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

Myungkou Shin*

December 18, 2025

1 Sieve maximum likelihood estimation

In this section, I propose a nonparametric estimation method to estimate the DTE parameters when U_i is continuous, using sieve maximum likelihood. Recall the integral decomposition:

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

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 is:

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

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

In particular, I propose tensor product spaces of Bernstein polynomials as sieves $\{\Theta_n\}_{n=1}^{\infty}$. For example, the conditional density $f_{Y(1)|U}$ approximated to a tensor product space with a given dimension of $\left(p^y+1, p^u+1\right)$ is as follows: with y, 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 $\theta_1 = \{\theta_{jk,1}\}_{0\le j\le p^j, 0\le k\le p^u}$.[1] 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 coefficients:

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

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

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

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

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

[1] 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.

fixing u:

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

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

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

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

Also, find that

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

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

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

Thus, the sum-to-one constraints are $\sum_{j=0}^{p^y}\frac{\theta_{j0,1}}{p^y+1} = 1$, $\sum_{l=0}^{k}\sum_{j=0}^{p^y}\frac{1}{p^y+1}(-1)^{k-l}\binom{p^u}{k}\binom{k}{l}\theta_{jl,1} = 0$ $\forall k = 1, \cdots, p^u$.

Lastly, for the monotonicity constraint, find that

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

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

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

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

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

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

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

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:

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

2 Additional simulation results

In this section, I present additional simulation results. Firstly, I expand Tables 1-3 of the main text to include more values of δ. Comparing Table 1 and 2, the use of nonnegative matrix factorization (NMF) compared to eigenvalue decomposition (EVD) improves the estimation performance in finite sample on the intensive margin, especially when the proxy variable is less informative: $\sigma_{\min}(\Lambda) = 0.337$. Also, Table 3 shows that the asymptotic standard error attaints the correct coverage probability when n is large, except when $F_{Y(1)-Y(0)}(\delta)$ is near zero or one.

Secondly, Table 4 contains proportions of the simulated samples where the NMF was successful and the same for EVD. For either estimation procedure, the estimation procedure was deemed 'unsuccessful' and halted, if any of the nuisance parameter matrices showed condition number bigger than 10^{10}, i.e. $\sigma_{\min}(A)/\sigma_{\max}(A) \leq 10^{-10}$. Simulation shows that the NMF estimation strategy almost always avoids singular nuisance parameters while the EVD estimation strategy fails for 15.4-47.2% of the simulated samples; the gain on the extensive margin from the additional regularization is huge. In addition, Table 4 contains average computation time for each estimation procedure, conditioning on estimation being successful.[2] The cost of the additional regularization is that the computation time increases by a factor of 1.59-5.09.

Thirdly, I provide visual illustration of estimation performance across $\sigma_{\min}(\Lambda)$ and n. Figure 1 contains the true marginal distribution of treatment effect and the conditional distributions of treatment effect given U_i for reference. Figure 2 plots the true marginal distribution of treatment and the mean of the DTE estimates based on NMF for $\left(\sigma_{\min}(\Lambda), n\right) = (0.377, 750)$ and $(0.806, 2000)$. In addition to the two plots, individual DTE estimates from the simulated samples are drawn as thin lines.

Then, Figure 3 zooms in on for three subsets of the support of δ, highlighting the bias. In addition to the mean of the main GMM estimates, Figure 3 also plots the means of one-sided GMM estimates. Find that the same DTE parameter $F_{Y(1)-Y(0)}(\delta)$ can be estimated using two different moments $\theta = \Pr\{Y_i(1) - Y_i(0) \leq \delta\}$ and $\theta = 1 - \Pr\{Y_i(1) - Y_i(0) > \delta\}$.

[2]The first-step NMF was implemented with 50 randomly generated initial values.

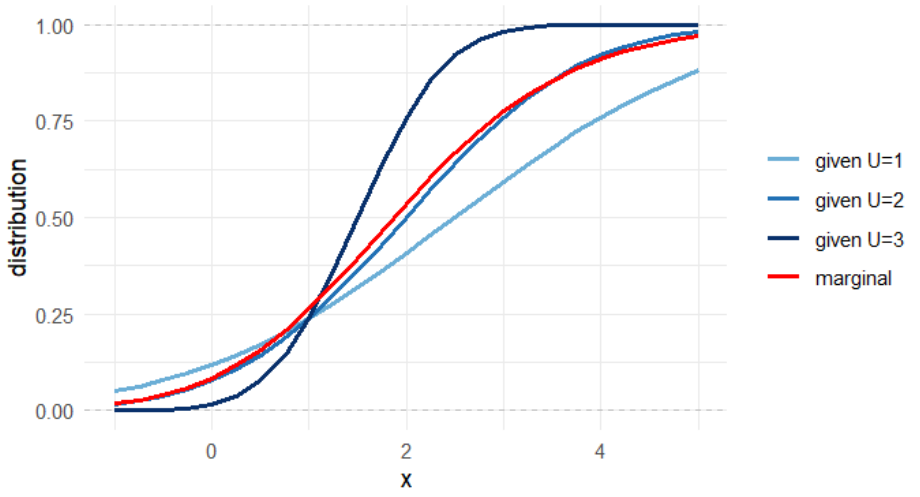


Figure 1: True distribution of treatment effect.

Though these two moments are identical when we observe both $Y_i(1)$ and $Y_i(0)$, they may lead to different feasible moment conditions due to the orthogonalization step. Let $\widehat{F}^1_{Y(1)-Y(0)}(\delta)$ denote the DTE estimator based on $\Pr\{Y_i(1) - Y_i(0) \leq \delta\}$ and $\widehat{F}^2_{Y(1)-Y(0)}(\delta)$ based on $1 - \Pr\{Y_i(1) - Y_i(0) > \delta\}$. The estimation results in the main text used the averaging estimator

$$\widehat{F}_{Y(1)-Y(0)}(\delta) = \frac{1}{2}\left(\widehat{F}^1_{Y(1)-Y(0)}(\delta) + \widehat{F}^2_{Y(1)-Y(0)}(\delta)\right)$$

The one-sided estimator $\widehat{F}^1_{Y(1)-Y(0)}(\delta)$ is denoted with green "m1" and the other one-sided estimator $\widehat{F}^2_{Y(1)-Y(0)}(\delta)$ is denoted with purple "m2" in Figure 3.

In general, the first one-sided estimator overestimates the DTE parameter while the second one-sided estimator underestimates. This contrast is most stark when Z_i is less informative and n is small. However, the averaging is mostly successful in controlling for the bias, suggesting that the averaging estimator be the preferred choice. The only exception is when $\delta \in [4, 5]$, for the DGP with $\left(\sigma_{\min}(\Lambda), n\right) = (0.377, 750)$. This is possibly due to the fact that the true conditional distributions of treatment effect are more heterogeneoues when δ is larger, as shown in Figure 1.

	true value	bias				rMSE			
$\widehat{F}_{Y(1)-Y(0)}(-1)$	0.018	-0.003	-0.001	-0.003	-0.001	0.008	0.005	0.006	0.003
$\widehat{F}_{Y(1)-Y(0)}(-0.5)$	0.041	-0.002	0.000	-0.003	-0.001	0.011	0.006	0.008	0.005
$\widehat{F}_{Y(1)-Y(0)}(0)$	0.084	0.000	0.000	-0.002	-0.001	0.014	0.009	0.011	0.007
$\widehat{F}_{Y(1)-Y(0)}(0.5)$	0.158	0.001	0.001	-0.001	-0.001	0.019	0.012	0.015	0.010
$\widehat{F}_{Y(1)-Y(0)}(1)$	0.264	0.001	0.001	-0.001	-0.001	0.023	0.015	0.019	0.012
$\widehat{F}_{Y(1)-Y(0)}(1.5)$	0.395	0.001	0.000	0.000	-0.001	0.025	0.016	0.021	0.013
$\widehat{F}_{Y(1)-Y(0)}(2)$	0.536	0.001	0.000	0.000	-0.001	0.025	0.016	0.022	0.014
$\widehat{F}_{Y(1)-Y(0)}(2.5)$	0.667	0.001	-0.001	0.001	0.000	0.023	0.015	0.020	0.013
$\widehat{F}_{Y(1)-Y(0)}(3)$	0.775	0.002	0.000	0.002	0.000	0.020	0.012	0.018	0.011
$\widehat{F}_{Y(1)-Y(0)}(3.5)$	0.855	0.004	0.001	0.003	0.001	0.017	0.010	0.015	0.009
$\widehat{F}_{Y(1)-Y(0)}(4)$	0.911	0.005	0.002	0.003	0.001	0.014	0.008	0.012	0.007
$\widehat{F}_{Y(1)-Y(0)}(4.5)$	0.947	0.006	0.002	0.004	0.001	0.012	0.007	0.010	0.006
$\widehat{F}_{Y(1)-Y(0)}(5)$	0.970	0.006	0.002	0.004	0.001	0.010	0.005	0.008	0.004
$\sigma_{\min}(\Lambda)$		0.337	0.337	0.806	0.806	0.337	0.337	0.806	0.806
n		750	2000	750	2000	750	2000	750	2000

Table 1: Bias and rMSE of DTE estimator $\widehat{F}_{Y(1)-Y(0)}(\delta)$ based on NMF.
Note: Bias and rMSE are computed among samples where the DTE estimation was successful.

	true value	bias				rMSE			
$\widehat{F}_{Y(1)-Y(0)}(-1)$	0.018	0.011	0.007	0.001	0.001	0.027	0.022	0.015	0.009
$\widehat{F}_{Y(1)-Y(0)}(-0.5)$	0.041	0.013	0.008	0.001	0.001	0.032	0.027	0.019	0.010
$\widehat{F}_{Y(1)-Y(0)}(0)$	0.084	0.014	0.008	0.002	0.001	0.034	0.029	0.022	0.012
$\widehat{F}_{Y(1)-Y(0)}(0.5)$	0.158	0.012	0.007	0.002	0.000	0.032	0.027	0.024	0.013
$\widehat{F}_{Y(1)-Y(0)}(1)$	0.264	0.006	0.004	0.002	0.000	0.030	0.021	0.024	0.014
$\widehat{F}_{Y(1)-Y(0)}(1.5)$	0.395	0.000	-0.001	0.000	0.000	0.031	0.022	0.024	0.015
$\widehat{F}_{Y(1)-Y(0)}(2)$	0.536	-0.006	-0.005	-0.001	0.000	0.037	0.029	0.025	0.015
$\widehat{F}_{Y(1)-Y(0)}(2.5)$	0.667	-0.009	-0.007	-0.001	-0.001	0.041	0.033	0.026	0.014
$\widehat{F}_{Y(1)-Y(0)}(3)$	0.775	-0.009	-0.007	-0.001	-0.001	0.040	0.032	0.025	0.012
$\widehat{F}_{Y(1)-Y(0)}(3.5)$	0.855	-0.008	-0.006	-0.001	-0.001	0.034	0.026	0.022	0.011
$\widehat{F}_{Y(1)-Y(0)}(4)$	0.911	-0.006	-0.004	0.000	-0.001	0.025	0.019	0.018	0.009
$\widehat{F}_{Y(1)-Y(0)}(4.5)$	0.947	-0.005	-0.003	0.001	0.000	0.017	0.012	0.014	0.007
$\widehat{F}_{Y(1)-Y(0)}(5)$	0.970	-0.005	-0.003	0.001	0.000	0.013	0.009	0.011	0.006
$\sigma_{\min}(\Lambda)$		0.337	0.337	0.806	0.806	0.337	0.337	0.806	0.806
n		750	2000	750	2000	750	2000	750	2000

Table 2: Bias and rMSE of DTE estimator $\widehat{F}_{Y(1)-Y(0)}(\delta)$ based on EVD.
Note: Bias and rMSE are computed among samples where the DTE estimation was successful.

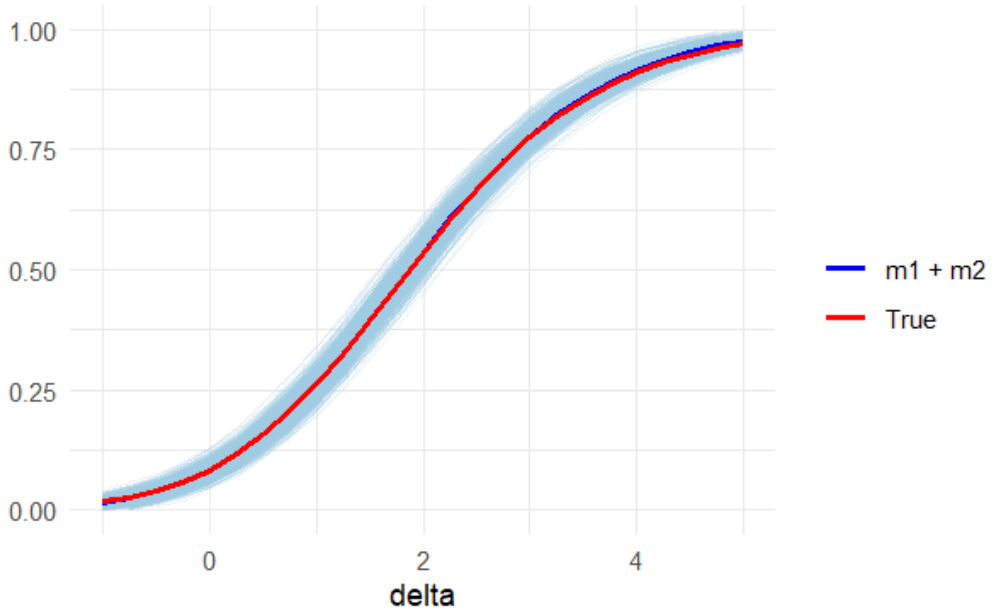
	true value	coverage probability			
$\widehat{F}_{Y(1)-Y(0)}(-1)$	0.018	0.938	0.940	0.864	0.918
$\widehat{F}_{Y(1)-Y(0)}(-0.5)$	0.041	0.959	0.945	0.924	0.933
$\widehat{F}_{Y(1)-Y(0)}(0)$	0.084	0.971	0.951	0.952	0.935
$\widehat{F}_{Y(1)-Y(0)}(0.5)$	0.158	0.971	0.954	0.957	0.945
$\widehat{F}_{Y(1)-Y(0)}(1)$	0.264	0.975	0.959	0.958	0.952
$\widehat{F}_{Y(1)-Y(0)}(1.5)$	0.395	0.971	0.957	0.965	0.951
$\widehat{F}_{Y(1)-Y(0)}(2)$	0.536	0.970	0.960	0.957	0.951
$\widehat{F}_{Y(1)-Y(0)}(2.5)$	0.667	0.968	0.967	0.960	0.952
$\widehat{F}_{Y(1)-Y(0)}(3)$	0.775	0.962	0.959	0.943	0.951
$\widehat{F}_{Y(1)-Y(0)}(3.5)$	0.855	0.953	0.956	0.938	0.953
$\widehat{F}_{Y(1)-Y(0)}(4)$	0.911	0.940	0.954	0.934	0.948
$\widehat{F}_{Y(1)-Y(0)}(4.5)$	0.947	0.923	0.944	0.904	0.938
$\widehat{F}_{Y(1)-Y(0)}(5)$	0.970	0.866	0.926	0.857	0.922
$\sigma_{\min}(\Lambda)$		0.337	0.337	0.806	0.806
n		750	2000	750	2000

Table 3: Coverage of 95% confidence interval based on NMF.
Note: Coverage probability is computed among samples where the DTE estimation was successful.

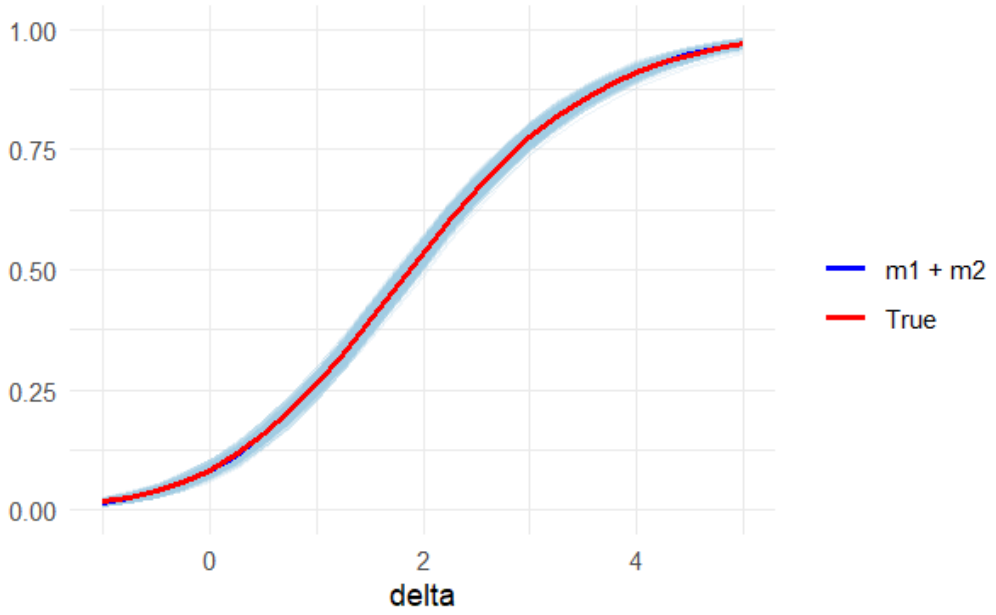
	success rate				computation time (sec)			
NMF	0.999	1.000	1.000	1.000	98.01	163.28	66.32	117.40
EVD	0.528	0.666	0.790	0.846	19.27	80.57	19.57	73.77
$\sigma_{\min}(\Lambda)$	0.337	0.337	0.806	0.806	0.337	0.337	0.806	0.806
n	750	2000	750	2000	750	2000	750	2000

Table 4: Success rate and computation time for DTE estimation based on NMF and EVD.

Note: Success rate is the proportion of the simulated samples where the estimation procedure was completed. Reasons for non-completion were singular first-step estimates $\widehat{\Lambda}_0, \widehat{\Lambda}_1$ and singular Jacobian matrix. Additionally, EVD estimation was halted whenever the eigenvalue decomposition led to complex eigenvectors.

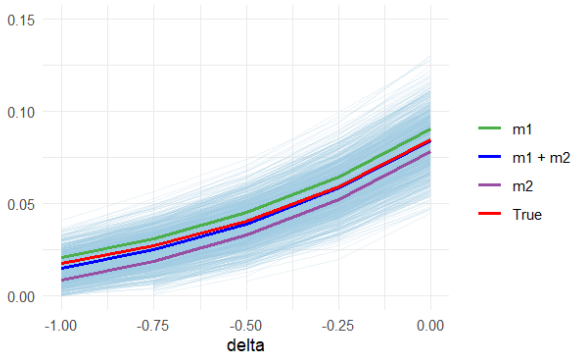


(a) $\sigma_{\min}(\Lambda) = 0.377$, $n = 750$.

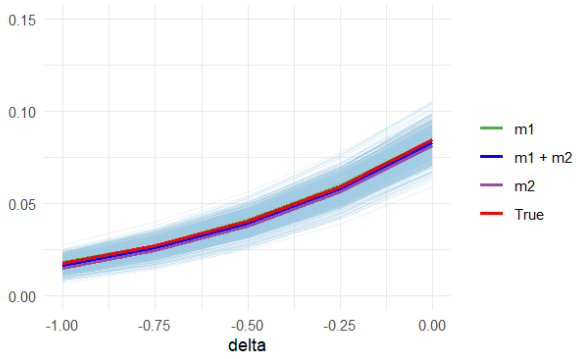


(b) $\sigma_{\min}(\Lambda) = 0.806$, $n = 2000$.

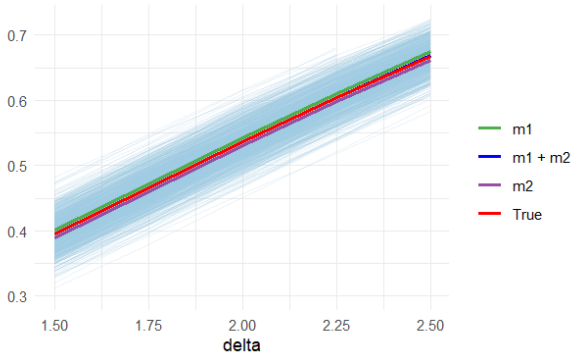
Figure 2: Distribution and mean of DTE estimates, compared to the true distribution.



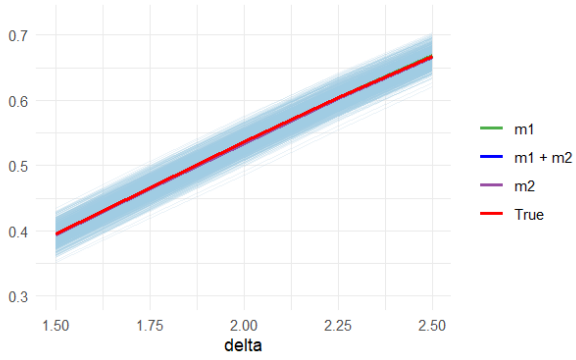
(a) $\sigma_{\min}(\Lambda) = 0.377$, $n = 750$, $\delta \in [-1, 0]$



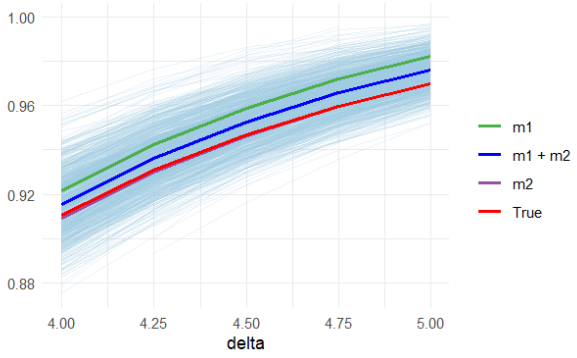
(b) $\sigma_{\min}(\Lambda) = 0.806$, $n = 2000$, $\delta \in [-1, 0]$



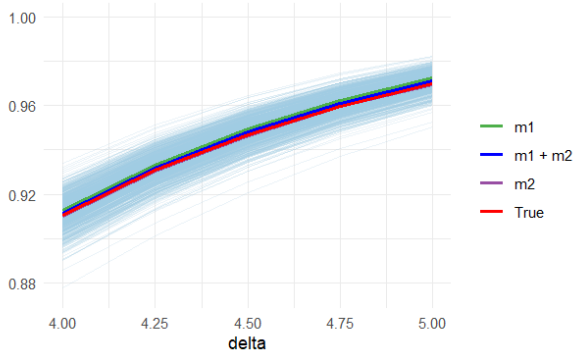
(c) $\sigma_{\min}(\Lambda) = 0.377$, $n = 750$, $\delta \in [1.5, 2.5]$



(d) $\sigma_{\min}(\Lambda) = 0.806$, $n = 2000$, $\delta \in [1.5, 2.5]$



(e) $\sigma_{\min}(\Lambda) = 0.377$, $n = 750$, $\delta \in [4, 5]$



(f) $\sigma_{\min}(\Lambda) = 0.806$, $n = 2000$, $\delta \in [4, 5]$

Figure 3: Comparison across three DTE estimators.

3 Additional discussion on empirical illustration

3.1 Choice of K

To choose K to be used in Section 5 of the main manuscript, I applied the eigenvalue ratio estimator and the Kleibergen-Paap rank test to a 12×10 matrix

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

To discretize X_i, I used an equal partition $(-\infty, F_X^{-1}(1/12)], \cdots, (F_X^{-1}(11/12), \infty)$ and for Z_i, I used an equal partition $(-\infty, F_Z^{-1}(1/5)], \cdots, (F_Z^{-1}(4/5), \infty)$. Thus, $\mathbf{H}_X$ has 12 rows and 10 columns. Under Assumptions 1-3, $\mathbf{H}_X$ is at most rank K. Tables 5-6 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 5: 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 6: Kleibergen-Paap rank test statistics for $H_0 : \text{rank} = K$ and their p-values

As a second step, I solved the NMF problem in the main text for $K = 3, 4, 5, 6$ and applied the falsification tests. In discretization, 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. Thus, the two matrices $\mathbf{H}_0$ and $\mathbf{H}_1$ used in the first-step NMF have 24 rows and K columns. Given each NMF, I tested the two testable implications:

$$\sum_{k=1}^{K} \sum_{m=1}^{M_X} \left(\Pr\left\{ X_i \in \mathcal{X}^m | D_i = 1, U_i = u^k \right\} - \Pr\left\{ X_i \in \mathcal{X}^m | D_i = 0, U_i = u^k \right\} \right)^2 = 0 \quad (2)$$

$$\sum_{k=1}^{K} \left(\Pr\left\{ U_i = u^k | D_i = 1 \right\} - \Pr\left\{ U_i = u^k | D_i = 0 \right\} \right)^2 = 0 \quad (3)$$

where $\mathcal{X}^m = \left[F_X^{-1}((m-1)/M_X), F_X^{-1}(m/M_X) \right]$ for $m = 1, \cdots, M_X$. Both (2) and (3) hold true when the treatment D_i is randomly assigned, independently of (X_i, U_i). Let T_n^1 denote the test statistic for the testable implication (2) and T_n^2 for the testable implication (3), as developed in the appendix Section B of the main text. Table 7 contains the test results. Overall, we reject neither (2) nor (3) at 0.1 singificance level, for $K = 3, 4, 5$. In particular, the two distributional equivalences seem to hold well when $K = 5$. In the case of $K = 6$, I suspect that the large test statistics are due to overfitting in the first-step NMF. This concern is addressed again in the next subsection.

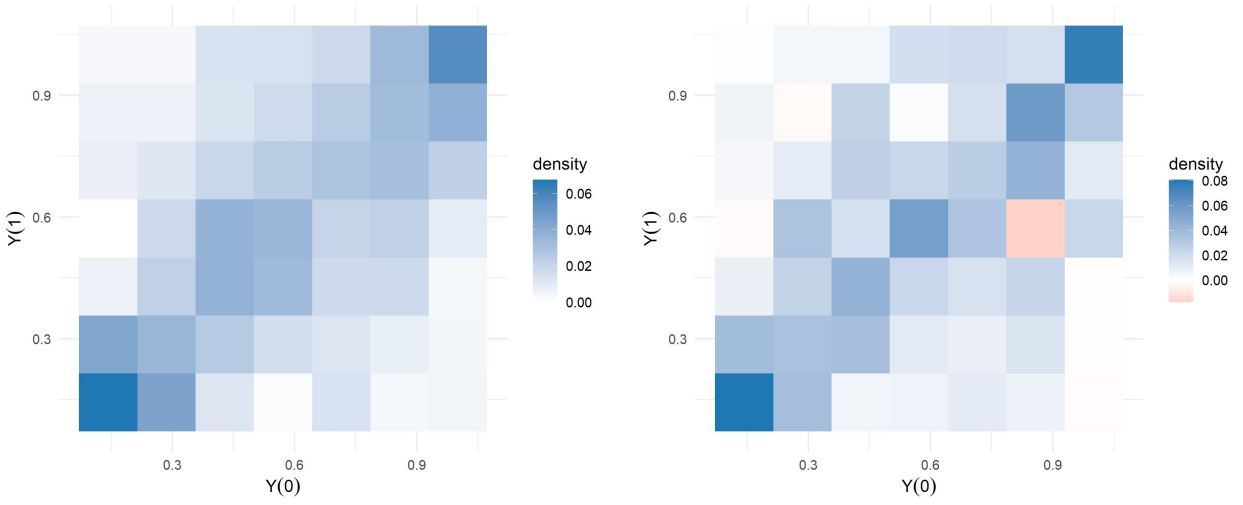
K	3	4	5	6
T_n^1	17.68	27.07	16.79	47.66
p-value	0.477	0.301	0.975	0.092
T_n^2	1.57	0.22	0.24	4.27
p-value	0.666	0.995	0.999	0.640

Table 7: Falsification test statistics $\left(T_n^1, T_n^2\right)$ 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 $K = 3, 4, 5, 6$. Firstly, Figure 4 plots the estimated joint distribution of the two potential outcomes from the NMF algorithm with $K = 4, 5$. For visibility, I first 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 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 5 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 $[-\$1000, \$1000]$.



(a) $K = 4$ (b) $K = 5$

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

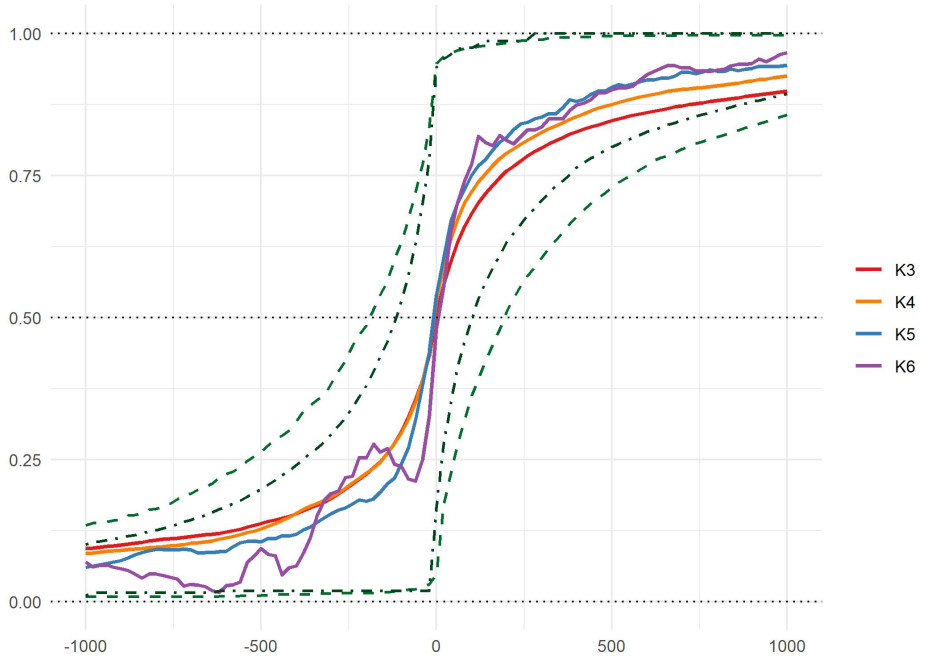


Figure 5: Marginal distribution of $Y_i(1) - Y_i(0)$, across $K = 3, \cdots, 6$.

Figure 5 is plotted by evaluating the DTE function $F_{Y(1)-Y(0)}(\delta)$ at 101 points; out of 100 increments, the estimated distribution function with $K = 6$ decreases for 37 increments. Moreover, I also constructed a 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 supports the conjecture that the NMF estimator overfits the data matrix $\mathbb{H}_0$ and $\mathbb{H}_1$ when $K = 6$.

Also, the right tail of the distribution function seems to be sensitive to the choice of K. In particular, it suggests that the direction of the misspecification/discretization bias from using a smaller-than-true K is negative; 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 the quadratic constraints that $Y_i(d)$ and X_i are independent of each other conditional on U_i. Fix some (y, d, x, k) and let

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

Then,

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

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

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

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

Secondly, ϕ_B is the part of ϕ that corresponds to the law of iterated expectation. Fix some (d, x) and let

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

Then,

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

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

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

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

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

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

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

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

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

Also,

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

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

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

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

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

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

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

It suffices to show that the submatrix

$$\begin{pmatrix} \mathbf{E}\left[\frac{\partial}{\partial \tilde{\lambda}}\phi_A\left(W_i, W_{i'}; \tilde{\lambda}, p\right)\right] & \mathbf{E}\left[\frac{\partial}{\partial \tilde{\lambda}}\phi_B\left(W_i, W_{i'}; \tilde{\lambda}, p\right)\right] \\ \mathbf{O}_{2K \times 2MK} & \mathbf{E}\left[\frac{\partial}{\partial p_{D,U}}\phi_B\left(W_i, W_{i'}; \tilde{\lambda}, p\right)\right] \end{pmatrix}. \tag{4}$$

is full rank. Assume to the contrary that the rows of the submatrix from (4) are linearly

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

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

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

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

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

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

$\mathbf{E}\left[\frac{\partial}{\partial p_{D,U}}\phi_B\left(W_i, W_{i'}; \tilde{\lambda}, p\right)\right]$ 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

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

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

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

Then,

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

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:

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

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

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

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

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

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

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

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

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

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

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

and

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

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

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

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

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

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

which leads to a contradiction. Thus, we have

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

Two elements of $a_{\cdot 1}$ being away from zero creates a contradiction to $\left\| \widehat{\Gamma}_0 - \Gamma_0 A \right\|_F = O_p\left(\frac{1}{\sqrt{n}} \right)$ since the convergence says that each column of $\Gamma_0 A$ can be well-approximated by

a column in $\widehat{\Gamma}_0$, which satisfies the quadratic constraints. To see this, let $\tilde{\Gamma}_{0,k}$ be a $M_X \times M_Y$ matrix whose m-th row and m'-th column element is

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

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

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

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

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

with $\widehat{\tilde{\Gamma}}_{0,k}$ constructed from $\widehat{\Gamma}_0$ in the same manner as $\tilde{\Gamma}_{0,k}$. The first inequality holds from the construction of the estimator $\widehat{\Gamma}_0$; the estimated mixture component distribution satisfies the exclusion restriction of $Y_i(0)$ and X_i given U_i and thus $\widehat{\tilde{\Gamma}}_{0,1}$ is a rank one matrix. The second inequality holds since $\sum_{k=1}^K \tilde{\Gamma}_{0,k} a_{k1}$ corresponds to the first column of $\Gamma_0 A$ and $\widehat{\tilde{\Gamma}}_{0,1}$ 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 $p_k(x) = \Pr\{X_i = x | U_i = u^k\}$, $q_{dk}(y) = \Pr\{Y_i(d) = y | U_i = u^k\}$ for $k = 1, \cdots, K$ and let

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

Then,

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

From Assumption 3.c,

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

WLOG let y^1 and y^2 satisfy that

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

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

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

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

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

Then,

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

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 $\begin{pmatrix} w(y^1) & w(y^2) \end{pmatrix}$ is bounded away from zero with a lower bound proportional to $\frac{1}{\sqrt{n}}$, I use the following result:

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

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

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

Find that

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

We have shown that

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

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

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

Thus, with probability going to one,

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

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

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

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

Step 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

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

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.

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

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