

# Distributional Treatment Effect with Latent Rank Invariance

Myungkou Shin

University of Surrey

Bristol econometrics study group conference

July 12, 2025

## Distributional treatment effect

Potential outcome setup: with  $D \in \{0, 1\}$ ,

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

We do not observe  $Y(1)$  and  $Y(0)$  simultaneously; focus on ATE, LATE, etc.

Some questions can only be answered with **distribution** of treatment effect  $Y(1) - Y(0)$ .

*“How many people are better off under the treatment?”*

*“How heterogeneous is the treatment effect at the individual level?”*

Various distributional treatment effect (DTE) parameters can be defined:

$\text{Var}(Y(1) - Y(0))$ ,  $\Pr\{Y(1) - Y(0) \geq 0\}$ , etc.

## Distributional treatment effect

### Existing approaches

- Partial identification: put a bound on  $\Pr \{Y(1) - Y(0) \leq y\}$

Heckman et al. (1997); Fan and Park (2010); Fan et al. (2014); Firpo and Ridder (2019)  
Frandsen and Lefgren (2021); Kaji and Cao (2023) and more

- Independence: assume  $Y(1) \perp\!\!\!\perp Y(0)$  or  $Y(0) \perp\!\!\!\perp (Y(1) - Y(0))$

Heckman et al. (1997); Gautier and Hoderlein (2015); Noh (2023)

In this paper, I follow the latter, assuming a latent variable  $U$  such that

$$Y(1) \perp\!\!\!\perp Y(0) \mid U$$

and two proxy variables  $X$  and  $Z$  to identify the cond. dist. of  $Y(d)$  given  $U$ .

## Distributional treatment effect: setup

An econometrician observes  $\{Y_i, D_i, X_i, Z_i\}_{i=1}^n$ :

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

$Y_i, X_i, Z_i \in \mathbb{R}$ ,  $D_i \in \{0, 1\}$  and  $(Y_i(1), Y_i(0), D_i, X_i, Z_i, U_i) \sim iid$ .

$X_i$  and  $Z_i$  are proxy variables for  $U_i$ .  $U_i \in \mathbb{R}$ .

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

- One of the proxy  $Z_i$  and the latent variable  $U_i$  are confounders.

- In proximal inference terminology,

$X_i$  is outcome-aligned proxy and  $Z_i$  is treatment-aligned proxy.

Hu and Schennach (2008); Miao et al. (2018); Deaner (2023); Nagasawa (2022) and more

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

## Distributional treatment effect: example 1 (*rank invariance*)

Assume **rank invariance** between  $Y_i(1)$  and  $Y_i(0)$ :

$$\Pr \{F_{Y(1)}(Y_i(1)) = F_{Y(0)}(Y_i(0))\} = 1.$$

When  $D_i$  is random, Assumptions 1-2 trivially hold with

$$U_i = X_i = Z_i = F_{Y(1)}(Y_i(1)) = F_{Y(0)}(Y_i(0)).$$

Rank invariance is a commonly used assumption in quantile treatment effect/IV literature:

Chernozhukov and Hansen (2005, 2006); Athey and Imbens (2006); Callaway and Li (2019) and more.

In this paper,  $U_i$  is not a deterministic function of  $Y_i(1)$  or  $Y_i(0)$ ; hence '*latent* rank invariance.'

## Distributional treatment effect: example 2 (*panel w/ latent, Markovian state*)

Deaner (2023): consider a panel where  $T = 3$  and  $D_i = 1$  means being treated for  $t = 2, 3$ .

$$Y_{it}(d) = g_d(U_{it}, \varepsilon_{it}(d)).$$

There are a common shock  $U_{it}$  and treatment-status-specific shocks  $(\varepsilon_{it}(0), \varepsilon_{it}(1))$ .

Assumption 2 holds when 1)  $\{U_{it}\}_{t=1}^3$  is first-order Markov and

2)  $\{U_{it}\}_{t=1}^3, \varepsilon_{i1}(0), \varepsilon_{i2}(1), \varepsilon_{i2}(0), \varepsilon_{i3}(1), \varepsilon_{i3}(0) \sim \text{ind.}$

Hidden Markov model: Kasahara and Shimotsu (2009); Hu and Shum (2012) and more.

I add treatment-status-specific shocks and assume  $\varepsilon_{it}(1) \perp\!\!\!\perp \varepsilon_{it}(0)$ .

## Distributional treatment effect: example 3 (*repeated measurements*)

In some empirical contexts,  $U_i$  has economic interpretation;

Assumption 2 holds with some repeated (independent) measurements of  $U_i$ .

Carneiro et al. (2003); Cunha et al. (2010) and more.

In Carneiro et al. (2003),  $Y_i(1), Y_i(0)$  are potential earnings and  $D_i$  is college education.  
 $U_i$  is the latent ability of a student and  $(X_i, Z_i)$  are test scores.

$$Y_i(1) = U_i^\top f^1 + \varepsilon_i(1),$$

$$Y_i(0) = U_i^\top f^0 + \varepsilon_i(0),$$

$$X_i = U_i^\top f^x + \varepsilon_i^x,$$

$$Z_i = U_i^\top f^z + \varepsilon_i^z.$$

I am relaxing the factor structure.

## Identification

Two identification results: one for finitely discrete  $U_i$  and another for continuous  $U_i$ .

Assumption 3 and 4 are **full rank/completeness** assumptions on  $f_{X|Z}$  in each case. A3 A4

For continuous  $U_i$ , I additionally invoke

**Assumption 5.**  $\mathbf{E}[Y_i(1)|U_i = u]$  and  $\mathbf{E}[Y_i(0)|U_i = u]$  are strictly increasing in  $u$ .

“Conditional expectation of  $Y_i(1)$  given  $U_i$  and that of  $Y_i(0)$  given  $U_i$  have the same rank.”

$U_i$  can be thought of as a ‘latent’ or ‘interim’ rank.



I apply Hu and Schennach (2008) to treated subpopulation and to untreated subpopulation.

[more](#)

### Theorem 1.

Let Assumptions 1-3 or Assumptions 1-2, 4-5 hold. Then, the joint distribution of  $(Y_i(1), Y_i(0))$  and thus the distribution of the treatment effect  $Y_i(1) - Y_i(0)$  are identified.

$$f_{Y(1), Y(0)}(y, y') = \int_{\mathbb{R}} f_{Y(1), Y(0)|U}(y, y'|u) du = \int_{\mathbb{R}} f_{Y(1)|U}(y|u) \cdot f_{Y(0)|U}(y'|u) du,$$
$$f_{Y(1) - Y(0)}(\delta) = \int_{\mathbb{R}} f_{Y(1) - Y(0)|U}(\delta|u) du = \int_{\mathbb{R}} \int_{\mathbb{R}} f_{Y(1)|U}(y + \delta|u) \cdot f_{Y(0)|U}(y|u) dy du.$$

I focus on two DTEs:  $F_{Y(1), Y(0)}$  and  $F_{Y(1) - Y(0)}$ .

## Identification: implicit restriction

A crucial step in the identification argument is that there exists some  $w$  such that

$$\begin{aligned}\mathbf{E}[Y_i(1)|Y_i(0) = y] &= \int_{\mathbb{R}} \frac{w(y, z)}{f_{Y(0)}(y)} \cdot \mathbf{E}[Y_i|D_i = 1, Z_i = z]dz, \\ \mathbf{E}[Y_i(1)Y_i(0)] &= \int_{\mathbb{R}} \int_{\mathbb{R}} w(y, z) \cdot y \mathbf{E}[Y_i|D_i = 1, Z_i = z]dydz.\end{aligned}$$

$\mathbf{E}[Y_i|D_i = 1, Z_i]$  replaces  $Y_i(1)$  and  $w(y, z)$  replaces the joint density of  $(Y_i(1), Y_i(0))$ .

“Proxy variable  $Z_i$  creates sufficient variation in the distribution of  $Y_i(1)$ .”

The implicit restriction is that

“conditional distribution of  $Y_i(1)$  given  $Y_i(0)$  is a linear combination of  $\{F_{Y|D=1, Z}(\cdot|z)\}_z$ .”

## Identification: falsification test

The conditional independence of  $Y_i(1)$  and  $Y_i(0)$  is fundamentally untestable.  
Instead, we can test the null

$$f_{X|D=1,U}(\cdot|u) = f_{X|D=0,U}(\cdot|u) \quad \forall u.$$

with

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

“Can we construct a latent variable  $U_i$  that satisfies 1) conditional independence  $X_i \perp\!\!\!\perp Y_i \mid U_i$  and  
2) random treatment  $X_i \perp\!\!\!\perp D_i \mid U_i$ ?”

In the short panel context,

- cannot test the conditional independence *across treatment regime*.
- can somewhat test the *intertemporal* conditional independence, given random treatment.

## Implementation

The estimation strategy is two-step:

**Step 1.** Estimate  $f_{U|D=d,Z}$  using *nonnegative matrix factorization*.

- Decompose  $\mathbf{H} = \left( f_{Y,X|D=d,Z}(y, x|z) \right)_{(y,x),z}$  into  $\left( f_{Y(d),X|U}(y, x|u) \right)_{(y,x),u}$  and  $\left( f_{U|D=d,Z}(u|z) \right)_{u,z}$ .

**Step 2.** *Plug-in GMM* to estimate DTE.

- DTE parameters are quadratic moments of  $(Y_i, D_i, Z_i)$ .
- $f_{U|D=d,Z}$  is the nuisance parameter of GMM.

Assume  $|\text{supp}_U| < \infty$  for asymptotic theory.

## Implementation: finite support assumption

To let  $U_i$  be continuous and apply the standard semiparametric theory, Shen (1997); Chen and Shen (1998); Ai and Chen (2003) and more we need different strong assumptions such as bounded support of  $Y_i$  and  $X_i$ . Why? DTE parameters are inner products of densities. sieve

Instead, I assume  $U_i \in \{u^1, \dots, u^K\}$  with known  $K < \infty$ .

Reasoning behind the finite support assumption:

1. Finite mixture: Henry et al. (2014) and more.  
Discretization as approximation: Bonhomme et al. (2022) and more.
2. DTE parameters are identified with quadratic moments;  
a limiting distribution is derived from U stat. theory and Neyman orthogonality.
3. Identification is not tied to the finite support;  
alternative asymptotic theory could be developed as well.

## Implementation: nonnegative matrix factorization

Fix  $K$ . Given some partitions  $\{\mathcal{Y}^m\}_m, \{\mathcal{X}^{m'}\}_{m'}, \{\mathcal{Z}^l\}_l$ , let

$$\begin{aligned}\mathbf{H}_d &= \left( \Pr \left\{ Y_i \in \mathcal{Y}^m, X_i \in \mathcal{X}^{m'} \mid D_i = d, Z_i \in \mathcal{Z}^l \right\} \right)_{(m,m'),l} \\ &= \Gamma_d \cdot \Lambda_d\end{aligned}$$

where  $\Gamma_d = \left( \Pr \left\{ Y_i(d) \in \mathcal{Y}^m, X_i \in \mathcal{X}^{m'} \mid U_i = u^k \right\} \right)_{(m,m'),k}$

$$\Lambda_d = \left( \Pr \left\{ U_i = u^k \mid D_i = d, Z_i \in \mathcal{Z}^l \right\} \right)_{k,l}.$$

$\mathbf{H}_d$  is a discretization of  $f_{Y,X|D=d,Z}$ .

The full rank condition implies  $|\text{supp}_Z| \geq K$ ; if  $|\text{supp}_Z| > K$ , use partition  $\{\mathcal{Z}^l\}_{l=1}^K$ .

## Implementation: nonnegative matrix factorization

Solve the following nonnegative matrix factorization problem:

$$\left(\hat{\Gamma}_0, \hat{\Gamma}_1, \hat{\Lambda}_0, \hat{\Lambda}_1\right) = \arg \min \|\mathbb{H}_0 - \Gamma_0 \cdot \Lambda_0\|_F + \|\mathbb{H}_1 - \Gamma_1 \cdot \Lambda_1\|_F \quad (1)$$

subject to 1)  $\Gamma_0, \Gamma_1, \Lambda_0, \Lambda_1$  are nonnegative.

Also, their columnwise sums are one.  $\dots$  (*linear constraints*)

2)  $\Gamma_0$  and  $\Gamma_1$  satisfy  $Y_i(d) \perp\!\!\!\perp X_i \mid U_i \dots$  (*quadratic constraints*)

3)  $\Gamma_0$  and  $\Gamma_1$  imply the same marginal distribution of  $X_i \dots$  (*linear constraints*)

The objective becomes quadratic once we fix  $(\Gamma_0, \Gamma_1)$  or  $(\Lambda_0, \Lambda_1)$ .

The quadratic constraint becomes linear once we fix  $\Gamma_X$  or  $(\Gamma_{Y0}, \Gamma_{Y1})$ .

(1) is solved iteratively. algorithm

## Implementation: nonnegative matrix factorization

**Theorem 2.** Under Assumptions 1-3,

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

as  $n \rightarrow \infty$ , up to some permutation on  $\{1, \dots, K\}$ .

The convergence rate is  $n^{-\frac{1}{2}}$ .

No additional assumptions needed; Assumptions 1-2 and full rank of  $\mathbf{H}_X$ .



## Implementation: plug-in GMM

$\Lambda_0$  and  $\Lambda_1$  nuisance parameters for DTE since

$$\begin{aligned} & \left( F_{Y(d)|U}(y|u^1) \quad \cdots \quad F_{Y(d)|U}(y|u^K) \right) \\ &= \left( F_{Y|D=d,Z}(y|\mathcal{Z}^1) \quad \cdots \quad F_{Y|D=d,Z}(y|\mathcal{Z}^K) \right) (\Lambda_d)^{-1}. \end{aligned}$$

Distribution of  $Y_i(d)$  given  $U_i$  are linear in (observed) distribution of  $Y_i$  given  $D_i = d, Z_i$ .

Let  $\tilde{\Lambda}_d = \left( \tilde{\lambda}_{lk,d} \right)_{l,k} := (\Lambda_d)^{-1}$  for  $d = 0, 1$ .

We always get  $\sum_{l=1}^K \tilde{\lambda}_{lk,d} = 1$  but  $\tilde{\lambda}_{lk,d}$  may be negative.

"Extrapolation may need to happen unless  $F_{Y|D=d,Z}(\cdot|\mathcal{Z}) = F_{Y(d)|U}(\cdot|u)$  for some  $\mathcal{Z}$ ."

## Implementation: plug-in GMM

Let  $p_U(k) := \Pr\{U_i = u^k\} \quad \forall k = 1, \dots, K$

$p_{D,Z}(d, l) := \Pr\{D_i = d, Z \in \mathcal{Z}^l\} \quad \forall d = 0, 1 \text{ and } l = 1, \dots, K.$

Then, quadratic moments identify DTE: with  $w_{klm} = \frac{p_U(k) \tilde{\lambda}_{lk,0} \tilde{\lambda}_{mk,1}}{p_{D,Z}(0, l) p_{D,Z}(1, m)},$

$$F_{Y(1), Y(0)}(y, y') = \sum_{k,l,m=1}^K w_{klm} \cdot \mathbf{E}[\mathbf{1}\{Y_i \leq y, D_i = 1, Z_i \in \mathcal{Z}^m, Y_j \leq y', D_j = 0, Z_j \in \mathcal{Z}^l\}]$$

$$F_{Y(1)-Y(0)}(\delta) = \sum_{k,l,m=1}^K w_{klm} \cdot \mathbf{E}[\mathbf{1}\{Y_i \leq Y_j + \delta, D_i = 1, Z_i \in \mathcal{Z}^m, D_j = 0, Z_j \in \mathcal{Z}^l\}]$$

for all  $(y, y') \in \mathbb{R}^2$  and  $\delta \in \mathbb{R}$ , with  $(Y_i, D_i, Z_i) \perp\!\!\!\perp (Y_j, D_j, Z_j).$

## Implementation: plug-in GMM

Our (naive) estimator is a plug-in U statistics.

$$\begin{aligned} & \widehat{F}_{Y(1)-Y(0)}(\delta) \\ &= \sum_{k,l,m=1}^K \hat{w}_{klm} \cdot \binom{n}{2}^{-1} \sum_{i \neq j} \left( \frac{1}{2} \mathbf{1}\{Y_i \leq Y_j + \delta, D_i = 1, Z_i \in \mathcal{Z}^m, D_j = 0, Z_j \in \mathcal{Z}^l\} \right) \end{aligned}$$

and similarly for  $\widehat{F}_{Y(1),Y(0)}$ .

When the nuisance parameters  $\{\tilde{\lambda}_{lk,0}, \tilde{\lambda}_{lk,1}\}_{l,k}$  and  $\{p_U(k), p_{D,Z}(d,k)\}_{d,k}$  are known, the standard U statistics asymptotic theory applies.

In fact, (uniform) consistency is a direct corollary of Theorem 1.

## Implementation: Neyman orthogonality

$\tilde{\Lambda}$  is estimated with  $n^{-\frac{1}{2}}$  rate.

To be robust to the first step estimation error, use an orthogonal score.

Three sets of nuisance parameters:  $\{p_{D,Z}(d, k)\}_{d,k}$ ,  $\{p_U(k)\}_k$  and  $\{\tilde{\lambda}_{lk,d}\}_{l,k,d}$ .

For  $\{p_U(k)\}_k$  and  $\{\tilde{\lambda}_{lk,d}\}_{l,k,d}$ , use the quadratic constraints of conditional independence: [more](#)

$$\Pr \{Y_i \in \mathcal{Y}, X_i \in \mathcal{X} | U_i = u\} = \Pr \{Y_i \in \mathcal{Y} | U_i = u\} \cdot \Pr \{X_i \in \mathcal{X} | U_i = u\}$$

## Implementation: Neyman orthogonality

Let  $m$  be the score function for a DTE parameter  
and  $\phi$  be the score function for the nuisance parameters.

The **orthogonalized score** is

$$m(W_i, W_j) - \mu^\top \phi(W_i, W_j)$$

where  $W_i = (Y_i, D_i, X_i, Z_i)$  and  $\mu = \begin{pmatrix} \mathbf{E}[\frac{\partial}{\partial \lambda} \phi] \\ \mathbf{E}[\frac{\partial}{\partial p} \phi] \end{pmatrix}^+ \begin{pmatrix} \mathbf{E}[\frac{\partial}{\partial \lambda} m] \\ \mathbf{E}[\frac{\partial}{\partial p} m] \end{pmatrix}$ .

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

$$\begin{pmatrix} \mathbf{E}[\frac{\partial}{\partial \lambda} \phi] \\ \mathbf{E}[\frac{\partial}{\partial p} \phi] \end{pmatrix}$$

has full rank.

## Implementation: Neyman orthogonality

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

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

as  $n \rightarrow \infty$ .

Asymptotic variances are consistently estimated.

## Implementation: choice of $K$

Choice of  $K$  is a nontrivial issue.

When using more partitions than needed,

$$\mathbf{H}_d = \left( \Pr \left\{ Y_i \in \mathcal{Y}^m, X_i \in \mathcal{X}^{m'} \mid D_i = d, Z_i \in \mathcal{Z}^l \right\} \right)_{(m,m'),l}$$

is not full rank.

Cragg and Donald (1997); Kwon and Mbakop (2021); Bai and Ng (2002) and more.

In the empirical illustration, I used smallest  $K$  such that

$$U_i \mid D_i = 0 \stackrel{d}{=} U_i \mid D_i = 1$$

since the treatment  $D_i$  was randomly assigned.

## Simulation

Monte Carlo simulations with a simple DGP with  $K = 3$  and  $Y_i, X_i, Z_i \in \{1, 2, 3\}$ .

Nonnegative matrix factorization is applied to two  $9 \times 3$  matrices.

Informativeness of the two proxy variables:

$$\Gamma_X = \left( \Pr\{X_i = x | U_i = u^k\} \right)_{x,k} = \begin{pmatrix} 0.800 & 0.100 & 0.067 \\ 0.133 & 0.800 & 0.133 \\ 0.067 & 0.100 & 0.800 \end{pmatrix}$$
$$\Lambda = \left( \Pr\{U_i = u^k | Z_i = z\} \right)_{z,k} = \begin{pmatrix} 0.840 & 0.091 & 0.040 \\ 0.077 & 0.772 & 0.055 \\ 0.083 & 0.137 & 0.905 \end{pmatrix}.$$

Their smallest singular values are 0.665 and 0.701.



## Simulation

As we shift  $\Lambda$ , estimation worsens:

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

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

First step NMF worsens as  $Z_i$  gets less informative.

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

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

## Empirical Illustration

I revisit Jones et al. (2019), which studies the effect of workplace wellness program. The program *eligibility* was randomly assigned to employees at UIUC; *intent-to-treat*. Using the University-provided health insurance data, Jones et al. (2019) estimates its effect on medical spending.

The variables in the dataset are:

$Y_i$  = monthly medical spending over August 2016-July 2017

$D_i = \mathbf{1}\{\textit{eligible for the wellness program starting in September 2016}\}$

$X_i$  = monthly medical spending over July 2015-July 2016

$Z_i$  = monthly medical spending over August 2017-January 2019

*“Underlying health status  $U_i$  depends on past health status, but not on medical spendings.”*

## Empirical Illustration: setup

I used  $K = 5$ .

Partitions are constructed with  $F_Y^{-1}(0), F_Y^{-1}(1/5), \dots, F_Y^{-1}(1)$  and so on.

The test statistic on the null hypothesis  $f_{X|D=1,U}(\cdot|u) = f_{X|D=0,U}(\cdot|u)$  for all  $u$ :  
with  $W_n = \left( \hat{f}_{X|D=1,U}(\mathcal{X}^m|u) - \hat{f}_{X|D=0,U}(\mathcal{X}^m|u) \right)_{m,u} \in \mathbb{R}^{25}$ ,

$$nW_n^\top Avar(W)^{-1}W_n = 16.435$$

The  $p$ -value is 0.901 .

## Empirical Illustration: joint distribution of potential outcomes

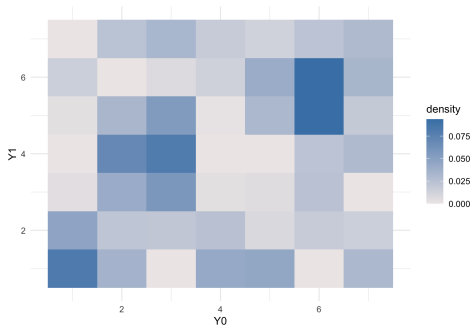


Figure 1: Joint density of  $(Y(1), Y(0))$ .

$y$ -axis is  $Y(1)$  and  $x$ -axis  $Y(0)$ ; each cell corresponds to  $F_X^{-1}(0), F_X^{-1}(1/7), \dots, F_X^{-1}(1)$ .

No noticeable treatment effect; in Jones et al. (2019),  $p$ -values for ATE are 0.86-0.94.

## Empirical Illustration: treatment effect distribution

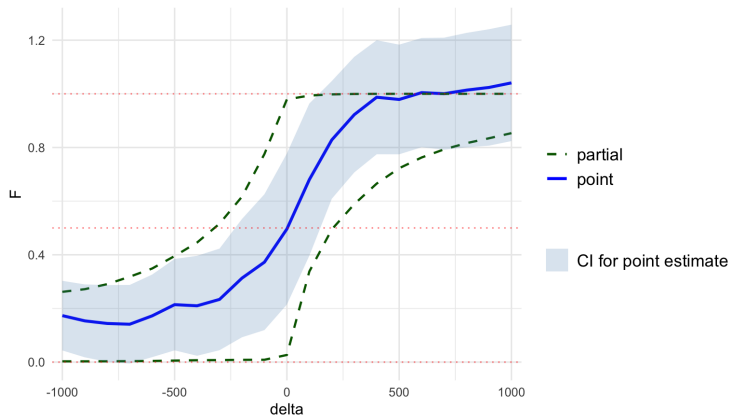


Figure 2: Marginal density of  $Y(1) - Y(0)$ .

Unclear whether the probability of getting benefited is bigger or smaller than 0.5. Thicker left tail.

## Summary

- Assume a latent variable  $U$  such that

$$Y_i(1) \perp\!\!\!\perp Y_i(0) \mid U_i.$$

This assumption could be thought of as a ‘latent rank invariance’ condition when  $\mathbf{E}[Y_i(d)|U_i = u]$  is monotone increasing in  $u$ .

- Use two proxy variables  $X_i$  and  $Z_i$  to identify the distribution of  $Y_i(d)|U_i$ .
- Nonnegative matrix factorization estimates distribution of  $U_i$  given  $(D_i, Z_i)$ .
- An asymptotic distribution is derived for the plug-in GMM estimator.

## Identification à la Hu and Schennach (2008)

An essential building block in the identification argument:  $f_{Y,X|D,Z}$ .

Fix  $y$  and  $d$  and discretize  $X_i$  and  $Z_i$ :

$$\mathbf{H} = \left( f_{Y=y, X|D=d, Z}(x|z) \right)_{x,z} = \left( f_{X|U}(x|u) \right)_{x,u} \cdot \text{diag} \left( f_{Y|U}(y|u) \right)_u \cdot \left( f_{U|D=d, Z}(u|z) \right)_{u,z}.$$

$H$  is a  $|\text{supp}_X| \times |\text{supp}_Z|$  matrix whose rows correspond to  $X_i$  and columns to  $Z_i$ .

Likewise, define  $\mathbf{H}_X = \left( f_{X|D=d, Z}(x|z) \right)_{x,z}$ .

Under Assumptions 1-2 and **full rank/completeness** of  $\mathbf{H}_X$ , A3 A4

$$\mathbf{H} \cdot (\mathbf{H}_X)^{-1} = \left( f_{X|U}(x|u) \right)_{x,u} \cdot \text{diag} \left( \{ f_{Y(d)=y|U}(u) \}_u \right) \cdot \left( \left( f_{X|U}(x|u) \right)_{x,u} \right)^{-1}$$

Spectral decomposition identifies  $f_{X|U}$ .



## Spectral Theorem of Hu and Schennach (2008)

Several deviations from Hu and Schennach (2008):

1. Two decomposition results; treated population and untreated population.

Need to connect  $\{f_{Y(1)|U}(\cdot|u)\}_u$  to  $\{f_{Y(0)|U}(\cdot|u)\}_u$ .

2. Mapping from  $\{f_{X|U}(\cdot|u)\}_u$  to  $u$  to have distribution of  $U_i$ .

1. is easily solved.

Firstly, split the sample into two subsamples  $\{i : D_i = 1\}$  and  $\{i : D_i = 0\}$  and we get  $\{f_{Y(1)|U}(\cdot|u), f_{X=1|U}(\cdot|u)\}_u$  and  $\{f_{Y(0)|U}(\cdot|u), f_{X=0|U}(\cdot|u)\}_u$ .

Under Assumption 1,  $f_{X|D=1,U}(\cdot|u)$  and  $f_{X|D=0,U}(\cdot|u)$  should be the same.

## Spectral Theorem of Hu and Schennach (2008)

A linear operator  $L_{Y=y, X|D=d, X}$  maps a density of  $Z_i$  to a density of  $(Y_i(d) = y, X_i)$ :

$$(L_{Y=y, X|D=d, Z} g)(x) = \int_{\mathbb{R}} f_{Y(d), X|D, Z}(y, x|d, z) g(z) dz.$$

From the decomposition based on Assumption 2, we get

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

with similarly defined operators  $L_{X|U}$ ,  $L_{U|D=d, Z}$  and a diagonal operator  $\Delta_{Y=y|U}$ . Thus,

$$\begin{aligned} L_{Y=y, X|D=d, Z} (L_{X|D=d, Z})^{-1} &= L_{X|U} \cdot \Delta_{Y=y|U} \cdot L_{U|D=d, Z} \cdot (L_{X|U} \cdot L_{U|D=d, Z})^{-1} \\ &= \underbrace{L_{X|U} \cdot \Delta_{Y=y|U}}_{\text{spectral decomposition}} \cdot (L_{X|U})^{-1}. \end{aligned}$$

## Assumption 3

### Assumption 3.

- a. (*finitely discrete  $U_i$* )  $U_i \in \{u^1, \dots, u^K\}$ .
- b. (*full rank*)  $\left(f_{U|D=1,Z}(u|z)\right)_{u,z}$ ,  $\left(f_{U|D=0,Z}(u|z)\right)_{u,z}$  and  $\left(f_{X|U}(x|u)\right)_{x,z}$  have rank  $K$ .
- c. (*no repeated eigenvalue*) For any  $k \neq k'$ , there exist some  $d \in \{0, 1\}$  and  $y$  such that

$$f_{Y(d)|U}(y|u^k) \neq f_{Y(d)|U}(y|u^{k'}).$$

"The latent heterogeneity  $U_i$  can be *at most* as rich/flexible as the proxy variables." [back](#)

## Assumption 4

### Assumption 4.

- a. (continuous  $U_i$ )  $U_i \in [0, 1]$ .
- b. (bounded density) All marginal and conditional densities of  $(Y_i(1), Y_i(0), X_i, Z_i, U_i)$  are bounded.
- c. (completeness) Let  $f_{X|Z,d}$  denote the conditional density of  $X_i$  given  $(D_i = d, Z_i)$ .

$$\int_{\mathbb{R}} |g(x)| dx \quad \text{and} \quad \int_{\mathbb{R}} g(x) f_{X|Z,d}(x|z) dx = 0 \quad \forall d, z$$

implies  $g(x) = 0$ . Assume similarly for  $f_{X|U}$ .

- d. (no repeated eigenvalue)  $\forall u \neq u'$ , there exists  $d \in \{0, 1\}$  such that

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

## Nonnegative matrix factorization

The objective function in (1) is quadratic with linear constraints, once we fix two out of the three matrices  $\Gamma_X, \Gamma_Y, \Lambda$ .

Thus, find the (local) minima by iterating among three objects:

1. Given  $(\Gamma_0^{(s)}, \Gamma_1^{(s)})$ , update  $(\Lambda_0, \Lambda_1)$ .
2. Given  $(\Lambda_0^{(s+1)}, \Lambda_1^{(s+1)}, \Gamma_{Y0}^{(s)}, \Gamma_{Y1}^{(s)})$ , update  $\Gamma_X$ .
3. Given  $(\Lambda_0^{(s+1)}, \Lambda_1^{(s+1)}, \Gamma_X^{(s+1)})$ , update  $(\Gamma_{Y0}, \Gamma_{Y1})$ .
4. Iterate **1-3** until convergence.

In practice, use many initial values to find the global minimum.

[back](#)

## Sieve MLE

To allow for a continuous  $U_i$ , we can directly construct a likelihood using sieves:

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

The latent rank interpretation is simple to impose with Bernstein polynomials:  
a Bernstein polynomial of degree  $m$  is

$$g_m(u) = \sum_{k=0}^m \theta_k u^k (1-u)^{m-k}.$$

Then, monotonicity of  $\int_0^1 u g_m(u) du$  is a set of linear constraints on  $\{\theta_k\}_{k=0}^m$ .

**Theorem 4.** Let Assumptions 1-2,4-6 hold. Then,

$$\left\| \hat{f}_{Y(1),Y(0)} - f_{Y(1),Y(0)} \right\|_{\infty} \xrightarrow{p} 0$$

as  $n \rightarrow \infty$  and for any  $(y, y') \in \mathbb{R}^2$  and  $\delta \in \mathbb{R}$ ,

$$\begin{aligned} \sqrt{n} \left( \hat{f}_{Y(1),Y(0)}(y, y') - f_{Y(1),Y(0)}(y, y') \right) &\xrightarrow{d} \mathcal{N}(0, \sigma(y, y')^2) \\ \sqrt{n} \left( \widehat{\Pr \{Y_i(1) - Y_i(0) \leq \delta\}} - \Pr \{Y_i(1) - Y_i(0) \leq \delta\} \right) &\xrightarrow{d} \mathcal{N}(0, \sigma(\delta)^2) \end{aligned}$$

as  $n \rightarrow \infty$ .

## Assumption 6 I

### Assumption 6

- a. Functions in  $\{\Theta_n\}_{n=1}^{\infty} \cup \Theta$  is uniformly bounded.  $\Theta$  is convex.
- b.  $f_{Y(1)|U}, f_{Y(0)|U}, f_{X|U}, f_{U|D=1,Z}, f_{U|D=0,Z}$  are in the interior of  $\Lambda_c^{\gamma_1}([0, 1]^2)$  with  $\gamma_1 > 1$ . Also, for any  $\theta \in \Theta_n$  for some  $n$ ,

$$f_{Y(1)|U,n}(\cdot; \theta), f_{Y(0)|U,n}(\cdot; \theta), f_{X|U,n}(\cdot; \theta), f_{U|D=1,Z,n}(\cdot; \theta), f_{U|D=0,Z,n}(\cdot; \theta) \in \Lambda_c^{\gamma_1}([0, 1]^2)$$

and  $\log f_{Y,X|D,Z}(\cdot; \theta) \in \Lambda_c^{\gamma}([0, 1]^4)$  with  $\gamma > 2$ .

- c.  $\mathbf{E} \left[ (\log f_{Y,X|D,Z}(Y_i, X_i | D_i, Z_i))^2 \right] < \infty$ . There exists measurable functions  $h_1, h_2$  such that

$$\begin{aligned} & h_1(y, d, x, z) \\ & \leq \frac{1}{f_{Y,X|D,Z}(y, x | d, z; \theta)} \left( \int_0^1 \frac{f_{Y(d)|U}(y|u; \theta) f_{X|U}(x|u; \theta) f_{U|D=d,Z}(u|z; \theta)}{f_{Y(d)|U}(y|u; \theta) + f_{X|U}(x|u; \theta) + f_{U|D=d,Z}(u|z; \theta)} du \right) \\ & \leq h_2(y, d, x, z) \end{aligned}$$



## Assumption 6 II

for all  $\theta \in \Theta$  and  $\mathbf{E} [(h_1(Y_i, D_i, X_i, Z_i,)) ^2] , \mathbf{E} [(h_2(Y_i, D_i, X_i, Z_i))^2] < \infty$ . Also, There exist a measurable function  $h_3$  such that

$$\begin{aligned} & \frac{1}{2f_{Y,X|D,Z}(y, x|d, z; \theta)^2} \left( \int_0^1 \frac{f_{Y(d)|U}(y|u; \theta) f_{X|U}(x|u; \theta) f_{U|D=d,Z}(u|z; \theta)}{f_{Y(d)|U}(y|u; \theta) + f_{X|U}(x|u; \theta) + f_{U|D=d,Z}(u|z; \theta)} du \right)^2 \\ & + \frac{1}{f_{Y,X|D,Z}(y, x|d, z; \theta)} \int_0^1 (f_{Y(d)|U}(y|u; \theta) + f_{X|U}(x|u; \theta) + f_{U|D=d,Z}(u|z; \theta)) du \\ & \leq h_3(y, d, x, z) \end{aligned}$$

for all  $\theta \in \Theta$  and  $\mathbf{E} [(h_3(Y_i, D_i, X_i, Z_i,)) ^2] < \infty$ .

**d.**  $\|\Pi_n \theta^0 - \theta^0\|_\infty = o(n^{-\frac{1}{4}})$  as  $n \rightarrow \infty$  where

$$\Pi_n \theta^0 = \arg \max_{\theta \in \Theta_n} \mathbf{E} [\log f_{Y,X|D,Z}(Y_i, X_i|D_i, Z_i; \theta)]$$

Also,  $p_n \rightarrow \infty, \frac{p_n \log n}{\sqrt{n}} \rightarrow 0$  as  $n \rightarrow \infty$ .

## Assumption 6 III

e. With some  $c_1, c_2 > 0$ ,

$$c_1 \mathbf{E} \left[ \log \frac{f_{Y,X|D,Z}(Y_i, X_i | D_i, Z_i; \theta^0)}{f_{Y,X|D,Z}(Y_i, X_i | D_i, Z_i; \theta)} \right] \leq \|\theta - \theta^0\|^2 \leq c_2 \mathbf{E} \left[ \log \frac{f_{Y,X|D,Z}(Y_i, X_i | D_i, Z_i; \theta^0)}{f_{Y,X|D,Z}(Y_i, X_i | D_i, Z_i; \theta)} \right]$$

holds for any  $\theta \in \Theta_n$  such that  $\|\theta - \theta^0\|_\infty = o(1)$ .

f. Let  $p_1$  be the degree of a tensor product Bernstein polynomial used in approximating  $f_{Y(1)|U}$  to  $\Theta_n$  and similarly define  $p_0, p_X, p_{1Z}$  and  $p_{0Z}$ ; for example,  $p_1 = (p^y + 1) \cdot (p^u + 1)$ . With some abuse of notation, let  $\{\theta_{j,1}\}_{j=1}^{p_1}$  denote the basis functions used in approximating  $f_{Y(1)|U}$  and similarly define  $\{p_{j,0}\}_{j=1}^{p_0}, \dots, \{p_{j,0Z}\}_{j=1}^{p_{0Z}}$ .

## Assumption 6 IV

Let

$$\frac{d}{d\theta_1} \log f_{Y,X|D,Z}(Y_i, X_i|D_i, Z_i; \theta^0) [\{\theta_{j,1}\}_{j=1}^{p_1}] = \begin{pmatrix} \frac{d}{d\theta_1} \log f_{Y,X|D,Z}(Y_i, X_i|D_i, Z_i; \theta^0) [\theta_{1,1}] \\ \vdots \\ \frac{d}{d\theta_1} \log f_{Y,X|D,Z}(Y_i, X_i|D_i, Z_i; \theta^0) [\theta_{p_1,1}] \end{pmatrix}$$
$$W_n(Y_i, D_i, X_i, Z_i) = \begin{pmatrix} \frac{d}{d\theta_1} \log f_{Y,X|D,Z}(Y_i, X_i|D_i, Z_i; \theta^0) [\{\theta_{j,1}\}_{j=1}^{p_1}] \\ \vdots \\ \frac{d}{d\theta_{0Z}} \log f_{Y,X|D,Z}(Y_i, X_i|D_i, Z_i; \theta^0) [\{\theta_{j,0Z}\}_{j=1}^{p_{0Z}}] \end{pmatrix}$$

and

$$\Omega_n = \mathbf{E} [W_n(Y_i, D_i, X_i, Z_i) (W_n(Y_i, D_i, X_i, Z_i))^{\top}].$$

Then, the smallest eigenvalue of  $\Omega_n$  is bounded away from zero uniformly across  $n$ .

## Additional moments

The quadratic moment is

$$\begin{aligned} & \sum_{l=1}^K \frac{\tilde{\lambda}_{lk,d}}{p_{D,Z}(d,l)} \cdot \mathbf{E} \left[ \frac{1}{2} \mathbf{1}\{Y_i \in \mathcal{Y}, D_i = d, X_i \in \mathcal{X}, Z_i \in \mathcal{Z}^l\} \right] \\ & + \sum_{m=1}^K \frac{\tilde{\lambda}_{mk,d}}{p_{D,Z}(d,m)} \cdot \mathbf{E} \left[ \frac{1}{2} \mathbf{1}\{Y_j \in \mathcal{Y}, D_j = d, X_j \in \mathcal{X}, Z_j \in \mathcal{Z}^m\} \right] \\ & - \sum_{l=1}^K \sum_{m=1}^K \frac{\tilde{\lambda}_{lk,d} \tilde{\lambda}_{mk,d}}{p_{D,Z}(d,l) \cdot p_{D,Z}(d,m)} \mathbf{E} [\mathbf{1}\{Y_i \in \mathcal{Y}, D_i = d, Z_i \in \mathcal{Z}^l, X_j \in \mathcal{X}, D_j = d, Z_j \in \mathcal{Z}^m\}] = 0 \end{aligned}$$

with  $(Y_i, D_i, Z_i) \perp\!\!\!\perp (Y_j, D_j, Z_j)$ . [back](#)

## References I

- Ai, Chunrong and Xiaohong Chen**, "Efficient estimation of models with conditional moment restrictions containing unknown functions," *Econometrica*, 2003, 71 (6), 1795–1843.
- Athey, Susan and Guido W Imbens**, "Identification and inference in nonlinear difference-in-differences models," *Econometrica*, 2006, 74 (2), 431–497.
- Bai, Jushan and Serena Ng**, "Determining the number of factors in approximate factor models," *Econometrica*, 2002, 70 (1), 191–221.
- Bonhomme, Stéphane, Thibaut Lamadon, and Elena Manresa**, "Discretizing unobserved heterogeneity," *Econometrica*, 2022, 90 (2), 625–643.
- Callaway, Brantly and Tong Li**, "Quantile treatment effects in difference in differences models with panel data," *Quantitative Economics*, 2019, 10 (4), 1579–1618.
- Carneiro, Pedro, Karsten T. Hansen, and James J. Heckman**, "2001 Lawrence R. Klein Lecture Estimating Distributions of Treatment Effects with an Application to the Returns to Schooling and Measurement of the Effects of Uncertainty on College Choice\*," *International Economic Review*, 2003, 44 (2), 361–422.
- Chen, Xiaohong and Xiaotong Shen**, "Sieve extremum estimates for weakly dependent data," *Econometrica*, 1998, pp. 289–314.
- Chernozhukov, Victor and Christian Hansen**, "An IV model of quantile treatment effects," *Econometrica*, 2005, 73 (1), 245–261.
- Chernozhukov, Victor and Christian Hansen**, "Instrumental quantile regression inference for structural and treatment effect models," *Journal of Econometrics*, 2006, 132 (2), 491–525.

## References II

- Cragg, John G and Stephen G Donald**, “Inferring the rank of a matrix,” *Journal of econometrics*, 1997, 76 (1-2), 223–250.
- Cunha, Flavio, James J Heckman, and Susanne M Schennach**, “Estimating the technology of cognitive and noncognitive skill formation,” *Econometrica*, 2010, 78 (3), 883–931.
- Deaner, Ben**, “Proxy controls and panel data,” 2023.
- Fan, Yanqin and Sang Soo Park**, “Sharp bounds on the distribution of treatment effects and their statistical inference,” *Econometric Theory*, 2010, 26 (3), 931–951.
- Fan, Yanqin, Robert Sherman, and Matthew Shum**, “Identifying treatment effects under data combination,” *Econometrica*, 2014, 82 (2), 811–822.
- Firpo, Sergio and Geert Ridder**, “Partial identification of the treatment effect distribution and its functionals,” *Journal of Econometrics*, 2019, 213 (1), 210–234.
- Frandsen, Brigham R and Lars J Lefgren**, “Partial identification of the distribution of treatment effects with an application to the Knowledge is Power Program (KIPP),” *Quantitative Economics*, 2021, 12 (1), 143–171.
- Gautier, Eric and Stefan Hoderlein**, “A triangular treatment effect model with random coefficients in the selection equation,” 2015.
- Heckman, James J, Jeffrey Smith, and Nancy Clements**, “Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts,” *The Review of Economic Studies*, 1997, 64 (4), 487–535.
- Henry, Marc, Yuichi Kitamura, and Bernard Salanié**, “Partial identification of finite mixtures in econometric models,” *Quantitative Economics*, 2014, 5 (1), 123–144.

## References III

- Hu, Yingyao and Matthew Shum**, “Nonparametric identification of dynamic models with unobserved state variables,” *Journal of Econometrics*, 2012, 171 (1), 32–44.
- Hu, Yingyao and Susanne M Schennach**, “Instrumental variable treatment of nonclassical measurement error models,” *Econometrica*, 2008, 76 (1), 195–216.
- Jones, Damon, David Molitor, and Julian Reif**, “What do workplace wellness programs do? Evidence from the Illinois workplace wellness study,” *The Quarterly Journal of Economics*, 2019, 134 (4), 1747–1791.
- Kaji, Tetsuya and Jianfei Cao**, “Assessing Heterogeneity of Treatment Effects,” 2023.
- Kasahara, Hiroyuki and Katsumi Shimotsu**, “Nonparametric identification of finite mixture models of dynamic discrete choices,” *Econometrica*, 2009, 77 (1), 135–175.
- Kwon, Caleb and Eric Mbakop**, “Estimation of the number of components of nonparametric multivariate finite mixture models,” *The Annals of Statistics*, 2021, 49 (4), 2178–2205.
- Miao, Wang, Zhi Geng, and Eric J Tchetgen Tchetgen**, “Identifying causal effects with proxy variables of an unmeasured confounder,” *Biometrika*, 2018, 105 (4), 987–993.
- Nagasawa, Kenichi**, “Treatment effect estimation with noisy conditioning variables,” *arXiv preprint arXiv:1811.00667*, 2022.
- Noh, Sungho**, “Nonparametric identification and estimation of heterogeneous causal effects under conditional independence,” *Econometric Reviews*, 2023, 42 (3), 307–341.
- Shen, Xiaotong**, “On methods of sieves and penalization,” *The Annals of Statistics*, 1997, 25 (6), 2555–2591.