Distributional Treatment Effect with Latent Rank Invariance

Myungkou Shin

University of Surrey

August 26, 2024

Setup

An econometrician is interested in the ***distribution*** of treatment effect $Y(1) - Y(0)$, given a binary treatment $D \in \{0, 1\}$ and a continuous outcome variable $Y \in \mathbb{R}$:

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

Randomness of D is not enough; we cannot observe $Y(1)$ and $Y(0)$ simultaneously.

Setup

An econometrician is interested in the ***distribution*** of treatment effect $Y(1) - Y(0)$,

given a binary treatment $D \in \{0, 1\}$ and a continuous outcome variable $Y \in \mathbb{R}$:

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

Randomness of D is not enough; we cannot observe $Y(1)$ and $Y(0)$ simultaneously.

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 $Y(0) \perp\!\!\!\perp \big(Y(1) - Y(0)\big)$

  Heckman et al. (1997); Carneiro et al. (2003); Gautier and Hoderlein (2015); Noh (2023)

Setup

When estimating quantile treatment effect with endogeneous treatment,

a type of **rank invariance/similarity** is often used to extrapolate $Y_i(d)$ on $\{i : D_i = 1 - d\}$ and/or vice versa.

Chernozhukov and Hansen (2005, 2006); Athey and Imbens (2006); Vuong and Xu (2017); Callaway and Li (2019) and more

Setup

When estimating quantile treatment effect with endogeneous treatment,

a type of **rank invariance/similarity** is often used to extrapolate $Y_i(d)$ on $\{i : D_i = 1 - d\}$ and/or vice versa.

Chernozhukov and Hansen (2005, 2006); Athey and Imbens (2006); Vuong and Xu (2017); Callaway and Li (2019) and more

Rank invariance is strong enough to extrapolate the entire distribution:

$Y(1) \perp\!\!\!\perp Y(0) \mid$ rank holds and thus point identification is implied.

$Y(d) \mid$ rank is nonrandom. Want to relax rank invariance.

Setup

When estimating quantile treatment effect with endogeneous treatment,

a type of **rank invariance/similarity** is often used to extrapolate $Y_i(d)$ on $\{i : D_i = 1 - d\}$ and/or vice versa.

Chernozhukov and Hansen (2005, 2006); Athey and Imbens (2006); Vuong and Xu (2017); Callaway and Li (2019) and more

Rank invariance is strong enough to extrapolate the entire distribution:

$Y(1) \perp\!\!\!\perp Y(0) \mid$ rank holds and thus point identification is implied.

$Y(d) \mid$ rank is nonrandom. Want to relax rank invariance.

Assume a latent variable U such that $Y(1) \perp\!\!\!\perp Y(0) \mid U$.

U is not a function of $Y(1)$ or $Y(0)$ anymore; $Y(d) \mid U$ is not degenerate.

When $Y(1) \mid U$ are $Y(0) \mid U$ are identified, treatment effect distribution is identified.

Assume two proxy variables to identify $Y(d) \mid U$.

Model

An econometrican 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 $\left(Y_i(1), Y_i(0), D_i, X_i, Z_i, U_i\right) \sim iid$.

X_i and Z_i are proxy variables for U_i, used to identify $Y(d) \mid U$.

Model

An econometrican 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 $\left(Y_i(1), Y_i(0), D_i, X_i, Z_i, U_i\right) \sim iid$.

X_i and Z_i are proxy variables for U_i, used to identify $Y(d) \mid U$.

Assumption 1. $\left(Y_i(1), Y_i(0), X_i, U_i\right) \perp\!\!\!\perp D_i$.

- The treatment is random. Z_i may depend on D_i.

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

- $\mathrm{Var}(Y_i(d)|U_i) > 0$ is allowed.
- (X_i, Z_i) relate to measurement error / proxy variable literature.
  Hu and Schennach (2008); Miao et al. (2018); Deaner (2023); Nagasawa (2022) and more

Model: U_i as latent rank

A simple example: assume

$$Y_i(1) = g_1\big(U_i, \varepsilon_i(1)\big),$$
$$Y_i(0) = g_0\big(U_i, \varepsilon_i(0)\big).$$

Model: U_i as latent rank

A simple example: assume

$$Y_i(1) = g_1\big(U_i, \varepsilon_i(1)\big),$$
$$Y_i(0) = g_0\big(U_i, \varepsilon_i(0)\big).$$

A common shock U_i is drawn first.

Conditioning on U_i, treatment-specific shocks $\varepsilon_i(1)$ and $\varepsilon_i(0)$ are drawn independently.

"U_i captures all of the dependence between $Y_i(1)$ and $Y_i(0)$."

First part of Assumption 2 is satisfied.

Model: U_i as latent rank

A simple example: assume

$$Y_i(1) = g_1\left(U_i, \varepsilon_i(1)\right),$$
$$Y_i(0) = g_0\left(U_i, \varepsilon_i(0)\right).$$

A common shock U_i is drawn first.

Conditioning on U_i, treatment-specific shocks $\varepsilon_i(1)$ and $\varepsilon_i(0)$ are drawn independently.

"U_i captures all of the dependence between $Y_i(1)$ and $Y_i(0)$."

First part of Assumption 2 is satisfied.

Suppose $\mathrm{E}\left[g_1(u, \varepsilon_i(1)) | U_i = u\right]$ and $\mathrm{E}\left[g_0(u, \varepsilon_i(0)) | U_i = u\right]$ are monotone in u.

"Rank invariance holds for conditional expectation of $Y_i(d)$ given U_i."

U_i can be thought of as a 'latent' or 'interim' rank.

Model: proxy variables *(past and future outcomes)*

For X_i and Z_i, extend the cross-section model to a short panel:

$T = 3$ and $D_i = 1$ means being treated for $t = 2, 3$.

$$Y_{it}(d) = g_d\left(V_{it}, \epsilon_{it}(d)\right). \tag{2}$$

Now, the common shock V_{it} is time-dependent.

Model: proxy variables *(past and future outcomes)*

For X_i and Z_i, extend the cross-section model to a short panel:

$T = 3$ and $D_i = 1$ means being treated for $t = 2, 3$.

$$Y_{it}(d) = g_d\left(V_{it}, \epsilon_{it}(d)\right). \tag{2}$$

Now, the common shock V_{it} is time-dependent.

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

2) $\{V_{it}\}_{t=1}^{3}, \varepsilon_{i1}(0), \varepsilon_{i2}(1), \varepsilon_{i2}(0), \varepsilon_{i3}(1), \varepsilon_{i3}(0)$ are mutually independent

by letting

$$Y_i = Y_{i2}, \quad X_i = Y_{i1}, \quad Z_i = Y_{i3} \ \text{ and } \ U_i = V_{i2}.$$

Y_{it} depends on Y_{it-1} only through V_{it} depending on V_{it-1}.

Model: proxy variables *(repeated measurements)*

Suppose some error-ridden measurements of the latent variable U_i: X_i and Z_i.

Carneiro et al. (2003) discusses a similar model, but with a factor structure:

$$
\begin{aligned}
Y_i(1) &= \lambda_i^\intercal f^1 + \varepsilon_i(1) \\
Y_i(0) &= \lambda_i^\intercal f^0 + \varepsilon_i(0) \\
X_i &= \lambda_i^\intercal f^x + \varepsilon_i^x \\
Z_i &= \lambda_i^\intercal f^z + \varepsilon_i^z
\end{aligned}
$$

$Y_i(1)$, $Y_i(0)$ are potential earnings, depending on college attendance D_i.

λ_i is the latent ability of a student and (X_i, Z_i) are test scores.

Carneiro et al. (2003) assumes $\varepsilon_i(1) \perp\!\!\!\perp \varepsilon_i(0) \mid \lambda_i$ as well.

λ_i is multidimensional but a factor structure is imposed across $Y_i(1), Y_i(0), X_i$ and Z_i.

Identification

Along with some additional assumptions Assumption 3,

Assumption 2 identifies $Y(1) \mid U$ and $Y(0) \mid U$.

Firstly, split sample into two subsamples $\{i : D_i = 1\}$ and $\{i : D_i = 0\}$.

For each subsample, construct conditional density of (Y_i, X_i) given $(D_i = d, Z_i)$: $f_{Y,X|Z,d}$.

From Assumption 2,

$$f_{Y,X|Z,d}(y, x|z) = \int_{[0,1]} f_{Y(d)|U}(y|u) f_{X|U}(x|u) f_{U|Z,d}(u|z) du. \tag{3}$$

Identification

Along with some additional assumptions Assumption 3,

Assumption 2 identifies $Y(1) \mid U$ and $Y(0) \mid U$.

Firstly, split sample into two subsamples $\{i : D_i = 1\}$ and $\{i : D_i = 0\}$.

For each subsample, construct conditional density of (Y_i, X_i) given $(D_i = d, Z_i)$: $f_{Y,X|Z,d}$.

From Assumption 2,

$$f_{Y,X|Z,d}(y,x|z) = \int_{[0,1]} f_{Y(d)|U}(y|u) f_{X|U}(x|u) f_{U|Z,d}(u|z) du. \tag{3}$$

Applying Hu and Schennach (2008) more to each of the two subsamples,

Assumptions 1-3 identify the conditional densities $(f_{Y(1)|U}, f_{X|U}, f_{U|Z,1})$ and $(f_{Y(0)|U}, f_{X|U}, f_{U|Z,0})$.

The key condition is the completeness of $f_{X|Z,d}$.

Identification

Theorem 1. Assumptions 1-3 hold. The joint density $f_{Y(1),Y(0)}$ and the treatment effect distribution are identified.

$$f_{Y(1),Y(0)}(y,y') = \int_{[0,1]} f_{Y(1),Y(0)|U}(y,y'|u)du = \int_{[0,1]} f_{Y(1)|U}(y|u) \cdot f_{Y(0)|U}(y'|u)du, \tag{4}$$

$$f_{Y(1)-Y(0)}(\delta) = \int_{[0,1]} f_{Y(1)-Y(0)|U}(\delta|u)du = \int_{[0,1]} \int_{\mathbb{R}} f_{Y(1)|U}(y+\delta|u) \cdot f_{Y(0)|U}(y|u)dydu. \tag{5}$$

Identification: roles of X_i and Z_i

The key condition for X_i is $(X_i, U_i) \perp\!\!\!\perp D_i$, which implies

$$X_i | (U_i, D_i = 1) \overset{d}{\equiv} X_i | (U_i, D_i = 0). \qquad (*)$$

The two identification results are connected through $X_i | U_i$.

Identification: roles of X_i and Z_i

The key condition for X_i is $(X_i, U_i) \perp\!\!\!\perp D_i$, which implies

$$X_i | (U_i, D_i = 1) \overset{d}{\equiv} X_i | (U_i, D_i = 0). \tag{$*$}$$

The two identification results are connected through $X_i | U_i$.

An alternative sufficient condition for $(*)$, other than random treatment, is

$$(Y_i(1), Y_i(0), X_i) \perp\!\!\!\perp D_i \,\big|\, (Z_i, U_i)\,.$$

Treatment endogeneity is allowed. more

Identification: roles of X_i and Z_i

The key condition for X_i is $(X_i, U_i) \perp\!\!\!\perp D_i$, which implies

$$X_i | (U_i, D_i = 1) \overset{d}{\equiv} X_i | (U_i, D_i = 0). \qquad (*)$$

The two identification results are connected through $X_i | U_i$.

An alternative sufficient condition for $(*)$, other than random treatment, is

$$\left(Y_i(1), Y_i(0), X_i\right) \perp\!\!\!\perp D_i \,\middle|\, (Z_i, U_i)\,.$$

Treatment endogeneity is allowed. more

The key condition for Z_i is completeness of $f_{X|Z,d}$ and $f_{X|U}$.

Both $\{f_{X|Z,1}(\cdot|z)\}_{z\in\mathbb{R}}$ and $\{f_{X|Z,0}(\cdot|z)\}_{z\in\mathbb{R}}$ span the same space as $\{f_{X|U}(\cdot|u)\}_{u\in[0,1]}$.

Implementation

Recall the decomposition from Assumption 2: for $d = 0, 1$,

$$f_{Y,X|Z,d}(y, x|z) = \int_{[0,1]} f_{Y(d),X|U}(y, x|u) f_{U|Z,d}(u|z) du.$$

1. Discretization of $f_{Y,X|Z,d}$.
2. Nonnegative matrix factorization of the discretized $f_{Y,X|Z,d}$: $f_{Y(1)|U}$ and $f_{Y(0)|U}$.
3. Construct treatment effect distribution from $f_{Y(1)|U}$ and $f_{Y(0)|U}$.

Implementation: 1. discretization of $f_{Y,X|Z,d}$

With parametric assumptions on the conditional densities,

$$f_{Y,X|Z,d}(y,x|z) = \int_{[0,1]} f_{Y(d),X|U}(y,x|u) f_{U|Z,d}(u|z) du.$$

directly motivates MLE.

Parametrization defeats the purpose of flexibility of the identification result.

Instead, we consider nonparametric estimation, through the discretization of $f_{Y,X|Z,d}$.

Implementation: 1. discretization of $f_{Y,X|Z,d}$

With parametric assumptions on the conditional densities,

$$f_{Y,X|Z,d}(y,x|z) = \int_{[0,1]} f_{Y(d),X|U}(y,x|u) f_{U|Z,d}(u|z) du.$$

directly motivates MLE.

Parametrization defeats the purpose of flexibility of the identification result.

Instead, we consider nonparametric estimation, through the discretization of $f_{Y,X|Z,d}$.

Discretize $f_{Y,X|Z,d}(y,x|z)$ to a matrix, where rows correspond to (y,x) and columns correspond to z.

Straightforward when Y_i, X_i and Z_i are discrete variables.

If not, partition $\mathbb{R}$: $\mathbb{R} = \cup_{m=1}^{M_y} \mathcal{Y}^m = \cup_{m=1}^{M_x} \mathcal{X}^m = \cup_{m=1}^{M_z} \mathcal{Z}^m$ and

$$H_d = \left(\Pr \left\{ Y_i \in \mathcal{Y}^m, X_i \in \mathcal{X}^{m'} \middle| D_i = d, Z_i \in \mathcal{Z}^l \right\} \right)_{(m,m'),l}.$$

Implementation: 1. discretization of $f_{Y,X|Z,d}$

If U_i is discrete, we get the following matrix decomposition:

$$\mathrm{H}_d = \Gamma_d \cdot \Lambda_d$$

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

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

If U_i is continuous, we consider pseudo-true Γ_d and Λ_d satisfying $H_d = \Gamma_d \cdot \Lambda_d$. approximation

Such Γ_d and Λ_d may not always exist when M_z is small.

Implementation: 2. nonnegative matrix factorization

The decomposition $H_d = \Gamma_d \cdot \Lambda_d$ motivates the nonnegative matrix factorization as the estimation method.

1. Given the partition on $\mathbb{R}$: $\mathbb{R} = \cup_{m=1}^{M_y} \mathcal{Y}^m = \cup_{m=1}^{M_x} \mathcal{X}^m = \cup_{m=1}^{M_z} \mathcal{Z}^m$, construct $\mathbb{H}_0$ and $\mathbb{H}_1$, sample analogues of H_0 and H_1.

Implementation: 2. nonnegative matrix factorization

The decomposition $H_d = \Gamma_d \cdot \Lambda_d$ motivates the nonnegative matrix factorization as the estimation method.

1. Given the partition on $\mathbb{R}$: $\mathbb{R} = \cup_{m=1}^{M_y} \mathcal{Y}^m = \cup_{m=1}^{M_x} \mathcal{X}^m = \cup_{m=1}^{M_z} \mathcal{Z}^m$, construct $\mathbb{H}_0$ and $\mathbb{H}_1$, sample analogues of H_0 and H_1.
2. Fix K= the number of columns of Γ_d= the number of rows of Λ_d.

Implementation: 2. nonnegative matrix factorization

The decomposition $H_d = \Gamma_d \cdot \Lambda_d$ motivates the nonnegative matrix factorization as the estimation method.

1. Given the partition on $\mathbb{R}$: $\mathbb{R} = \cup_{m=1}^{M_y} \mathcal{Y}^m = \cup_{m=1}^{M_x} \mathcal{X}^m = \cup_{m=1}^{M_z} \mathcal{Z}^m$,
   construct $\mathbb{H}_0$ and $\mathbb{H}_1$, sample analogues of H_0 and H_1.
2. Fix K= the number of columns of Γ_d= the number of rows of Λ_d.
3. Solve the following nonnegative matrix factorization problem: algorithm

$$\left(\widehat{\Gamma}_0, \widehat{\Gamma}_1, \widehat{\Lambda}_0, \widehat{\Lambda}_1\right) = \arg\min \|\mathbb{H}_0 - \Gamma_0 \cdot \Lambda_0\|_F + \|\mathbb{H}_1 - \Gamma_1 \cdot \Lambda_1\|_F \tag{6}$$

subject to 1) $\Gamma_0, \Gamma_1, \Lambda_0, \Lambda_1$ satisfy the nonnegative, sum-to-one constraints $\cdots$ *(linear constraints)*

2) Γ_0 and Γ_1 satisfy $Y_i(d) \perp\!\!\!\perp X_i \mid U_i \cdots$ *(quadratic constraints)*

3) Γ_0 and Γ_1 imply the same marginal distribution of X_i w.r.t. $\{\mathcal{X}_m\}_{m=1}^{M_x} \cdots$ *(linear constraints)*

Implementation: 3. treatment effect distribution

Γ_d contains information on the conditional distribution of $Y_i(d)$ given U_i, but only discretely.

Want to extend Γ_d for a full density of $Y_i(d)$ given U_i.

Implementation: 3. treatment effect distribution

Γ_d contains information on the conditional distribution of $Y_i(d)$ given U_i, but only discretely.

Want to extend Γ_d for a full density of $Y_i(d)$ given U_i.

Taking a row of $H_d = \Gamma_d \cdot \Lambda_d$ to a limit, we get

$$\begin{pmatrix} f_{Y|Z,d}(y|\mathcal{Z}^1) & \cdots & f_{Y|Z,d}(y|\mathcal{Z}^{M_z}) \end{pmatrix} = \begin{pmatrix} f_{Y(d)|U}(y|u^1) & \cdots & f_{Y(d)|U}(y|u^K) \end{pmatrix} \underbrace{\begin{pmatrix} f_{U|Z,d}(u^1|\mathcal{Z}^1) & \cdots & f_{U|Z,d}(u^1|\mathcal{Z}^{M_z}) \\ \vdots & \ddots & \vdots \\ f_{U|Z,d}(u^K|\mathcal{Z}^1) & \cdots & f_{U|Z,d}(u^K|\mathcal{Z}^{M_z}) \end{pmatrix}}_{=\Lambda_d}.$$

When Λ_d is invertible,

(pseudo-true) $f_{Y(d)|U}(y|u)$ is a linear combination of $f_{Y|Z,d}(y|\mathcal{Z}^1), \cdots, f_{Y|Z,d}(y|\mathcal{Z}^{M_z})$, by multiplying $\Lambda_d{}^{-1}$

Implementation: 3. treatment effect distribution

1. Estimate $f_{Y|Z,0}$, $f_{Y|Z,1}$ with kernel estimation.

Implementation: 3. treatment effect distribution

1. Estimate $f_{Y|Z,0}$, $f_{Y|Z,1}$ with kernel estimation.
2. Estimate (pseudo-true) $f_{Y(d)|U}$ with

$$\left(\hat{f}_{Y(d)|U}(y|u^1) \quad \cdots \quad \hat{f}_{Y(d)|U}(y|u^K)\right) := \left(\hat{f}_{Y|Z,d}(y|\mathcal{Z}^1) \quad \cdots \quad \hat{f}_{Y|Z,d}(y|\mathcal{Z}^{M_{\mathbf{z}}})\right) \left(\widehat{\Lambda}_d\right)^{-1}$$

for each $d = 0, 1$.

Implementation: 3. treatment effect distribution

1. Estimate $f_{Y|Z,0}$, $f_{Y|Z,1}$ with kernel estimation.
2. Estimate (pseudo-true) $f_{Y(d)|U}$ with

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

   for each $d = 0, 1$.
3. Estimate the joint density of the potential outcomes and the marginal density of treatment effect:

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

$$\hat{f}_{Y(1)-Y(0)}(\delta) = \sum_{k=1}^{K} \int_{\mathbb{R}} \hat{f}_{Y(1)|U}(y+\delta|u^k) \cdot \hat{f}_{Y(0)|U}(y|u^k) dy \cdot \Pr\left\{\widehat{U_i = u^k}\right\}$$

   Likewise, estimate $F_{Y(1),Y(0)}$ and $F_{Y(1)-Y(0)}$ with empirical distribution functions.
   $\Pr\left\{U_i = u^k\right\}$ is estimated from the marginal distribution of Z_i and $\widehat{\Lambda}_d$.

Asymptotic theory: consistency 1

Assumption 4. U_i has a finite support: $\mathcal{U} = \{u^1, \cdots, u^K\}$.

Under Assumption 4, Γ_d and Λ_d can be thought of as 'true' distributional parameters.

Theorem 2. Under Assumptions 1-2 and 4-5, A5

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

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

Corollary 1. Under Assumptions 1-2 and 4-5,

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

$$\sup_{\delta \in \mathbb{R}} \left| \hat{F}_{Y(1) - Y(0)}(\delta) - F_{Y(1) - Y(0)}(\delta) \right| \xrightarrow{p} 0$$

as $n \to \infty$.

Asymptotic theory: consistency 2 (in development)

The nonnegative matrix factorization can be understood as a sieve GMM estimation:
the basis used in the estimation are step functions, constructed with partitions $\mathbb{R} = \cup_{m=1}^{M_y} \mathcal{Y}^m = \cdots$.

Theorem 3. Under Assumptions 1-3 and 6, A6

$$\left\| \hat{F}_{Y(d),X|U}(\cdot|u) - F_{Y(d),X|U}(\cdot|u) \right\|_2 \xrightarrow{p} 0,$$
$$\left\| \hat{F}_{U|Z,d}(\cdot|z) - F_{U|Z,d}(\cdot|z) \right\|_2 \xrightarrow{p} 0,$$

as $n \to \infty$.

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; random treatment, 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:

$$Y_i = \text{monthly medical spending over August 2016-July 2017}$$

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

$$X_i = \text{monthly medical spending over July 2015-July 2016}$$

$$Z_i = \text{monthly medical spending over August 2017-January 2019}$$

"Underlying health status U_i depends on past health status, but not on realized past medical spendings."

Empirical Illustration

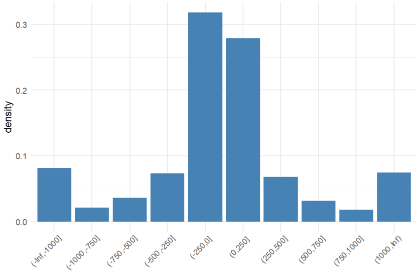

Figure 1: Marginal density of $Y_i(1) - Y_i(0)$, $K = 5$.

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

Empirical Illustration

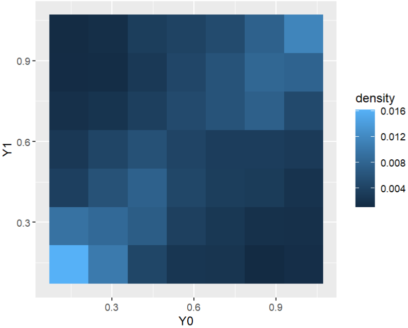

Figure 2: Joint density of $F_Y(Y_i(1))$ and $F_Y(Y_i(0))$, $K = 5$.

Higher dependence around the two ends of the spectrum.

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.

- Assume two measurements of U_i / proxy variables X_i and Z_i:

  a) $X_i | U_i, D_i = 1 \overset{d}{\equiv} X_i | U_i, D_i = 0$, which connects treated sample and untreated sample;

  b) $Z_i \perp\!\!\!\perp X_i | U_i$ and Z_i shifts U_i for a given $X_i = x$.

- A (sieve) estimation method based on nonnegative matrix factorization estimates distributional treatment effect parameters such as $F_{Y(1),Y(0)}$ or $F_{Y(1)-Y(0)}$.

Spectral Theorem of Hu and Schennach (2008)

Consider a linear operator $L_{y,X|Z,d}$ which maps a density of $Z_i|D_i = d$ to a density of $(Y_i(d) = y, X_i)$: with some density g,

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

From the decomposition based on Assumption 2, we get

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

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

Also, by integrating over y, we get $L_{X|Z,d} = L_{X|U} \cdot L_{U|Z,d}$. From Assumption 3, $L_{X|Z,d}$ exists. Thus,

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

back

Identification

Assumption 3.

a. *(bounded density)* All marginal and conditional densities of $\left(Y_i(1), Y_i(0), X_i, Z_i, U_i\right)$ are bounded.

b. *(completeness)* Let $f_{X|Z,d}$ denote the conditional density of X_i given $(D_i = d, Z_i)$.

Then, $\forall d = 0, 1$,

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

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

c. *(no repeated eigenvalue)* $\forall d = 0, 1$, $u \neq u' \Rightarrow \Pr\left\{f_{Y(d)|U}(Y_i(d)|u) \neq f_{Y(d)|U}(Y_i(d)|u')\right\} > 0$.

d. *(normalization of U_i)* $h(u) := \mathrm{E}\left[Y_i(d)|U_i = u\right]$ is monotone increasing in u and continuously differentiable.

back

Endogeneous treatment

The random treatment assumption is not crucial to identification.

Assumption 4. $(Y_i(1), Y_i(0), X_i) \perp\!\!\!\perp D_i \mid (Z_i, U_i)$.

"Z_i and U_i contain sufficient information on treatment assignment."

Let $f_{Y,X|Z,d}$ denote the conditional density of (Y_i, X_i) given $(D_i = d, Z_i)$. Then, $\forall d = 0, 1$,

$$\begin{aligned}
f_{Y,X|Z,d}(y, x|z) &= \int_{[0,1]} f_{Y(d),X|U,Z,d}(y, x|u, z) \cdot f_{U|Z,d}(u|z)du \\
&= \int_{[0,1]} f_{Y(d),X|U,Z}(y, x|u, z) \cdot f_{U|Z,d}(u|z)du \quad \because \text{Assumption 4} \\
&= \int_{[0,1]} f_{Y(d)|U}(y|u) \cdot f_{X|U}(x|u) \cdot f_{U|Z,d}(u|z)du \quad \because \text{Assumption 2}
\end{aligned}$$

U_i (and thus Z_i) should be rich enough for cond. ind. of potential outcomes AND unconfoundedness.

back

Endogeneous treatment: example

Let us go back to the nonlinear panel model with $T = 3$.

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

Recall that we set $U_i = V_{i2}$.

At $t = 2$, individuals select into treatment by solving

$$\max_d d\left(\mathrm{E}\left[Y_{i2}(1) + \beta Y_{i3}(1) | V_{i2}\right] - \eta_i\right) + (1-d)\mathrm{E}\left[Y_{i2}(0) + \beta Y_{i3}(0) | V_{i2}\right].$$

Individuals observe signal V_{i2} and a latent cost η_i, but not V_{i3} or $\{\varepsilon_{it}(d)\}_{d,t\geq 2}$.

When $\eta_i \perp\!\!\!\perp \{V_{it}, \varepsilon_{it}(d)\}_{d,t}$, Assumption 4 holds.

back

Finite mixture approximation

Note that the integral decomposition

$$f_{Y,X|Z,d}(y,x|z) = \int_{[0,1]} f_{Y(d),X|U}(y,x|u) f_{U|Z,d}(u|z) du.$$

can be thought of as a (infinite) mixture model.

Hall and Zhou (2003); Kasahara and Shimotsu (2009); Henry et al. (2014); Kedagni (2023) and more

The latent variable U_i denotes the mixture component of unit i.

$f_{Y(d),X|U}(\cdot|u)$ denotes observation density associated with mixture component u.

The proxy variable Z_i partitions the population into subpopulations.

$f_{U|Z,d}(\cdot|z)$ denotes mixture component distribution for subpopulation $\{i : Z_i = z\}$.

Z_i plays a role as an instrument in shifting mixture component distribution.

Finite mixture approximation

Note that the integral decomposition

$$f_{Y,X|Z,d}(y,x|z) = \int_{[0,1]} f_{Y(d),X|U}(y,x|u) f_{U|Z,d}(u|z) du.$$

can be thought of as a (infinite) mixture model.

Hall and Zhou (2003); Kasahara and Shimotsu (2009); Henry et al. (2014); Kedagni (2023) and more

The latent variable U_i denotes the mixture component of unit i.

$f_{Y(d),X|U}(\cdot|u)$ denotes observation density associated with mixture component u.

The proxy variable Z_i partitions the population into subpopulations.

$f_{U|Z,d}(\cdot|z)$ denotes mixture component distribution for subpopulation $\{i : Z_i = z\}$.

Z_i plays a role as an instrument in shifting mixture component distribution.

The matrix decomposition $H_d = \Gamma_d \cdot \Lambda_d$ exists

when the true model admits a good finite mixture approximation w.r.t. the given partition. back

Nonnegative matrix factorization

The objective function in (6) is quadratic once we fix either (Γ_0, Γ_1) or (Λ_0, Λ_1).

Thus, I find the (local) minima by iterating between the two: with some initial values $\left(\Gamma_0^{(0)}, \Gamma_1^{(0)}\right)$,

1. Given $\left(\Gamma_0^{(s)}, \Gamma_1^{(s)}\right)$, let $\left(\Lambda_0^{(s)}, \Lambda_1^{(s)}\right)$ be the solution to

$$\min_{\Lambda} \left\| \mathbb{H}_0 - \Gamma_0^{(s)} \Lambda_0 \right\|_F + \left\| \mathbb{H}_1 - \Gamma_1^{(s)} \Lambda_1 \right\|_F .$$

2. Given $\left(\Lambda_0^{(s)}, \Lambda_1^{(s)}\right)$, let $\left(\tilde{\Gamma}_0, \tilde{\Gamma}_1\right)$ be the solution to

$$\min_{\Gamma} \left\| \mathbb{H}_0 - \Gamma_0 \Lambda_0^{(s)} \right\|_F + \left\| \mathbb{H}_1 - \Gamma_1 \Lambda_1^{(s)} \right\|_F .$$

   Construct $\left(\Gamma_0^{(s+1)}, \Gamma_1^{(s+1)}\right)$ from marginal probabilities of $\left(\tilde{\Gamma}_0, \tilde{\Gamma}_1\right)$.

3. Iterate between **1** and **2** until convergence.

back

Assumption 5

Assumption 5.

a. Λ_0 and Λ_1 have rank K.

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

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

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

A5.a and linear independence of $\{p_k\}_k$ in **A5.b** relate to completeness;

variation in $\{q_{0k}\}_k$ and $\{q_{1k}\}_k$ in **A5.b** relate to no repeated eigenvalue.

back

Assumption 6

Assumption 6

a. The densities are in Hölder class with an exponent in $[\varepsilon, 1)$ with some $\varepsilon > 0$.

b. There exists a sequence of partitions $\left(\{\mathcal{Y}_n^m\}_{m=1}^{M_{y,n}}, \{\mathcal{X}_n^m\}_{m=1}^{M_{x,n}}, \{\mathcal{Z}_n^m\}_{m=1}^{M_{z,n}}\right)$ and corresponding parameter space for the (bounded) conditional densities, denoted by Θ_n. A norm on Θ_n is uniformly defined by $\|\theta\| = \sum_{d=0,1} \left(\int_{\mathbb{R}} \left(\int_{\mathbb{R}^2} f_{Y,X|Z,d}(y,x|z;\theta)^2 d(y,x)\right) f_{Z|d}(z) dz\right)^{\frac{1}{2}}$ where $f_{Y,X|Z,d}(y,x|z;\theta) = \int_{[0,1]} f_{Y(d),X|U}(y,x|u;\theta) f_{U|Z,d}(u|z;\theta) du$.

Then, $\{\Theta_n\}_n$ satisfies that

i. $\forall \theta \in \Theta$, there exists $\{\theta_n\}_n$ such that $\theta_n \in \Theta_n$ and $\lim_{n\to\infty} \theta_n = \theta$ w.r.t. $\|\cdot\|$.

ii. there exists some $s > 0$ such that for any $\delta > 0$,

$$\sup_{\theta,\theta' \in \Theta_n : \|\theta-\theta'\| \le \delta} \sup_{y,x,z} \left| f_{Y,X|Z,d}(y,x|z;\theta) - f_{Y,X|Z,d}(y,x|z;\theta') \right| \le \delta^s.$$

iii. $\log N\left(\delta^{\frac{1}{s}}, \Theta_n, \|\cdot\|\right) = o(n)$ for any $\delta > 0$; $N(\cdot,\cdot,\cdot)$ is the covering number.

back

References I

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

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

Carneiro, Pedro, Karsten T. Hansen, and James J. Heckman, "2001 Lawrence R. Klein Lecture Estimating Distributions of Treatment Effects with an Application to the Returns to Schooling and Measurement of the Effects of Uncertainty on College Choice*," *International Economic Review*, 2003, *44* (2), 361–422.

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

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

Deaner, Ben, "Proxy Controls and Panel Data," 2023.

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.

References II

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

Hall, Peter and Xiao-Hua Zhou, "Nonparametric estimation of component distributions in a multivariate mixture," *The annals of statistics*, 2003, *31* (1), 201–224.

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.

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.

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

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

References III

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

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

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