Distributional Treatment Effect with Latent Rank Invariance

Myungkou Shin

University of Surrey

University of Warwick econometrics seminar

November 17, 2025

Distributional treatment effect: background

Many important questions in policy learning:

"On the average, does the treatment raise the outcome?"

"Is the median outcome higher in the treated subpopulation?"

"How heterogeneous is the treatment effect at the individual level?"

"How many people would opt into treatment at cost c?"

"Should Myungkou be treated?"

Distributional treatment effect: background

Many important questions in policy learning:

"On the average, does the treatment raise the outcome?" $\Rightarrow \mathbf{E}[Y_i(1) - Y_i(0)]$

"Is the median outcome higher in the treated subpopulation?" $\Rightarrow F^{-1}_{Y(1)}(0.5) - F^{-1}_{Y(0)}(0,5)$

"How heterogeneous is the treatment effect at the individual level?" $\Rightarrow \mathrm{Var}\left(Y_i(1) - Y_i(0)\right)$

"How many people would opt into treatment at cost c*?"* $\Rightarrow \Pr\{Y_i(1) - Y_i(0) \geq c\}$

"Should Myungkou be treated?" $\Rightarrow \mathbf{1}\{Y_{ms}(1) - Y_{ms}(0) \geq 0\}$

Distributional treatment effect: background

Spectrum of treatment effect heterogeneity

- individual-level treatment effect: $\{Y_i(1) - Y_i(0)\}_{i=1}^n$
- distributional treatment effect (DTE): $\delta \mapsto \Pr\{Y_i(1) - Y_i(0) \leq \delta\}$
  $y \mapsto \Pr\{Y_i(1) \leq y | Y_i(0) = y'\}$
- summary measures:

ATT	$\mathbf{E}[Y_i(1) - Y_i(0)]$
CATE(x)	$\mathbf{E}[Y_i(1) - Y_i(0) \mid X_i = x]$
QTE(q)	$F^{-1}_{Y(1)}(q) - F^{-1}_{Y(0)}(q)$

The goal of the paper: estimate DTE.

Key challenge: how to obtain joint distribution $\Pr\{Y_i(0) \leq y, Y_i(1) \leq y'\}$
from marginals $\Pr\{Y_i(1) \leq y\}$ and $\Pr\{Y_i(0) \leq y'\}$?

Distributional treatment effect

Existing approaches

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

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

  Makarov bound; optimal transport with additional constraints

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

  Heckman et al. (1997); Carneiro et al. (2003); Wu and Perloff (2006); Noh (2023)

  Direct multiplication; deconvolution.

Distributional treatment effect

In this paper, I assume a latent variable U_i such that

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

U_i models individual-level heterogeneity and explains the dependence between $Y_i(1)$ and $Y_i(0)$.

$$\Pr\left\{Y_i(0) \leq y, Y_i(1) \leq y'\right\} = \mathbf{E}\left[\Pr\left\{Y_i(0) \leq y | U_i\right\} \cdot \Pr\left\{Y_i(1) \leq y' | U_i\right\}\right].$$

Three distributions to estimate:

1. conditional dist. of $Y_i(1)$ given U_i;
2. conditional dist. of $Y_i(0)$ given U_i;
3. marginal dist. of U_i.

To estimate the distributions, I assume two proxy variables X_i and Z_i, which shift U_i.

Role of latent heterogeneity U_i and proxy variables X_i and Z_i

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

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

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

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

- In proximal inference literature,
  $X_i =$ outcome-aligned proxy and $Z_i =$ treatment-aligned proxy.
  Miao et al. (2018); Deaner (2023); Nagasawa (2022) and more.
- Variation in $Y_i(d), X_i \,\big|\, Z_i$ comes from changes in "posterior" of U_i and vice versa for $D_i, Z_i \,\big|\, X_i$.
- Set identification for $(Y_i(d), X_i) \,\big|\, U_i$: Henry et al. (2014).

Role of latent heterogeneity U_i and proxy variables X_i and Z_i

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

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

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

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

- $Y_i(d) \perp\!\!\!\perp X_i \mid U_i$ reduces the set to a point.
  Standard assumption in nonclassical measurement error literature.
  Hu (2008); Hu and Schennach (2008) and more.

- $Y_i(1) \perp\!\!\!\perp Y_i(0) \mid U_i$ recovers the joint dist. of $Y_i(1)$ and $Y_i(0)$.
  Strong but necessary for point identification of DTE.

Role of latent heterogeneity U_i and proxy variables X_i and Z_i

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

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

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

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

- $Y_i(d) \perp\!\!\!\perp X_i \mid U_i$ reduces the set to a point.
  Standard assumption in nonclassical measurement error literature.
  Hu (2008); Hu and Schennach (2008) and more.

- $Y_i(1) \perp\!\!\!\perp Y_i(0) \mid U_i$ recovers the joint dist. of $Y_i(1)$ and $Y_i(0)$.
  Strong but necessary for point identification of DTE.

$\Rightarrow Y_i(1), Y_i(0), X_i, \big(D_i, Z_i\big)$ are mutually independent of each other given U_i.

Example 1 *(repeated measurements)*

In some empirical contexts, economic model finds us U_i and its proxy X_i and Z_i.

One example is "U_i = the innate ability of an individual,

(X_i, Z_i) = repeated measures of the innate ability."

Carneiro et al. (2003); Cunha et al. (2010); Attanasio et al. (2020) and more.

Attanasio et al. (2020): early childhood intervention's effect on children's development.

Y_i is test score at follow-up, U_i is innate ability at baseline, and (X_i, Z_i) are test scores at baseline.

$$
\begin{aligned}
Y_i(d) &= \mu^d + \alpha^d U_i + \varepsilon_i^d \quad \text{for } d = 0, 1, \\
X_i &= \mu^X + \alpha^X U_i + \varepsilon_i^X, \\
Z_i &= \mu^Z + \alpha^Z U_i + \varepsilon_i^Z.
\end{aligned}
$$

Assumptions 1-2 hold when

- $\varepsilon_i^0, \varepsilon_i^1, \varepsilon_i^X$ and ε_i^Z are mutually independent given U_i.
- D_i is randomly assigned.

Example 2 *(hidden Markov model)*

In a panel data model with Markovian latent state,

we can let "U_i = the contemporaneous latent state,

(X_i, Z_i) = past and future outcomes."

Kasahara and Shimotsu (2009); Hu and Shum (2012); Deaner (2023) and more

There are a common shock process $\{V_{it}\}_{t=1}^3$ and random shocks $\left(\varepsilon_{i1}^0, \varepsilon_{i2}^0, \varepsilon_{i2}^1, \varepsilon_{i3}^0, \varepsilon_{i3}^1\right)$.

$$Y_{it}(d) = g_d\left(V_{it}, \varepsilon_{it}^d\right) \quad \text{for } d = 0, 1 \text{ and } t = 1, 2, 3,$$

$$Y_{it} = \begin{cases} Y_{i1}(0) & \text{if } t = 1 \\ D_i \cdot Y_{it}(1) + (1 - D_i) \cdot Y_{it}(0) & \text{if } t \geq 2 \end{cases}.$$

Assumptions 1-2 hold with $\left(Y_i, X_i, Z_i, U_i\right) = \left(Y_{i2}, Y_{i1}, Y_{i3}, V_{i2}\right)$ when

- $\left(\{V_{it}\}_{t=1}^3, D_i\right), \varepsilon_{i1}^0, \varepsilon_{i2}^0, \varepsilon_{i2}^1, \varepsilon_{i3}^0, \varepsilon_{i3}^1$ are mutually independent.
- $\{V_{it}\}_{t=1}^3$ is first-order Markovian given D_i.
- D_i is randomly assigned at time $t = 2$: $\{V_{it}\}_{t=1}^2 \perp\!\!\!\perp D_i$.

Conditional independence: regime-changing treatment mechanism

In both examples,

1. Two separate outcome generating processes for $Y_i(1)$ and $Y_i(0)$: *regime-changing*.

$$Y_i(0) = \mu^0 + \alpha^0 U_i + \varepsilon_i^0$$
$$Y_i(1) = \mu^1 + \alpha^1 U_i + \varepsilon_i^1.$$

   In contrast to *input-changing* treatment mechanism. more

2. The regime-specific random shocks are purely random, satistfying $\varepsilon_i^1 \perp\!\!\!\perp \varepsilon_i^0 \mid U_i$.

   U_i must explain everything systemic.
   For less burden on U_i, everything today can be conditioning on extra control covariates.

Conditional independence: regime-changing treatment mechanism

Thus, Assumption 2 is most plausible when the treatment induces systemic changes:

- Attanasio et al. (2020):
  Treatment provided parenting guidance, changing how parents interacted with children.
- Jones et al. (2019): ← my empirical example
  Treatment provided information sessions on healthy lifestyle,
  changing how participants sought medical service and took self-care measures.
- Job assignment: e.g. the National Supported Work Demonstration.
  Treatment changes how worker skill U_i leads to outcome Y_i such as income.
- Teaching methodology: e.g. Banerjee et al. (2007); Muralidharan et al. (2019).
  Treatment changes how student aptitude U_i leads to outcome Y_i such as academic achievement.

Preview of results

Identification

1. Conditional independence framework: $Y_i(1), Y_i(0), X_i, \big(D_i, Z_i\big) \mid U_i \sim$ ind.
2. Apply diagonalization (Hu and Schennach, 2008) to untreated and treated subpopulations:

$$\left\{f_{Y(1)|U}(\cdot|u), f_{X|U}\right\}_u \qquad \text{and} \qquad \left\{f_{Y(0)|U}(\cdot|u), f_{X|U}\right\}_u$$

3. Connect the two subpopulations using $f_{X|U}(\cdot|u)$.

Estimation

finite support for $U_i \Rightarrow$ conditional independence becomes finite mixture

1. First-step: nonnegative matrix factorization.
   improved finite sample performance.
2. Second-step: plug-in GMM for DTE.
   first-step NMF as nuisance parameters.
   asymptotic normality thanks to Neyman orthogonality.

Distributional treatment effect

- Mostly focus on partial identification.

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

- A few notable point identification exceptions:

  Heckman et al. (1997); Carneiro et al. (2003); Wu and Perloff (2006); Noh (2023).

  Existing estimators use either parametric distributions or unconditional independence.

Nonclassical measurement error/proximal inference/finite mixture

- Draws from point identification result in the literature.

  Hu and Schennach (2008); Henry et al. (2014); Miao et al. (2018); Deaner (2023); Kedagni (2023); Nagasawa (2022) and more.

  Proposes a new estimator based on nonnegative matrix factorization;
  explicit nonnegativity constraint improves finite-sample performance.

Identification

Identification

The goal is to identify

$$\Pr\left\{Y_i(0) \leq y, Y_i(1) \leq y'\right\} = \mathbf{E}\left[\Pr\left\{Y_i(0) \leq y | U_i\right\} \cdot \Pr\left\{Y_i(1) \leq y' | U_i\right\}\right].$$

Identification strategy:

1. Identify the conditional distribution of $Y_i(d) \mid U_i$, within each subpopulation.
2. Identify the distribution of U_i.
3. Integrate out U_i from the conditional distribution of $\big(Y_i(1), Y_i(0)\big) \mid U_i$.

Identification: decomposition à la Hu and Schennach (2008)

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

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

When Y_i, X_i, Z_i, U_i are discrete, we get matrix representation for $d = 0, 1$,

$$\mathbf{H}_d = \begin{pmatrix} \Pr\left\{Y_i = y^1, X_i = x^1 \middle| D_i = d, Z_i = z^1\right\} & \cdots & \Pr\left\{Y_i = y^1, X_i = x^1 \middle| D_i = d, Z_i = z^{M_Z}\right\} \\ \vdots & \ddots & \vdots \\ \Pr\left\{Y_i = y^{M_Y}, X_i = x^{M_X} \middle| D_i = d, Z_i = z^1\right\} & \cdots & \Pr\left\{Y_i = y^{M_Y}, X_i = x^{M_X} \middle| D_i = d, Z_i = z^{M_Z}\right\} \end{pmatrix}$$
$$= \Gamma_d \cdot \Lambda_d$$

where $\Gamma_d = \left(\Pr\left\{Y_i(d) = y^m, X_i = x^{m'} \middle| U_i = u^k\right\} \right)_{(m,m'),k}$ *(mixture component density)*

$\Lambda_d = \left(\Pr\left\{U_i = u^k \middle| D_i = d, Z_i = z^l\right\} \right)_{k,l}$. *(mixture weight)*

Identification: decomposition à la Hu and Schennach (2008)

Recall Assumption 2 $\left(Y_i(1), Y_i(0)\right) \perp\!\!\!\perp X_i \mid U_i$.

By collecting rows of Γ_d that correspond to a specific value of y,

$$\Gamma_d(y) = \begin{pmatrix} \Pr\left\{Y_i(d) = y, X_i = x^1 \middle| U_i = u^1\right\} & \cdots & \Pr\left\{Y_i(d) = y, X_i = x^1 \middle| U_i = u^K\right\} \\ \vdots & \ddots & \vdots \\ \Pr\left\{Y_i(d) = y, X_i = x^{M_X} \middle| U_i = u^1\right\} & \cdots & \Pr\left\{Y_i(d) = y, X_i = x^{M_X} \middle| U_i = u^K\right\} \end{pmatrix}$$
$$= \Gamma_X \cdot \Delta_d(y)$$

where $\Gamma_X = \left(\Pr\left\{X_i = x^m \middle| U_i = u^k\right\}\right)_{m,k}$

$$\Delta_d(y) = \text{diag}\left(\Pr\left\{Y_i(d) = y \middle| U_i = u^1\right\}, \cdots, \Pr\left\{Y_i(d) = y \middle| U_i = u^K\right\}\right).$$

Then, $\mathbf{H}_d(y) = \Gamma_X \cdot \Delta_d(y) \cdot \Lambda_d$

$\sum_y \mathbf{H}_d(y) = \Gamma_X \cdot \Lambda_d.$

Identification: decomposition à la Hu and Schennach (2008)

When Γ_X is invertible, we get

$$\mathbf{H}_d(y)\left(\sum_{y'}\mathbf{H}_d(y')\right)^{-1} = \Gamma_X \cdot \Delta_d(y) \cdot \Lambda_d \cdot \left(\Gamma_X \cdot \Lambda_d\right)^{-1}$$
$$= \Gamma_X \cdot \Delta_d(y) \cdot \left(\Gamma_X\right)^{-1}$$

Eigenvalue decomposition finds Γ_X up to sign and scale.
Γ_X should be nonnegative and its columnwise sums should be one.
Repeating this across y finds Γ_d and thus Λ_d: Hu (2008)

Hu and Schennach (2008) develops its counterpart for continuous U_i. more
Conditional densities $f_{Y(d),X|U}, f_{U|D=d,Z}$ are identified.

Identification: sketchy of proof

1. Apply Hu and Schennach (2008) to the two subpopulations: $\Gamma_0, \Gamma_1, \Lambda_0, \Lambda_1$.
   Labelings on U_i are connected using $X_i \perp\!\!\!\perp D_i \mid U_i$.
2. $\{\Delta_0(y)\}_y$ is the distribution of $Y_i(0)$ given U_i and similarly for $\{\Delta_1(y)\}_y$:

$$\{\Delta_d(y)\}_y = \left\{\text{diag}\left(f_{Y(d)|U}(y|u)\right)_u\right\}_y$$

3. Λ_0, Λ_1 and the observed distribution of (D_i, Z_i) identifies the distribution of U_i:

$$\Lambda_d = \left(f_{U|D,Z}(u|d,z)\right)_{u,z}$$
$$f_U(u) = \mathbf{E}\left[f_{U|D,Z}(u|D_i, Z_i)\right].$$

$\Rightarrow$ $f_{Y(1)|U}, f_{Y(0)|U}, f_U$ are identified.
Assumption 2 $Y_i(1) \perp\!\!\!\perp Y_i(0) \mid U_i$ completes the proof.

Assumption 3/4. full rank/completeness of $f_{X|Z}$ when U_i is discrete/continuous: A3 A4

"Both of the proxy variables are informative for the latent variable U_i."

In the case of continuous U_i,

Assumption 5. $\mathbf{E}[Y_i(1)|U_i = u]$ or $\mathbf{E}[Y_i(0)|U_i = u]$ is 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."

$U_i =$ 'latent' or 'interim' rank.

Motivated from quantile treatment effect/IV literature:

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

Identification: identifying assumption—latent rank

Goal: to connect $\mathbf{H}_0 = \Gamma_0 \cdot \Lambda_0$ and $\mathbf{H}_1 = \Gamma_1 \cdot \Lambda_1$

When U_i is discrete, $X_i \perp\!\!\!\perp D_i \mid U_i$ and Γ_X being invertible is sufficient;
Any pair of permutations (π_0, π_1) s.t.

$$X_i \mid (D_i, U_i) = \left(0, u^{\pi_0(k)}\right) \stackrel{d}{\equiv} X_i \mid (D_i, U_i) = \left(1, u^{\pi_1(k)}\right) \qquad \forall k.$$

are equivalent.

When U_i is continuous, need to order the infinite collection $\{f_{Y(1)|U}(\cdot|u), f_{Y(0|U)}(\cdot|u), f_{X|U}(\cdot|u)\}$.
Why? Now we are in density territory.
Assumption 5 orders them using $\tilde{u} = \mathbf{E}[Y_i(d)|U_i = u]$.

Theorem 1.

Assumptions 1-3 or Assumptions 1-2, 4-5 hold.

Then, the distribution of $\left(Y_i(1), Y_i(0), D_i, X_i, Z_i\right)$ is identified.

For example:

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

is identified.

Identification: multidimensional U_i

So far, all of Y_i, X_i, Z_i, U_i are scalar random variables.
Identification holds the same with $U_i \in \mathbb{R}^p$ with $p > 1$.
Assumption 3 implies X_i, Z_i are at least p-dimensional.

Skill formation/human capital accumulation literature often model two-dimensional U_i:
Carneiro et al. (2003); Cunha et al. (2010); Attanasio et al. (2020) and more

- $U_i = \left(U_i^C, U_i^N\right)$: cognitive and noncognitive ability of a child.
- $X_i = \left(X_i^C, X_i^N\right)$: cognitive ability test scores and noncognitive ability test scores.
- $Z_i = \left(Z_i^C, Z_i^N\right)$: another set of ability scores, measured independently.

Components of X_i, Z_i need not match components of U_i.

Helps in assuming that any remaining heterogeneity after controlling for U_i is purely random

Implementation

The estimation strategy is two-step:

Step 0. Discretize X_i, Z_i and U_i.

- $\mathbf{H}_d = \Gamma_d \cdot \Lambda_d$ is exact.

Step 1. Estimate $f_{U|D=d,Z}$.

- Decompose $\mathbf{H}_d = \Gamma_d \cdot \Lambda_d$ for $d = 0, 1$ using nonnegative matrix factorization.

Step 2. Plug-in GMM to estimate DTE.

- DTE parameters are identified from quadratic moments of (Y_i, D_i, Z_i), with Λ_d as nuisance parameters.

Implementation: finite support assumption

For continuous U_i, sieve MLE and semiparametric estimation theory:

Shen (1997); Chen and Shen (1998); Ai and Chen (2003) and more.

Need strong assumptions; e.g. bounded support of Y_i and X_i.

Why? DTE parameters are complex nonlinear functionals of conditional densities. sieve

Instead, I assume $U_i \in \{u^1, \cdots, u^K\}$ with $K < \infty$. choice of K

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 parameter becomes linear in observable quantities;
   a limiting distribution is derived from U stat. theory and Neyman orthogonality.

Implementation: premise

With $U_i \in \{u^1, \cdots, u^K\}$,

$$\begin{aligned}
&\begin{pmatrix} F_{Y|D=d,Z}(y|z^1) & \cdots & F_{Y|D=d,Z}(y|z^K) \end{pmatrix} \\
&\quad = \begin{pmatrix} F_{Y(d)|U}(y|u^1) & \cdots & F_{Y(d)|U}(y|u^K) \end{pmatrix} \\
&\qquad\qquad \underbrace{\cdot \begin{pmatrix} \Pr\{U_i = u^1 | D_i = d, Z_i = z^1\} & \cdots & \Pr\{U_i = u^1 | D_i = d, Z_i = z^K\} \\ \vdots & \ddots & \vdots \\ \Pr\{U_i = u^K | D_i = d, Z_i = z^1\} & \cdots & \Pr\{U_i = u^K | D_i = d, Z_i = z^K\} \end{pmatrix}}_{=\Lambda_d} \cdot
\end{aligned}$$

With $\tilde{\Lambda}_d = {\Lambda_d}^{-1}$,

$$F_{Y(d)|U}(y|u^k) = \sum_{j=1}^{K} \tilde{\lambda}_{jk} F_{Y|D=d,Z}(y|z^j)$$

Implementation: nonnegative matrix factorization

With a known K, use a K-way partition for Z_i and construct

$$\mathbf{H}_d = \left(\Pr \left\{ Y_i = y^m, X_i = x^{m'} \middle| D_i = d, Z_i = z^k \right\} \right)_{(m,m'),k}.$$

$\mathbf{H}_d$ has K columns.

Estimate $\mathbf{H}_d$ with sample analogue:

$$\mathbb{H}_d = \left(\frac{\sum_{i=1}^n \mathbf{1}\{Y_i = y^m, D_i = d, X_i = x^{m'}, Z_i = z^k\}}{\sum_{i=1}^n \mathbf{1}\{D_i = d, Z_i = z^k\}} \right)_{(m,m'),k}$$

- Using a larger K = finer partition for Z_i has tradeoff.
- Using an overlapping partition for Z_i in general can't help;
  overlaps in the columns of Λ_d and thus Λ_d becomes more singular.

Implementation: nonnegative matrix factorization

Solve the following nonnegative matrix factorization problem:

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

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

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

This optimization is principal component analysis + additional constraint v. PCA

The objective becomes quadratic once we fix (Γ_0, Γ_1) or (Λ_0, Λ_1).

The quadratic constraint becomes linear once we fix Γ_X or $(\{\Delta_0(y)\}_y, \{\Delta_1(y)\}_y)$.

(1) is solved iteratively. algorithm

Theorem 2. Assumptions 1-3 hold. Up to some permutation on $\{u^1, \cdots, u^K\}$,

$$\left\|\widehat{\Lambda}_0 - \Lambda_0\right\|_F = O_p\left(\frac{1}{\sqrt{n}}\right) \quad \text{and} \quad \left\|\widehat{\Lambda}_1 - \Lambda_1\right\|_F = O_p\left(\frac{1}{\sqrt{n}}\right)$$

as $n \to \infty$.

The convergence rate is $n^{-\frac{1}{2}}$.
A direct corollary is that $\left(\widehat{\Lambda}_d\right)^{-1}$ is consistent for $(\Lambda_d)^{-1}$ at the same rate.

Asymptotic normality is likely for $\sqrt{n}\left(\text{vec}\left(\widehat{\Lambda}_d\right) - \text{vec}\left(\Lambda_d\right)\right)$, but out of scope here.

Implementation: comparison to diagonalization

(E) for eigenvalue decomposition more, (N) for nonnegative matrix factorization.

- Independence between $Y_i(d)$ and X_i given U_i.

  (E) imposed by diagonal representation $\Gamma_X \cdot \Delta_d(y) \cdot \left(\Gamma_X\right)^{-1}$
  (N) imposed as quadratic constraints.

- Independence between X_i and D_i given U_i.

  (E) not imposed. **(N)** imposed as linear constraints.

- $f_{Y(d),X|U}, f_{U|D=d,Z}$ are nonnegative.

  (E) not imposed. **(N)** imposed as linear constraints.

- $f_{Y(d),X|U}, f_{U|D=d,Z}$ sum to one.

  (E) imposed by rescaling eigenvectors, w.r.t. Γ_X but not imposed for eigenvalues, w.r.t. $\Delta_d(y)$
  (N) imposed as linear constraints.

Implementation: plug-in GMM

$\tilde{\lambda}_{lk,d}$ is l-th row k-th column component of $\tilde{\Lambda}_d := (\Lambda_d)^{-1}$.

$$
\begin{aligned}
&F_{Y(1)-Y(0)}(\delta) \\
&\quad = \mathbf{E}\left[\Pr\left\{Y_i(1) \le Y_i(0) + \delta | U_i\right\}\right] \\
&\quad = \sum_{k=1}^{K} p_U(k) \underbrace{\int_{\mathbb{R}}\int_{\mathbb{R}} \mathbf{1}\{y \le y' + \delta\} f_{Y(1)|U}(y|u^k) f_{Y(0)|U}(y'|u^k) dy dy'}_{=\Pr\left\{Y_i(1) \le Y_i(0)+\delta | U_i = u^k\right\}} \quad \because Y_i(1) \perp\!\!\!\perp Y_i(0) \,\middle|\, U_i \\
&\quad = \sum_{k=1}^{K} p_U(k) \int_{\mathbb{R}}\int_{\mathbb{R}} \mathbf{1}\{y \le y' + \delta\} \left(\sum_{j=1}^{K} \tilde{\lambda}_{jk,1} f_{Y|D=1,Z}(y|z^j)\right) \quad \because \text{multiplying } \tilde{\Lambda}_d \text{ to } \mathbf{H}_d = \Gamma_d \cdot \Lambda_d \\
&\qquad\qquad \left(\sum_{j'=1}^{K} \tilde{\lambda}_{j'k,0} f_{Y|D=0,Z}(y'|z^{j'})\right) dy dy' \\
&\quad = \sum_{k=1}^{K}\sum_{j=1}^{K}\sum_{j'=1}^{K} p_U(u^k) \tilde{\lambda}_{jk,1} \tilde{\lambda}_{j'k,0} \int_{\mathbb{R}}\int_{\mathbb{R}} \mathbf{1}\{y \le y' + \delta\} f_{Y|D=1,Z}(y|z^j) f_{Y|D=0,Z}(y'|z^{j'}) dy dy'
\end{aligned}
$$

Then, quadratic moments identify DTE:

$$F_{Y(1)-Y(0)}(\delta) = \sum_{k,j,j'=1}^{K} w_{kjj'} \cdot \mathbf{E}\big[\mathbf{1}\{Y_i \le Y_{i'} + \delta, D_i = 1, Z_i = z^j, D_{i'} = 0, Z_{i'} = z^{j'}\}\big]$$

$$F_{Y(1),Y(0)}(y, y') = \sum_{k,j,j'=1}^{K} w_{kjj'} \cdot \mathbf{E}\big[\mathbf{1}\{Y_i \le y, D_i = 1, Z_i = z^j, Y_{i'} \le y', D_{i'} = 0, Z_{i'} = z^{j'}\}\big]$$

for all $(y, y') \in \mathbb{R}^2$ and $\delta \in \mathbb{R}$, with $(Y_i, D_i, Z_i) \perp\!\!\!\perp (Y_{i'}, D_{i'}, Z_{i'})$ and

$$w_{kjj'} = \frac{p_U(k)\tilde{\lambda}_{jk,1}\tilde{\lambda}_{j'k,0}}{p_{D,Z}(1,j)p_{D,Z}(0,j')}$$

where $p_U(k) = \Pr\{U_i = u^k\}$ and $p_{D,Z}(d,j) = \Pr\{D_i = d, Z_i = z^j\}$.

Implementation: plug-in GMM

Nuisance parameter estimation for $w_{kjj'} = \dfrac{p_U(k)\tilde{\lambda}_{jk,1}\tilde{\lambda}_{j'k,0}}{p_{D,Z}(1,j)p_{D,Z}(0,j')}$:

- For $\tilde{\Lambda}_0, \tilde{\Lambda}_1$, use $\widehat{\tilde{\Lambda}}_d = \left(\widehat{\Lambda}_d\right)^{-1}$.
- For $p_{D,Z}(d,j)$, use $\hat{p}_{D,Z}(d,k) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{D_i = d, Z_i = z^k\}$.
- For $p_U(k)$, use $p_U(k) = \mathbf{E}[p_{U|D,Z}(k|D_i, Z_i)]$ and

$$\begin{pmatrix} \hat{p}_U(1) \\ \vdots \\ \hat{p}_U(K) \end{pmatrix} = \widehat{\Lambda}_0 \begin{pmatrix} \hat{p}_{D,Z}(0,1) \\ \vdots \\ \hat{p}_{D,Z}(0,K) \end{pmatrix} + \widehat{\Lambda}_1 \begin{pmatrix} \hat{p}_{D,Z}(1,1) \\ \vdots \\ \hat{p}_{D,Z}(1,K) \end{pmatrix}$$

First-order impact of nuisance parameter $\Rightarrow$ orthogonalization. Neyman

Implementation: plug-in GMM

The DTE estimators are plug-in GMM estimator from the orthogonalized moment:

$$\widehat{F}_{Y(1)-Y(0)}(\delta) = \sum_{k,j,j'=1}^{K} \widehat{w}_{kjj'} \cdot \binom{n}{2}^{-1} \sum_{i \neq j} \left(\frac{1}{2} \mathbf{1}\{Y_i \leq Y_{i'} + \delta, D_i = 1, Z_i = z^j, D_{i'} = 0, Z_{i'} = z^{j'}\} \right)$$
$$+ \text{orthogonalization term.}$$

Theorem 3. Assumptions 1-3 hold. Then, as $n \to \infty$,

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

Implementation: falsification test

Firstly, we can test $X_i \perp\!\!\!\perp D_i \mid U_i$ from Assumption 1 as a falsification test: more

$$\sum_{k=1}^{K}\sum_{m=1}^{M_X}\left(f_{X|D=1,U}\left(x^m|u^k\right)-f_{X|D=0,U}\left(x^m|u^k\right)\right)^2=0.$$

Alternatively, we can test $D_i \perp\!\!\!\perp U_i$ when D_i is randomly assigned:

$$\sum_{k=1}^{K}\left(p_{U|D=1}\left(u^k\right)-p_{U|D=0}\left(u^k\right)\right)^2=0.$$

Theorem 4. Under Assumptions 1-3, $T_n^1 \xrightarrow{d} \chi^2(K \cdot M_X)$ as $n \to \infty$.
Additionally, when $D_i \perp\!\!\!\perp U_i$, $T_n^2 \xrightarrow{d} \chi^2(K)$ as $n \to \infty$.

Simulation

Simulation

Monte Carlo simulations ($B = 1000$) with a simple DGP where $X_i, Z_i, U_i \in \{1, 2, 3\}$ and

$$Y_i(d) \mid U_i = k \sim \mathcal{N}\left(\mu^k(d), \sigma^k(d)^2\right).$$

Informativeness of the two proxy variables:

$$\Gamma_X = \Big(\Pr\{X_i = x | U_i = k\}\Big)_{x,k} = \begin{pmatrix} 0.911 & 0.050 & 0.022 \\ 0.067 & 0.900 & 0.067 \\ 0.022 & 0.050 & 0.911 \end{pmatrix}$$

$$\Lambda = \Big(\Pr\{U_i = k | Z_i = z\}\Big)_{z,k} = \begin{pmatrix} 0.712 & 0.195 & 0.066 \\ 0.181 & 0.544 & 0.149 \\ 0.107 & 0.260 & 0.784 \end{pmatrix}, \quad \begin{pmatrix} 0.920 & 0.063 & 0.017 \\ 0.050 & 0.853 & 0.039 \\ 0.030 & 0.084 & 0.944 \end{pmatrix}.$$

The smallest singular values for Λ is 0.377 and 0.806. specifics

Simulation

	true value	bias				rMSE			
$\widehat{F}_{Y(1)-Y(0)}(0)$	0.084	0.000	0.000	-0.002	-0.001	0.014	0.009	0.011	0.007
$\widehat{F}_{Y(1)-Y(0)}(1)$	0.264	0.001	0.001	-0.001	-0.001	0.023	0.015	0.019	0.012
$\widehat{F}_{Y(1)-Y(0)}(2)$	0.536	0.001	0.000	0.000	-0.001	0.025	0.016	0.022	0.014
$\widehat{F}_{Y(1)-Y(0)}(3)$	0.775	0.002	0.000	0.002	0.000	0.020	0.012	0.018	0.011
$\widehat{F}_{Y(1)-Y(0)}(4)$	0.911	0.005	0.002	0.003	0.001	0.014	0.008	0.012	0.007
$\sigma_{\min}(\Lambda)$		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.

Estimation performance improves as Z_i gets more informative, i.e. $\sigma_{\min}(\Lambda)$ goes up.

Simulation

	true value	coverage probability			
$\widehat{F}_{Y(1)-Y(0)}(0)$	0.084	0.971	0.951	0.952	0.935
$\widehat{F}_{Y(1)-Y(0)}(1)$	0.264	0.975	0.959	0.958	0.952
$\widehat{F}_{Y(1)-Y(0)}(2)$	0.536	0.970	0.960	0.957	0.951
$\widehat{F}_{Y(1)-Y(0)}(3)$	0.775	0.962	0.959	0.943	0.951
$\widehat{F}_{Y(1)-Y(0)}(4)$	0.911	0.940	0.954	0.934	0.948
$\sigma_{\min}(\Lambda)$		0.337	0.337	0.806	0.806
n		750	2000	750	2000

Table 2: Coverage of 95% confidence interval based on NMF.

Slight conservatism when $\sigma_{\min}(\Lambda)$ is low and n is small.

Simulation

	true value	bias				rMSE			
$\widehat{F}_{Y(1)-Y(0)}(0)$	0.084	0.000	0.000	0.014	0.008	0.014	0.009	0.034	0.029
$\widehat{F}_{Y(1)-Y(0)}(1)$	0.264	0.001	0.001	0.006	0.004	0.023	0.015	0.030	0.021
$\widehat{F}_{Y(1)-Y(0)}(2)$	0.536	0.001	0.000	-0.006	-0.005	0.025	0.016	0.037	0.029
$\widehat{F}_{Y(1)-Y(0)}(3)$	0.775	0.002	0.000	-0.009	-0.007	0.020	0.012	0.040	0.032
$\widehat{F}_{Y(1)-Y(0)}(4)$	0.911	0.005	0.002	-0.006	-0.004	0.014	0.008	0.025	0.019
first-step		NMF	NMF	EVD	EVD	NMF	NMF	EVD	EVD
n		750	2000	750	2000	750	2000	750	2000

Table 3: Comparison between first-step NMF and EVD, when $\sigma_{\min}(\Lambda) = 0.337$.

For EVD, we get nonzero bias and rMSE 1.25-4.77 times larger: intensive margin.
Also, the estimation halted for 15.4-47.2% of the samples: extensive margin.

Empirical Illustration

Empirical Illustration: setup

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:

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

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

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

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

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

Empirical illustration: choice of K

1. Consider a $M_X \times 2M_Z$ matrix $\mathbf{H}_X$:

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

$\mathbf{H}_X$ pools information from $\{i : D_i = 0\}$ and $\{i : D_i = 1\}$ and should have at most rank K.
Apply eigenvalue ratio estimator and rank test.

2. With true K, estimated densities should satisfy

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

Apply falsification tests.

Empirical illustration: choice of K

Both rank test and eigenvalue ratio estimator suggest $K = 3$.

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 4: 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 5: Kleibergen-Paap rank test statistics for H_0 : rank $= K$ and their p-values

Empirical Illustration: choice of K

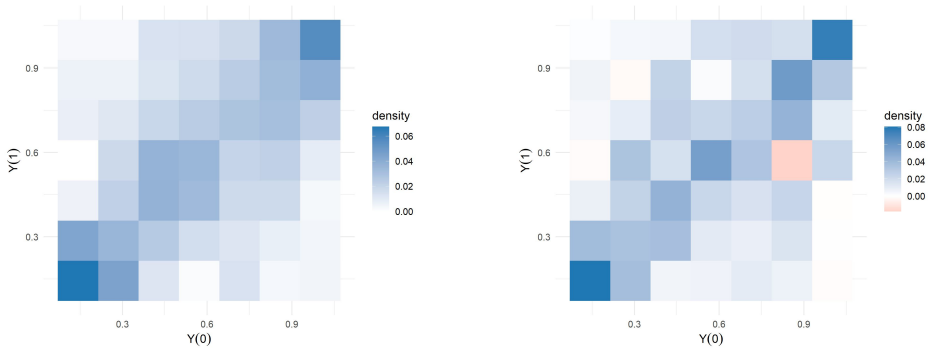
Two falsification test statistics:

$$T_n^1 = \chi^2 \text{ test statistic for } f_{X|D=1,U}(x|u) = f_{X|D=0,U}(x,u) \qquad \forall x, u,$$
$$T_n^2 = \chi^2 \text{ test statistic for } f_{U|D=1}(u) = f_{U|D=0}(u) \qquad \forall u.$$

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 6: Falsification test statistics (T_n^1, T_n^2) and their p-values

Empirical Illustration: joint density of potential outcomes



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

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

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

High correlation on the two ends of the spectrum.

Empirical Illustration: marginal distribution of treatmente effect

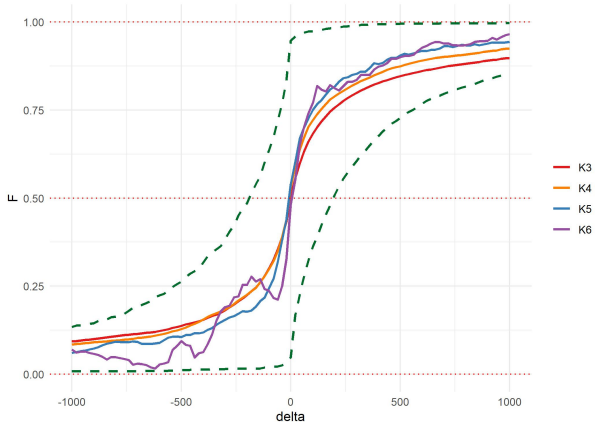


Figure 2: Marginal density of $Y(1) - Y(0)$, across K.

For 37% of the support, $\widehat{F}_{Y(1)-Y(0)}$ with $K = 6$ was decreasing.

Positive misspecification/discretization bias for on the right tail.

Empirical Illustration: treatment effect distribution

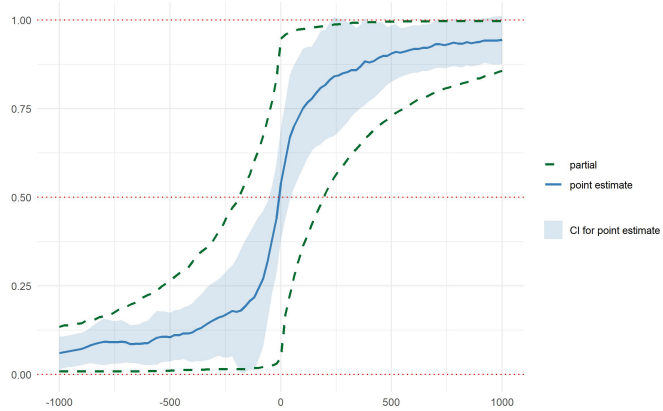


Figure 3: Marginal density of $Y(1) - Y(0)$.

Information gain from partial identification (Fan and Park, 2010).

Large negative effect is rejected while large positive effect is not.

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.

Input-changing treatment mechanism

Two common independence assumptions:

$$Y_i(1) \perp\!\!\!\perp Y_i(0) \quad \text{and} \quad Y_i(0) \perp\!\!\!\perp \big(Y_i(1) - Y_i(0)\big)$$

The latter can be motivated by *input-changing* treatment mechanism: with $V_i \perp\!\!\!\perp \varepsilon_i \mid U_i$,

$$Y_i(d) = \alpha + \mu^0 U_i + \mu^1 d V_i + \varepsilon_i$$

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

Treatment turns on a new source of individual-level heterogeneity V_i,
which is (conditionally) independent of the existing heterogeneity ε_i.

For example, providing new infrastructure: computers in teaching environment. back

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 $(Y_i(d) = y, X_i)$:

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

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,

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

back

Assumption 3

Assumption 3.

a. *(finitely discrete U_i)* $U_i \in \{u^1, \cdots, u^K\}$.

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.

a. *(continuous U_i)* $U_i \in [0,1]$.

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

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

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

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

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

back

Identification: implicit restriction

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

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

$$\mathbf{E}[Y_i(1)Y_i(0)] = \int_{\mathbb{R}} \int_{\mathbb{R}} w(y,z) \cdot y\mathbf{E}[Y_i|D_i = 1, Z_i = z]dydz.$$

$\mathbf{E}[Y_i|D_i = 1, Z_i]$ replaces $Y_i(1)$ and $w(y,z)$ replaces the joint density of $\big(Y_i(1), Y_i(0)\big)$.
"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 $\{F_{Y|D=1,Z}(\cdot|z)\}_z$."

Sieve MLE

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

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

Nonnegativity, sum-to-one, monotonicity conditions are easy 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 $\{\theta_k\}_{k=0}^m$.

back

Choice of K

Under Assumption 3, the rank of the following $M_X \times 2M_Z$ matrix is K:

$$\mathbf{H}_X = \begin{pmatrix} \Pr\left\{X_i \in \mathcal{X}^1 \middle| D_i = 0, Z_i \in \mathcal{Z}^1\right\} & \cdots & \Pr\left\{X_i \in \mathcal{X}^1 \middle| D_i = 1, Z_i \in \mathcal{Z}^{M_Z}\right\} \\ \vdots & \ddots & \vdots \\ \Pr\left\{X_i \in \mathcal{X}^{M_X} \middle| D_i = 0, Z_i \in \mathcal{Z}^1\right\} & \cdots & \Pr\left\{X_i \in \mathcal{X}^{M_X} \middle| D_i = 1, Z_i \in \mathcal{Z}^{M_Z}\right\} \end{pmatrix}$$

We can apply the Kleibergen-Paap rank test or the eigenvalue ratio test. back

Kleibergen and Paap (2006); Ahn and Horenstein (2013)

Principal component analysis vs. nonnegative matrix factorization

Principal component analysis:

- given a $M \times K$ matrix $\mathbf{H}$ and an integer $R > 0$, find a rank R matrix $\tilde{\mathbf{H}}$ such that

$$\min \left\| \mathbf{H} - \tilde{\mathbf{H}} \right\|_F$$

Nonnegative matrix factorization:

- given a $M \times K$ matrix $\mathbf{H}$ and an integer $R > 0$, find rank R ***nonnegative*** matrices Γ, Λ such that

$$\min \|\mathbf{H} - \Gamma\Lambda\|_F$$

NMF adds one more constraint: the low-rank representation should factor into nonnegative matrices.

back

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 $\left(\Gamma_0^{(s)}, \Gamma_1^{(s)}\right)$, update (Λ_0, Λ_1).
2. Given $\left(\Lambda_0^{(s+1)}, \Lambda_1^{(s+1)}, {\Gamma_{Y0}}^{(s)}, {\Gamma_{Y1}}^{(s)}\right)$, update Γ_X.
3. Given $\left(\Lambda_0^{(s+1)}, \Lambda_1^{(s+1)}, {\Gamma_X}^{(s+1)}\right)$, update $(\Gamma_{Y0}, \Gamma_{Y1})$.
4. Iterate **1**-**3** until convergence.

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

Eigenvalue decomposition estimator

Diagonalization-based estimation methods (Hu, 2008; Bonhomme et al., 2016):

1. For each d, y, construct
$$\mathbb{H}_d(y)\Big(\sum_{y'}\mathbb{H}_d(y')\Big)^{-1}$$
where $\mathbb{H}_d(y)$ estimates $\Pr\{Y_i = y, X_i = x | D_i = d, Z_i = z\}$ and $\sum_{y'}\mathbb{H}_d(y')$ estimates $\Pr\{X_i = x | D_i = d, Z_i = z\}$.
2. Diagonalize $\mathbb{H}_d(y)\Big(\sum_{y'}\mathbb{H}_d(y')\Big)^{-1}$ across d, y since
$$\mathbf{H}_d(y)\left(\sum_{y'}\mathbf{H}_d(y')\right)^{-1} = \Gamma_X \cdot \Delta_d(y) \cdot \Big(\Gamma_X\Big)^{-1}.$$
Sum-to-one will pin down eigenvectors, i.e. Γ_X.

The same $n^{-\frac{1}{2}}$ rate is established for the diagonalization estimator. back

Neyman orthogonality

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

For $\left\{\tilde{\lambda}_{lk,d}\right\}_{l,k,d}$, use the quadratic constraints of conditional independence:

$$\Pr\{Y_i = y, X_i = x | U_i = u\} = \Pr\{Y_i = y | U_i = u\} \cdot \Pr\{X_i = x | U_i = u\}$$

For $\{p_U(k)\}_k$, use the linear constraints of law of iterated expectation:

$$\Pr\{X_i = x\} = \sum_{k=1}^{k} p_U(k) \Pr\left\{X_i = x \middle| U_i = u^k\right\}.$$

For $\{p_{D,Z}(d,k)\}_{d,k}$, simply use $p_{D,Z}(d,k) = \Pr\left\{D_i = d, Z_i = z^k\right\}$. back

Neyman orthogonality

The quadratic moment is

$$\frac{1}{2}\sum_{l=1}^{K}\frac{\tilde{\lambda}_{lk,d}}{p_{D,Z}(d,l)}\cdot\mathbf{E}\left[\mathbf{1}\{Y_i=y,D_i=d,X_i=x,Z_i=z^l\}\right]$$

$$+\frac{1}{2}\sum_{m=1}^{K}\frac{\tilde{\lambda}_{mk,d}}{p_{D,Z}(d,m)}\cdot\mathbf{E}\left[\mathbf{1}\{Y_j=y,D_j=d,X_j=x,Z_i=z^m\}\right]$$

$$-\frac{1}{2}\sum_{l=1}^{K}\sum_{m=1}^{K}\frac{\tilde{\lambda}_{lk,d}\tilde{\lambda}_{mk,d}}{p_{D,Z}(d,l)\cdot p_{D,Z}(d,m)}\mathbf{E}\left[\mathbf{1}\{Y_i=y,D_i=d,Z_i=z^l,X_j=x,D_j=d,Z_j=z^m\}\right]$$

$$-\frac{1}{2}\sum_{l=1}^{K}\sum_{m=1}^{K}\frac{\tilde{\lambda}_{lk,d}\tilde{\lambda}_{mk,d}}{p_{D,Z}(d,l)\cdot p_{D,Z}(d,m)}\mathbf{E}\left[\mathbf{1}\{X_i=x,D_i=d,Z_i=z^l,Y_j=y,D_j=d,Z_j=z^m\}\right]=0$$

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

Neyman orthogonality

Let m be the score function for the DTE parameter
and ϕ be the score function for the additional moments.

The **orthogonalized score** is

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

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

Lemma 1. Assumptions 1-3 hold. Then,

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

has full rank.

Even with rich data, need to use exactly K-way partition on Z_i; just enough moments for $\tilde{\Lambda}_d$. back

Falsification test

$Y_i(1) \perp\!\!\!\perp Y_i(0) \mid U_i$ from Assumption 2 is fundamentally untestable.

Instead, I test $X_i \perp\!\!\!\perp D_i \mid U_i$ with estimators assuming $Y_i(d) \perp\!\!\!\perp X_i \mid U_i$.
"Can we construct a latent variable U_i that satisfies 1) conditional independence $Y_i(d) \perp\!\!\!\perp X_i \mid U_i$ and
2) random treatment $X_i \perp\!\!\!\perp D_i \mid U_i$?"
For this test, do not impose $X_i \perp\!\!\!\perp D_i \mid U_i$ in the NMF.

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

back

Data generating process

The specifics of the DGPs are as follows:

- $n = 750, 2000$.
- $D_i \perp\!\!\!\perp \big(Y_i(1), Y_i(0), X_i, Z_i, U_i\big)$ and $\Pr\{D_i = 1\} = 0.5$.
- $(p_U(1), p_U(2), p_U(3)) = (0.3, 0.3, 0.4)$.
- $Y_i(d) \mid U_i = k \sim \mathcal{N}\big(\mu^k(d), \sigma^k(d)^2\big)$ and

$$\Big(\mu^k(0), \sigma^k(0)\Big) = \begin{cases} (-1, 1) & \text{if } k = 1 \\ (0, 1) & \text{if } k = 2 \\ (1, 1) & \text{if } k = 3 \end{cases} \quad \text{and} \quad \Big(\mu^k(1), \sigma^k(1)\Big) = \begin{cases} (1.5, 1.5) & \text{if } k = 1 \\ (2, 1) & \text{if } k = 2 \\ (2.5, 0.5) & \text{if } k = 3 \end{cases}.$$

- Since Y_i is continuous, a three-way partition is used: $(-\infty, 0], (0, 2], (2, \infty)$.

back

Data generating process

Conditional distribution of $Y_i(1) - Y_i(0)$ given U_i:

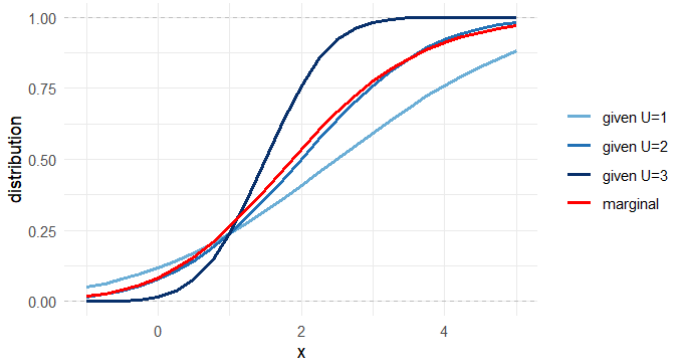


Figure 4: Marginal and conditional distributions of $Y_i(1) - Y_i(0)$.

back

References I

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

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.

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

Banerjee, Abhijit V, Shawn Cole, Esther Duflo, and Leigh Linden, "Remedying education: Evidence from two randomized experiments in India," *The quarterly journal of economics*, 2007, *122* (3), 1235–1264.

Bonhomme, Stéphane, Koen Jochmans, and Jean-Marc Robin, "Estimating multivariate latent-structure models," *The Annals of Statistics*, 2016, pp. 540–563.

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.

References II

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.

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.

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.

References III

Henry, Marc, Yuichi Kitamura, and Bernard Salanié, "Partial identification of finite mixtures in econometric models," *Quantitative Economics*, 2014, *5* (1), 123–144.

Hu, Yingyao, "Identification and estimation of nonlinear models with misclassification error using instrumental variables: A general solution," *Journal of Econometrics*, 2008, *144* (1), 27–61.

Hu, Yingyao and 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.

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

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

References IV

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.

Muralidharan, Karthik, Abhijeet Singh, and Alejandro J Ganimian, "Disrupting education? Experimental evidence on technology-aided instruction in India," *American Economic Review*, 2019, *109* (4), 1426–1460.

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.

Wu, Ximing and Jeffrey M Perloff, "Information-theoretic deconvolution approximation of treatment effect distribution," *Available at SSRN 903982*, 2006.