with y, u 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 $\theta_1 = \{\theta_{jk,1}\}_{0\le j\le p^j, 0\le k\le p^u}$.[2] The tensor product construction and the properties of Bernstein polynomials make it remarkably straightforward to impose that the approximated functions are densities. The constraints that $f_{Y(1)|U,n}(y|u;\theta_1)$ is nonnegative and integrate to one correspond to the following linear constraints on the polynomial

[2]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.

coefficients:

$$
\begin{aligned}
\theta_{jk,1} \geq 0 \quad & \forall j,k & \textit{(nonnegative)} \\
\sum_{j=0}^{p^y} \frac{\theta_{j0,1}}{p^y+1} = 1 \quad & & \textit{(sum-to-one)} \\
\sum_{l=0}^{k}\sum_{j=0}^{p^y} \frac{1}{p^y+1}(-1)^{k-l}\binom{p^u}{k}\binom{k}{l}\theta_{jl,1} = 0 \quad & \forall k = 1,\cdots,p^u & \textit{(sum-to-one)}
\end{aligned}
$$

Moreover, when the latent rank interpretation from Assumption 5 is assumed with conditional expectation, 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

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

with some weights $\{w_j\}_{j=0}^{p_y}$ such that

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

Plug-in DTE estimation follows directly once we have $\hat{\theta}$. For example, the marginal treatment effect distribution estimator can be constructed as follows: for any δ,

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

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

3 Additional discussion on empirical illustration

3.1 Choice of K

To choose K to be used in Section 6 of the main manuscript, I applied the eigenvalue ratio estimator and the Kleibergen-Paap rank test to a 12×10 matrix

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

To discretize X_i and Z_i, I used equal partitions $(-\infty, F_X^{-1}(1/12)], \cdots, (F_X^{-1}(11/12), \infty)$ and $(-\infty, F_Z^{-1}(1/5)], \cdots, (F_Z^{-1}(4/5), \infty)$. Thus, $\mathbf{H}_X$ has 12 rows and 10 columns. Under Assumptions 1-3, $\mathbf{H}_X$ is at most rank K. Tables 1-2 contain the estimation/test results from the eigenvalue ratio estimator from Ahn and Horenstein [2013] and the Kleibergen-Paap rank test from Kleibergen and Paap [2006]. Both the eigenvalue ratio estimator and the Kleibergen-Paap rank test suggest $K \geq 3$.

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

Table 1: Eigenvalue ratios and growth ratios

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

Table 2: Kleibergen-Paap rank test statistics for H_0 : rank $= K$ and their p-values

As a second step, I estimated the nonparametric finite mixture for $K = 3, 4, 5, 6$ and applied the specification test from Appendix B of the main text. To discretize Y_i

and X_i, I set $M_Y = 4$ and $M_X = 6$, using equal partitions

$$(-\infty, F_Y^{-1}(1/4)], \cdots, (F_Y^{-1}(3/4), \infty)$$

for Y_i and

$$(-\infty, F_X^{-1}(1/6)], \cdots, (F_X^{-1}(5/6), \infty)$$

for X_i. For each value of $K = 3, \cdots, 6$, I also used equal partitions

$$(-\infty, F_Z^{-1}(1/K)], \cdots, (F_Z^{-1}((K-1)/K), \infty)$$

for Z_i. The test statistic tests the null hypothesis that

$$\Pr\{X_i \in \mathcal{X}^{m'} | D_i = 1, U_i = u^k\} = \Pr\{X_i \in \mathcal{X}^{m'} | D_i = 0, U_i = u^k\},$$

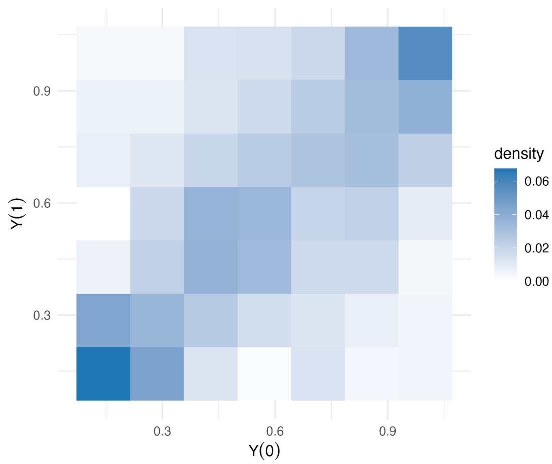
for every $m' = 1, \ldots, M_X - 1$ and $k = 1, \ldots, K$. Table 3 contains the test results. We find that the conditional distributions of X_i given U_i are well-balanced across the two treatment statuses, for all $K = 3, \ldots, 6$.

K	3	4	5	6
T_n	8.576	10.871	9.800	11.460
p-value	0.899	0.949	0.997	0.999

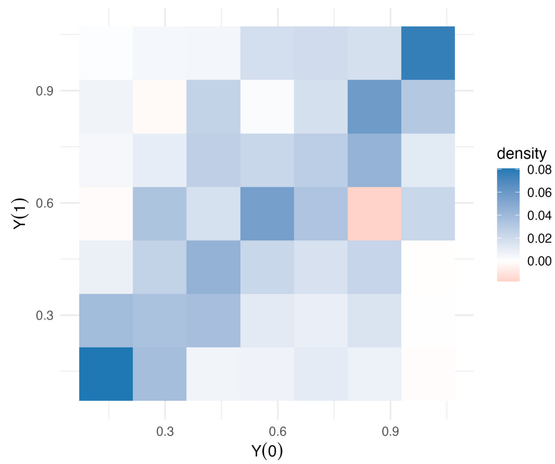
Table 3: Specification test statistics T_n and their p-values

3.2 Additional figures

In this subsection, I present estimates for the joint distribution of $Y_i(1)$ and $Y_i(0)$ for $K = 4, 5$ and estimates for the marginal distribution of $Y_i(1) - Y_i(0)$ for $K = 3, 4, 5, 6$, as sensitivity analysis. Firstly, Figure 1 plots the estimated joint distribution of the two potential outcomes for $K = 4, 5$. For visibility, I partitioned $Y_i(1)$ and $Y_i(0)$ with quantiles $F_Y^{-1}(1/7), \cdots, F_Y^{-1}(6/7)$ and plotted the joint distribution of



(a) $K = 4$



(b) $K = 5$

Figure 1: Joint density of $Y_i(1)$ and $Y_i(0)$, across $K = 4, 5$.

partitioned potential outcomes. Since the treated potential outcomes are plotted on the vertical axis, higher mass on the upper-left triangle means that the treatment reduces the medical spending. Overall, there is no distinctive pattern of positive treatment effect or negative treatment effect. One notable observation is that the joint density is higher where $F_Y(Y_i(1)) \approx F_Y(Y_i(0)) \approx 0$ and $F_Y(Y_i(1)) \approx F_Y(Y_i(0)) \approx 1$. This is intuitive since on the two ends of the underlying health status spectrum, the effectiveness of the workplace wellness program must be limited. Additionally, comparison between the joint density estimate when $K = 4$ and that when $K = 5$ shows that there may be slight overfitting when $K = 5$, leading to occasional negative density estimates.

Secondly, Figure 2 contains the estimated marginal distribution of the individual-level treatment effect $Y_i(1) - Y_i(0)$, across $K = 3, 4, 5, 6$. Notably, the point estimates are highly volatile with $K = 6$, decreasing on a significant subset of the support $[-\$1000, \$1000]$. Figure 2 is plotted by evaluating the DTE function $F_{Y(1)-Y(0)}(\delta)$ on a 101-point-grid on $[-\$1000, \$1000]$; out of 100 increments, the estimated distribution function decreases for 37 increments, when $K = 6$. Moreover, I also constructed a

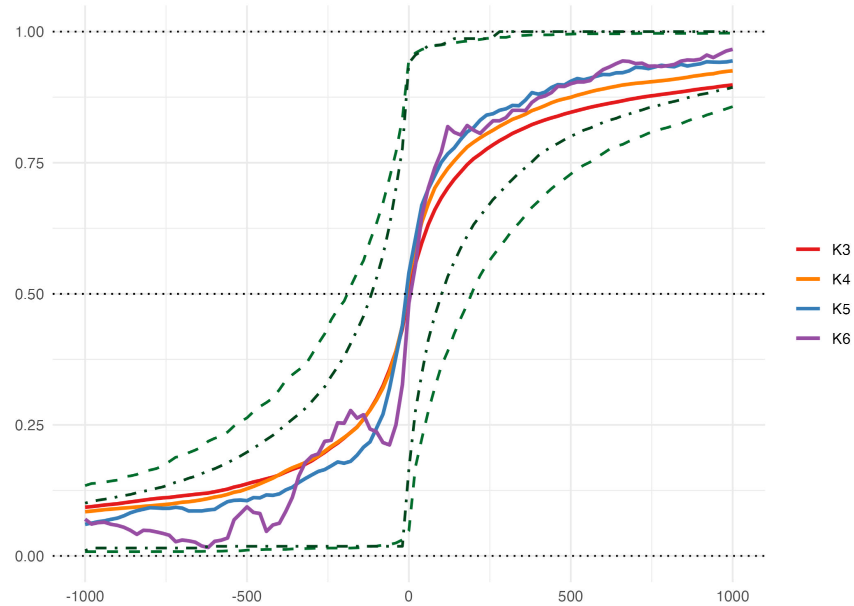


Figure 2: Marginal distribution of $Y_i(1) - Y_i(0)$, across $K = 3, \cdots, 6$.

measure of the monotonicity violation for comparison:

$$\sum_{d=1}^{100} \left| \widehat{F}_{Y(1)-Y(0)}(\delta^d) - \widehat{F}_{Y(1)-Y(0)}(\delta^{d-1}) \right| \mathbf{1}\left\{ \widehat{F}_{Y(1)-Y(0)}(\delta^d) - \widehat{F}_{Y(1)-Y(0)}(\delta^{d-1}) \right\}$$

with $\left(\delta^0, \delta^1, \cdots, \delta^{100}\right) = (-1000, -980, \cdots, 1000)$. The monotonicity violation measure is 0, 0, 0.029, 0.237 for $K = 3, 4, 5, 6$, respectively. This provides another anecdotal evidence that the nonlinear least-square estimator overfits the estimated probability $\hat{h}(\cdot|\cdot)$ when $K = 6$. Hence, I reported $K = 5$ result in the main text.

Also, the right tail of the distribution function seems to be more sensitive to the choice of K. In particular, the right tail suggests that the direction of the misspecification/discretization bias from using a smaller-than-true K is positive; as I increase K, the estimated right tail of the distribution approaches zero. This suggests that the negative impact of the treatment, i.e. an increase in medical spending from the treatment, may be even lower than what is suggested by $K = 5$ estimation results.

4 Proofs

4.1 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 Equation (9) of the main text. Fix some (m, d, m', k) and let

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

Then, for $j = 1, \ldots, K$, $\mathbf{E}\left[\frac{\partial}{\partial \tilde{\lambda}_{d'}(j,k')} \phi_A\big(W_i, W_{i'}; \tilde{\lambda}, p\big)\right]$ equals

$$
\begin{aligned}
&\Pr\{Y_i \in \mathcal{Y}^m, X_i \in \mathcal{X}^{m'} | D_i = d, Z_i \in \mathcal{Z}^j\} - \Pr\{Y_i \in \mathcal{Y}^m | D_i = d, Z = z^j\} \cdot \Pr\{X_i \in \mathcal{X}^{m'} | U_i = u^k\} \\
&\qquad - \Pr\{X_i \in \mathcal{X}^{m'} | D_i = d, Z = z^j\} \cdot \Pr\{Y_i(d) = y | U_i = u^k\}.
\end{aligned}
$$

if $k' = k$ and $d' = d$ and zero otherwise. $\mathbf{E}\Big[\frac{\partial}{\partial p_{D,U}(d',k')} \phi_A\big(W_i, W_{i'}; \tilde{\lambda}, p\big)\Big] = 0$ for every d', k'. Lastly, for $j = 1, \ldots, K$, $\mathbf{E}\left[\frac{\partial}{\partial p_{D,Z}(d',j)} \phi_A\big(W_i, W_{i'}; \tilde{\lambda}, p\big)\right]$ equals

$$
\begin{aligned}
&- \frac{\tilde{\lambda}_d(j,k)}{p_{D,Z}(d,j)} \cdot \Pr\{Y_i \in \mathcal{Y}^m, X_i \in \mathcal{X}^{m'} | D_i = d, Z_i \in \mathcal{Z}^j\} \\
&\quad + \frac{\tilde{\lambda}_d(j,k)}{p_{D,Z}(d,j)} \cdot \Pr\{Y_i \in \mathcal{Y}^m | D_i = d, Z_i \in \mathcal{Z}^j\} \cdot \Pr\{X_i \in \mathcal{X}^{m'} | U_i = u^k\} \\
&\quad + \frac{\tilde{\lambda}_d(j,k)}{p_{D,Z}(d,j)} \cdot \Pr\{X_i \in \mathcal{X}^{m'} | D_i = d, Z_i \in \mathcal{Z}^j\} \cdot \Pr\{Y_i(d) = y | U_i = u^k\}
\end{aligned}
$$

if $d' = d$ and zero otherwise.

Secondly, ϕ_B is the part of ϕ that corresponds to the Equation (10) of the main text. Fix some (d, m') and let

$$\begin{aligned}
&\phi_B(W_i, W_{i'}; \tilde{\lambda}, p) \\
&= \frac{\mathbf{1}\{D_i = d, X_i \in \mathcal{X}^{m'}\} + \mathbf{1}\{D_{i'} = d, X_{i'} \in \mathcal{X}^{m'}\}}{2} \\
&\quad - \sum_k p_{D,U}(d,k) \sum_j \frac{\tilde{\lambda}_d(j,k)}{p_{D,Z}(d,j)} \frac{\mathbf{1}\{D_i = d, X_i \in \mathcal{X}^{m'}, Z_i \in \mathcal{Z}^j\} + \mathbf{1}\{D_{i'} = d, X_{i'} \in \mathcal{X}^{m'}, Z_{i'} \in \mathcal{Z}^j\}}{2}.
\end{aligned}$$

Then, for $j = 1, \ldots, K$ and $k = 1, \ldots, K$,

$$\mathbf{E}\left[\frac{\partial}{\partial \tilde{\lambda}_{d'}(j,k)} \phi_B\big(W_i, W_{i'}; \tilde{\lambda}, p\big)\right] = -p_{D,U}(d,k) \cdot \Pr\{X_i \in \mathcal{X}^{m'} | D_i = d, Z_i \in \mathcal{Z}^j\}$$

if $d' = d$ and zero otherwise. Also, for $k = 1, \ldots, K$,

$$\mathbf{E}\left[\frac{\partial}{\partial p_{D,U}(d',k)} \phi_B\big(W_i, W_{i'}; \tilde{\lambda}, p\big)\right] = -\Pr\{X_i \in \mathcal{X}^{m'} | U_i = u^k\}$$

if $d' = d$ and zero otherwise. Lastly,

$$\mathbf{E}\left[\frac{\partial}{\partial p_{D,Z}(d,j)} \phi_B\big(W_i, W_{i'}; \tilde{\lambda}, p\big)\right] = \sum_{k=1}^{K} \frac{p_U(k) \tilde{\lambda}_d(j,k)}{p_{D,Z}(d,j)} \cdot \Pr\{X_i \in \mathcal{X}^{m'} | D_i = d, Z_i \in \mathcal{Z}^j\}$$

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

Thirdly, ϕ_C is the moment condition for $p_{D,Z}$ from Equation (11) of the main text. Fix some (d, j) and let

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

$\mathbf{E}\left[\frac{\partial}{\partial \tilde{\lambda}_{d'}(j,k)} \phi_C\big(W_i, W_{i'}; \tilde{\lambda}, p\big)\right]$ and $\mathbf{E}\left[\frac{\partial}{\partial p_{D,U}(d',k)} \phi_C\big(W_i, W_{i'}; \tilde{\lambda}, p\big)\right]$ are zero for every (d', j, k). Also,

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

if $d' = d$ and $j' = j$ and zero otherwise.

With some order of stacking ϕ_A, ϕ_B, ϕ_C and flattening $\tilde{\Lambda}_0, \tilde{\Lambda}_1$,[3] the Jacobian matrix becomes

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

It suffices to show that the submatrix

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

is full rank. Assume to the contrary that the rows of the submatrix from (2) are linearly dependent: with linear coefficients $\alpha = \left(\alpha_{A,1}, \cdots, \alpha_{A,2K^2}, \alpha_{B,1}, \cdots, \alpha_{B,2K}\right)^\intercal$,

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

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

[3] The order of ϕ_A, ϕ_B and ϕ_C across different values of (m, m', d, j, k) in ϕ is as follows. Firstly, stack ϕ_A across every value of (m, m') for $(d = 0, k = 1)$, looping m' firstly and then m, and then for $(d = 1, k = 1)$. Then, repeat this for $k = 2, \cdots, K$. These will be the first $2M_Y M_X K$ 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.

For the order of $\tilde{\lambda}_d(j,k)$ in vectorized $\tilde{\lambda}$, collect $\tilde{\lambda}_d(j,k)$ across j for $(d = 0, k = 1)$ and then for $(d = 1, k = 1)$. Then, repeat this for $k = 2, \cdots, K$. This will be the $2K^2$-dimensional vector $\tilde{\lambda}$. For p, collect $p_{D,U}(0,k)$ across k, $p_{D,U}(1,k)$ across k, $p_{D,Z}(0,j)$ across j, and then $p_{D,Z}(1,j)$ across j.

matrices, each of which is a $K \times M$ matrix. For example, the first block matrix is

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

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

$$\alpha_B{}^{\intercal}\mathbf{E}\left[\frac{\partial}{\partial p_{D,U}}\phi_B\left(W_i, W_{i'}; \tilde{\lambda}, p\right)\right] = \alpha_B{}^{\intercal}\begin{pmatrix} -\Gamma_X{}^{\intercal} & \mathbf{O}_{K\times M_X} \\ \mathbf{O}_{K\times M_X} & -\Gamma_X{}^{\intercal} \end{pmatrix} = \mathbf{0}.$$

$\mathbf{E}\left[\frac{\partial}{\partial p_{D,U}}\phi_B\left(W_i, W_{i'}; \tilde{\lambda}, p\right)\right]$ is also a diagonal block matrix, where each of the two blocks is a $K \times M_X$ matrix $-\Gamma_X{}^{\intercal}$. From Assumption 3.b, α_B must be a zero vector. The Jacobian matrix has full rank. □

4.2 Proof for Lemma 2

From iid-ness of observations, we have $\|\mathbb{H}_0 - \mathbf{H}_0\|_F = O_p\left(\frac{1}{\sqrt{n}}\right)$ and $\|\mathbb{H}_1 - \mathbf{H}_1\|_F = O_p\left(\frac{1}{\sqrt{n}}\right)$. 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 $\left\|P\Gamma_1\Lambda_1 - P\widehat{\Gamma}_1\widehat{\Lambda}_1\right\|_F = \left\|P\Gamma_0\Lambda_1 - P\widehat{\Gamma}_0\widehat{\Lambda}_1\right\|_F = O_p\left(\frac{1}{\sqrt{n}}\right)$. □

4.3 Proof for Lemma 3

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

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

from Lemma 2. 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. When ${\Gamma_0}^{\intercal}\widehat{\Gamma}_0\widehat{\Lambda}_0$ is invertible,

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

with A as defined in Lemma 3. 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 $\widehat{\Gamma}_0$. □

4.4 Proof for Lemma 4

The proof for Lemma 4 consists of two steps. For notational convenience, 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 1. 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:

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

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 $M_Y M_X \times (K-1)$ matrix. Let

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

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

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

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

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

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

$$\left| \Pr\{Y_i(0) = y, X_i \in \mathcal{X}^{m'} | U_i = u^1\} - \sum_{j=2}^{K} \Pr\{Y_i(0) = y, X_i \in \mathcal{X}^{m'} | U_i = u^j\} \cdot \frac{a_{jk}}{\sum_{j=2}^{K} a_{jk}} \right|$$

is greater than or equal to $\frac{\delta^*}{2\sqrt{M_Y M_X}}$. Thus, $\sum_{j=1}^{K} \Pr\{Y_i(0) = y, X_i \in \mathcal{X}^{m'} | U_i = u^j\} a_{jk}$ lies outside of

$$\Pr\{Y_i(0) = y, X_i \in \mathcal{X}^{m'} | U_i = u^1\} + \left[-\frac{|1 - a_{1k}|\delta^*}{2\sqrt{M_Y M_X}}, \frac{|1 - a_{1k}|\delta^*}{2\sqrt{M_Y M_X}} \right]$$

and

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

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 $\Pr\left\{\|a_{\cdot k}\|_\infty \geq \frac{4\sqrt{M_Y M_X}}{\delta^*} + 1\right\} = o(1)$ 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. 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

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

$\tilde{\Gamma}_{0,k}$ 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

$$p_k = \left(\Pr\left\{X_i \in \mathcal{X}^1 | U_i = u^k\right\} \quad \cdots \quad \Pr\left\{X_i \in \mathcal{X}^{M_X} | U_i = u^k\right\}\right)^{\intercal},$$
$$q_{dk} = \left(\Pr\left\{Y_i(d) \in \mathcal{Y}^1 | U_i = u^k\right\} \quad \cdots \quad \Pr\left\{Y_i(d) \in \mathcal{Y}^{M_Y} | U_i = u^k\right\}\right)^{\intercal},$$

for $k = 1, \ldots, K$. 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 $\widehat{\tilde{\Gamma}}_{0,1}$ 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 $p_k(m') = \Pr\{X_i \in \mathcal{X}^{m'} | U_i = u^k\}$, $q_{dk}(m) = \Pr\{Y_i(d) \in \mathcal{Y}^m | U_i = u^k\}$ for $k = 1, \cdots, K$ and let

$$w(m) = \left(a_{11} q_{01}(m) \quad \cdots \quad a_{K1} q_{0K}(m)\right)^{\intercal}.$$

Then,

$$\sum_{k=1}^{K} \tilde{\Gamma}_{0,k} a_{k1} = \sum_{k=1}^{K} a_{k1} p_k q_{0k}^{\intercal} = \Gamma_X \left(w(1) \quad \cdots \quad w(M_Y)\right).$$

From Assumption 3.c,

$$c^* := \min_{k \neq k'} \left\{ \max_{1 \leq m \leq M_Y} \left(q_{0k}(m) - q_{0k'}(m) \right) \right\} > 0.$$

WLOG let m^1 and m^2 satisfy that

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

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

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

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

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

Then,

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

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 $\begin{pmatrix} w(m^1) & w(m^2) \end{pmatrix}$ 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.

The smallest eigenvalue of $\begin{pmatrix} w(m^1) & w(m^2) \end{pmatrix}$ equals $\min_{p\in\mathbb{R}^{M_X},q\in\mathbb{R}^2} \left\| \begin{pmatrix} w(m^1) & w(m^2) \end{pmatrix} - pq^\intercal \right\|_F$. Thus, the samllest eigenvalue is bounded from below by

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

We have shown that $\det \begin{pmatrix} w_j(m^1) & w_j(m^2) \\ w_{j'}(m^1) & w_{j'}(m^2) \end{pmatrix} \geq \frac{\varepsilon c^{*2}}{2\sqrt{n}K^2}$. With probability going to one, $w(m^1)$ and $w(m^2)$ is bounded by $\frac{4\sqrt{M_Y M_X}}{\delta^*} + 1$ and therefore

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

Thus, with probability going to one, the smallest eigenvalue of $\begin{pmatrix} w(m^1) & w(m^2) \end{pmatrix}$ is bounded from below by $\frac{1}{\sqrt{n}} \cdot \frac{\varepsilon c^{*2}}{2K^2} \cdot \left(\frac{4\sqrt{2M_Y M_X}}{\delta^*} + \sqrt{2}\right)^{-1}$. Consequently, with some constant $C^* > 0$ which does not depend on ε,

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

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

Step 2. 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

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

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 $\widehat{\Lambda}_0$,

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

□

References

Seung C Ahn and Alex R Horenstein. Eigenvalue ratio test for the number of factors. *Econometrica*, 81(3):1203–1227, 2013.

Orazio Attanasio, Sarah Cattan, Emla Fitzsimons, Costas Meghir, and Marta Rubio-Codina. Estimating the production function for human capital: results from a randomized controlled trial in colombia. *American Economic Review*, 110(1):48–85, 2020.

Flavio Cunha, James J Heckman, and Susanne M Schennach. Estimating the technology of cognitive and noncognitive skill formation. *Econometrica*, 78(3):883–931, 2010.

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

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

Frank Kleibergen and Richard Paap. Generalized reduced rank tests using the singular value decomposition. *Journal of econometrics*, 133(1):97–126, 2006.