Distributional Treatment Effect with Latent Rank Invariance

Myungkou Shin

University of Surrey

McGill University Job Market Seminar

January 9, 2026

Distributional treatment effect

Motivation: in discussions of treatment effect heterogeneity, *distribution* of treatment effect lies at the core.

Goal: estimate distributional treatment effect parameters under intuitive and easily interpretable assumptions.

Distributional treatment effect

Motivation: in discussions of treatment effect heterogeneity,
distribution of treatment effect lies at the core.

Goal: estimate distributional treatment effect parameters
under intuitive and easily interpretable assumptions.

A potential outcome setup with a binary treatment: with $D_i \in \{0, 1\}$,

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

We only observe $Y_i(1)$ or $Y_i(0)$ for a given unit i.

Distributional treatment effect

Motivation: in discussions of treatment effect heterogeneity,
distribution of treatment effect lies at the core.

Goal: estimate distributional treatment effect parameters
under intuitive and easily interpretable assumptions.

A potential outcome setup with a binary treatment: with $D_i \in \{0, 1\}$,

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

We only observe $Y_i(1)$ or $Y_i(0)$ for a given unit i.

Existing frameworks mostly focus on some summary measures of treatment effect:

e.g.,	Average treatment effect on treated units (ATT)	$\mathbf{E}\left[Y_i(1) - Y_i(0) \mid D_i = 1\right]$
	Average treatment effect (ATE)	$\mathbf{E}\left[Y_i(1) - Y_i(0)\right]$
	Quantile treatment effect (QTE(τ))	$F^{-1}_{Y(1)}(\tau) - F^{-1}_{Y(0)}(\tau)$

Distributional treatment effect

Summary measures are invariant the dependence structure between $Y_i(1)$ and $Y_i(0)$.

Can't answer questions on treatment effect distribution or joint distribution of $Y_i(1)$ and $Y_i(0)$:

Distributional treatment effect

Summary measures are invariant the dependence structure between $Y_i(1)$ and $Y_i(0)$.

Can't answer questions on treatment effect distribution or joint distribution of $Y_i(1)$ and $Y_i(0)$:

1. Fairness concern: Levy and Markowitz (1979); Epstein and Segal (1992) and more

"Aggregate treatment effect with inequality aversion."

"I'd like to look at treatment effect heterogeneity w.r.t. baseline."

- Variance-adjusted ATE $\quad \mathbf{E}[Y_i(1) - Y_i(0)] + \frac{\gamma}{2} \mathrm{Var}\left(Y_i(1) - Y_i(0)\right)$.
- Conditional ATE given $Y_i(0) \in \mathcal{Y} \quad \mathbf{E}[Y_i(1) - Y_i(0) | Y_i(0) \in \mathcal{Y}]$.

Distributional treatment effect

Summary measures are invariant the dependence structure between $Y_i(1)$ and $Y_i(0)$.

Can't answer questions on treatment effect distribution or joint distribution of $Y_i(1)$ and $Y_i(0)$:

1. **Fairness concern:** Levy and Markowitz (1979); Epstein and Segal (1992) and more

   "Aggregate treatment effect with inequality aversion."

   "I'd like to look at treatment effect heterogeneity w.r.t. baseline."

2. **Probabilistic guarantee:** Reeve et al. (2023); Chernozhukov et al. (2025)

   "How many people are better off under treatment?"

   "What is $100 \cdot \alpha$% worst-case scenario?"

   - Share of winners $1 - F_{Y(1)-Y(0)}(0)$.
   - α-th quantile of treatment effect $F^{-1}_{Y(1)-Y(0)}(\alpha)$.

Distributional treatment effect

Summary measures are invariant the dependence structure between $Y_i(1)$ and $Y_i(0)$.
Can't answer questions on treatment effect distribution or joint distribution of $Y_i(1)$ and $Y_i(0)$:

1. Fairness concern: Levy and Markowitz (1979); Epstein and Segal (1992) and more

"Aggregate treatment effect with inequality aversion."

"I'd like to look at treatment effect heterogeneity w.r.t. baseline."

2. Probabilistic guarantee: Reeve et al. (2023); Chernozhukov et al. (2025)

"How many people are better off under treatment?"

"What is $100 \cdot \alpha$% worst-case scenario?"

3. Voluntary take-up: Heckman and Vytlacil (2005); Mogstad et al. (2018) and more

"How many people would opt into treatment at the cost of c."

"Policy-relevant treatment effect when people can opt out with full information."

- Take-up at cost c — $F_{Y(1)-Y(0)}(c)$.
- Conditional ATE given $Y_i(1) \geq Y_i(0)$ — $\mathbf{E}[Y_i(1) - Y_i(0) | Y_i(1) \geq Y_i(0)]$.

Distributional treatment effect

Existing approaches on $F_{Y(1)-Y(0)}$

- Partial identification: put bounds on $F_{Y(1)-Y(0)}(\delta)$.

  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

  $\Rightarrow$ Bounds often uninformatively large.

- 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)

  Multiplication and integration; deconvolution.

  $\Rightarrow$ Rely on restrictive independence or functional form assumptions.

Distributional treatment effect

Assume a **conditional independence** framework: Carneiro et al. (2003)

- Nonparametric identification and estimation.

Point identify and estimate **distributional treatment effect (DTE)** θ such that

$$\mathbf{E}\left[m\big(Y_i(1), Y_i(0); \theta\big)\right] = 0.$$

- Moment-identified DTE parameter θ include:
  $\mathrm{Var}\left(Y_i(1) - Y_i(0)\right)$, $F_{Y(1)-Y(0)}(\delta)$ for some δ, $F_{Y(0),Y(1)}\left(y, y'\right)$ for some $\left(y, y'\right)$ and many more.

Conditional independence

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

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

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

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

Conditional independence

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 $Y_i(0)$.

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

Once I identify **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

I identify $F_{Y(0),Y(1)}$ and thus any moment-identified DTE parameter θ.

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

Conditional independence: setup

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

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

There exists a latent variable U_i.

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

$\big(Y_i(1), Y_i(0), D_i, X_i, Z_i, U_i\big) \sim iid$.

Conditional independence: setup

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

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

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

Assumptions 1 and 2. (*conditional independence*)
$Y_i(1), Y_i(0), X_i, \big(D_i, Z_i\big)$ are mutually independent of each other given U_i.

- Only one proxy Z_i may depend on D_i given U_i.
- Can be connected to proximal inference and nonclassical measurement error literature. more
- $Y_i(1) \perp\!\!\!\perp Y_i(0) \mid U_i$ gives us key identifying power. more

Conditional independence: what is U_i and where do we find X_i and Z_i?

1. **Measurement error model**

   In some empirical contexts, natural interpretation on U_i.

   For example, D_i is early childhood intervention and Y_i is cognitive development score.

   Then, "$U_i =$ the innate ability of a child,

   $(X_i, Z_i) =$ repeated measures of the innate ability, such as test scores"

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

Conditional independence: what is U_i and where do we find X_i and Z_i?

1. **Measurement error model**

   In some empirical contexts, natural interpretation on U_i.

   For example, D_i is early childhood intervention and Y_i is cognitive development score.

   Then, "$U_i =$ the innate ability of a child,

   $(X_i, Z_i) =$ repeated measures of the innate ability, such as test scores"

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

2. **Hidden Markov model for panel data**

   A hidden Markov model for potential outcomes:

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

   and $\{V_{it}\}_t$ is first-order Markovian.

   Then, "$U_i =$ the contemporaneous common shock (V_{i2}),

   $(X_i, Z_i) =$ past and future outcomes (Y_{i1} and Y_{i3})."

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

Preview of results

Identification

1. Conditional independence framework: $Y_i(1), Y_i(0), X_i, \big(D_i, Z_i\big) \,\big|\, U_i \sim$ ind.
2. Apply diagonalization (Hu and Schennach, 2008) to untreated and treated subpopulations.
3. Connect the two decomposition results to identify $F_{Y(0),Y(1)}$.

Estimation

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

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

Literature reivew

Nonclassical measurement error/proximal inference/finite mixture

- Estimation mostly relies on diagonalization

  Hu (2008); Kasahara and Shimotsu (2009); Bonhomme et al. (2016) and more.

$\Rightarrow$ finite mixture estimator with additional regularization; better finite sample performance.
$\sqrt{n}$-consistency and orthogonalization procedure proposed.

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).

$\Rightarrow$ DTE estimator not relying on parametric distributions nor unconditional independence.
Huge information gain compared to partial bounds.

Preview of results

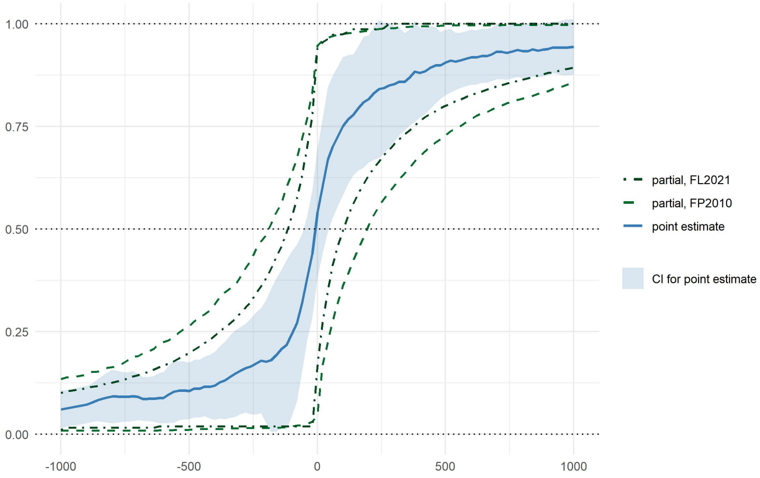


Figure 1: Marginal distribution of $Y_i(1) - Y_i(0)$.

Identification

Identification

The goal is to identify

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

Then, any DTE parameter θ s.t. $\mathbf{E}\left[m\big(Y_i(1), Y_i(0); \theta\big)\right] = 0$ is identified.

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: diagonalization à la Hu and Schennach (2008)

Given (y, d), construct

$$\mathbf{H}_d(y) = \begin{pmatrix} \Pr\left\{Y_i = y, X_i = x^1 \middle| D_i = d, Z_i = z^1\right\} & \cdots & \Pr\left\{Y_i = y, X_i = x^1 \middle| D_i = d, Z_i = z^J\right\} \\ \vdots & \ddots & \vdots \\ \Pr\left\{Y_i = y, X_i = x^{M_X} \middle| D_i = d, Z_i = z^1\right\} & \cdots & \Pr\left\{Y_i = y, X_i = x^{M_X} \middle| D_i = d, Z_i = z^J\right\} \end{pmatrix}.$$

Y_i, X_i, Z_i can be discretized when continuous.

Suppose U_i is discrete. Under Assumptions 1-2 that $Y_i(1), Y_i(0), X_i, (D_i, Z_i) \mid U_i \sim$ ind

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

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

$$\Delta_d(y) = \text{diag}\left(\Pr\left\{Y_i(d) = y \middle| U_i = u^k\right\} \right)_k$$

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

Identification: diagonalization à la Hu and Schennach (2008)

$$\sum_{y'} \mathbf{H}_d(y') = \Gamma_X \cdot \Big(\sum_{y'} \Delta_d(y') \Big) \cdot \Lambda_d = \Gamma_X \cdot \Lambda_d.$$

When $\Gamma_X \cdot \Lambda_d$ is invertible, we get

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

Eigenvalue decomposition finds $\Delta_d(y)$ and Γ_X up to sign and scale.
Repeating this across y completes identification: Hu (2008)

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

d is fixed; needs to extend Hu and Schennach (2008) to potential outcome setup.

Identification: sketchy of proof

1. Apply Hu and Schennach (2008) to the two subpopulations:

$$(\Gamma_X, \{\Delta_0(y)\}_y) \quad \text{and} \quad (\Gamma_X, \{\Delta_1(y)\}_y)\,.$$

   Labelings on U_i are connected using Γ_X since $X_i \perp\!\!\!\perp D_i \mid U_i$.

2. $\{\Delta_0(y)\}_y$ give us $f_{Y(0)|U}$ and $\{\Delta_1(y)\}_y$ give us $f_{Y(1)|U}$.
3. Given Γ_X, we get

$$\Lambda_d = \left(f_{U|D,Z}(u|d,z)\right)_{u,z} = (\Gamma_X)^+ \sum_{y'} \mathbf{H}_d(y').$$

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

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

$\Rightarrow f_{Y(1)|U}, f_{Y(0)|U}, f_U$ are identified. $Y_i(1) \perp\!\!\!\perp Y_i(0) \mid U_i$ identifies $F_{Y(0),Y(1)}$.

Identification: identifying assumptions

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

"Both of the proxy variables are fully informative about 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.

"For either $d = 0$ or 1, the latent variable U_i is the rank of $\mathbf{E}[Y_i(d)|U_i]$."

Motivated from quantile treatment effect/IV literature:

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

Rank invariance assumes a deterministic relationship between $Y_i(1)$ and $Y_i(0)$.
In my framework, only the systemic parts of $Y_i(1)$ and $Y_i(0)$ are connected through U_i:
latent rank invariance.

Assumption 5 says U_i *is* the latent rank for either $Y_i(1)$ or $Y_i(0)$.

Identification: Theorem 1

> **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.

Any moment-identified DTE parameter is trivially identified.

- Marginal distribution of treatment effect $\theta = F_{Y(1)-Y(0)}(\delta)$ as a working example.

Can be extended to multidimensional U_i. more

- The dimension of X_i and Z_i needs to be at least equal to the dimension of U_i.

Implementation

Implementation: finite support assumption

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

Discretization as approximation: Bonhomme et al. (2022) and more.

Reasoning behind the finite support assumption:

1. Conditional independence becomes finite mixture: Henry et al. (2014) and more.
2. The infeasible moment function $m\big(Y_i(1), Y_i(0); \theta\big)$ becomes linear in feasible moment functions; a plug-in GMM estimator for DTE parameters.

For continuous U_i, sieve MLE and semiparametric estimation theory:

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

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

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

Implementation: finite mixture representation

Under Assumptions 1-2 that $Y_i(1), Y_i(0), X_i, (D_i, Z_i) \mid U_i \sim$ ind,
conditional density of (Y_i, X_i) given (D_i, Z_i) admits a mixture interpretation:

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

$$f_{Y,X|D,Z}(y, x|d, z) = \int_{\mathbb{R}} \underbrace{f_{X|U}(x|u) \cdot f_{Y(d)|U}(y|u)}_{=\text{mixture component density}} \cdot \underbrace{f_{U|D,Z}(u|d, z)}_{=\text{mixture weights}} du.$$

A change in z only shifts mixture weights $f_{U|D,Z}(\cdot|d, z)$,
keeping the same mixture component densities $\{f_{Y(d),X}(\cdot, \cdot|u)\}_u$.

A finite support assumption on U_i gives us a finite mixture.
K is the number of mixture components.

Implementation: finite mixture representation

With some $M \geq K$, construct a K-way partition for Z_i and a M-way partition for (Y_i, X_i).

$$\mathbf{H}_d = \begin{pmatrix} \Pr\left\{Y_i = y^1, X_i = x^1 \middle| D_i = d, Z_i = z^1\right\} & \cdots & \Pr\left\{Y_i = y^1, X_i = x^1 \middle| D_i = d, Z_i = z^K\right\} \\ \vdots & \ddots & \vdots \\ \Pr\left\{Y_i = y^M, X_i = x^M \middle| D_i = d, Z_i = z^1\right\} & \cdots & \Pr\left\{Y_i = y^M, X_i = x^M \middle| D_i = d, Z_i = z^K\right\} \end{pmatrix}.$$

Then,

$$\underbrace{\mathbf{H}_d}_{M \times K} = \underbrace{\Gamma_d}_{M \times K} \cdot \underbrace{\Lambda_d}_{K \times K}$$

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

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

Theorem 1 says Γ_d and Λ_d are identified, up to some relabeling.

Implementation

A GMM model for DTE parameter θ with

$$\mathbf{E}\left[m\big(Y_i(1), Y_i(0); \theta\big)\right] = 0.$$

Step 0. Construct a feasible moment function $\tilde{m}$ with Λ_d as nuisance parameter.

Step 1. Estimate $\Lambda_d = \Big(f_{U|D,Z}(u|d,z)\Big)_{u,z}$.

- Estimate the finite mixture model using nonnegative matrix factorization
  $\Leftrightarrow$ decompose $\mathbf{H}_d = \Gamma_d \cdot \Lambda_d$ for $d = 0, 1$.

Step 2. Plug-in GMM to estimate θ.

Implementation, step 0: feasible GMM

Conditional independence/finite mixture $\mathbf{H}_d = \Gamma_d \cdot \Lambda_d$ implies

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

Let $\tilde{\lambda}_{jk,d}$ be the j-th row, k-th column entry of $\tilde{\Lambda}_d = \left(\Lambda_d\right)^{-1}$. Then,

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

Implementation, step 0: feasible GMM

By substituting for $f_{Y(d)|U}$,

$$
\begin{aligned}
0 &= \mathbf{E}\left[m\big(Y_i(1), Y_i(0); \theta\big)\right] \\
&= \int_{\mathbb{R}}\int_{\mathbb{R}} m\left(y', y; \theta\right) f_{Y(0),Y(1)}\left(y, y'\right) dy dy' \\
&= \int_{\mathbb{R}}\int_{\mathbb{R}} m\left(y', y; \theta\right) \left(\sum_{k=1}^{K} p_U(k) \cdot f_{Y(0)|U}\left(y|u\right) \cdot f_{Y(1)|U}\left(y'|u\right)\right) dy dy' \\
&= \int_{\mathbb{R}}\int_{\mathbb{R}} m\left(y', y; \theta\right) \left(\sum_{k=1}^{K} p_U(k) \cdot \sum_{j=1}^{K} \tilde{\lambda}_{jk,0} f_{Y|D=0,Z}\left(y|z^{j}\right) \cdot \sum_{j'=1}^{K} \tilde{\lambda}_{j'k,1} f_{Y|D=1,Z}\left(y|z^{j'}\right)\right) dy dy' \\
&= \sum_{k=1}^{K}\sum_{j=1}^{K}\sum_{j'=1}^{K} p_U(k) \tilde{\lambda}_{jk,0} \tilde{\lambda}_{j'k,1} \underbrace{\int_{\mathbb{R}}\int_{\mathbb{R}} m\left(y', y; \theta\right) \left(f_{Y|D=0,Z}\left(y|z^{j}\right) \cdot f_{Y|D=1,Z}\left(y|z^{j'}\right)\right) dy dy'}_{\text{quadratic moment of } (Y_i, D_i, X_i, Z_i)}
\end{aligned}
$$

where $p_U(k) = \Pr\left\{U_i = u^k\right\}$

Implementation, step 0: feasible GMM

For example,

$$\begin{aligned} F_{Y(1)-Y(0)}(\delta) &= \mathbf{E}\left[\mathbf{1}\{Y_i(1) \le Y_i(0) + \delta\}\right] \\ &= \sum_{k=1}^{K} \sum_{j=1}^{K} \sum_{j'=1}^{K} \frac{p_U(k)\tilde{\lambda}_{jk,1}\tilde{\lambda}_{j'k,0}}{p_{D,Z}(1,j)p_{D,Z}(0,j')} \\ &\qquad \cdot \mathbf{E}\Big[\mathbf{1}\{Y_i \le Y_{i'} + \delta, D_i = 1, Z_i = z^j, D_{i'} = 0, Z_{i'} = z^{j'}\}\Big] \end{aligned}$$

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

The nuisance parameters are $\tilde{\lambda} = \text{vec}\left((\Lambda_0)^{-1}, (\Lambda_1)^{-1}\right)$ and

$$p = \text{vec}\Big(\{p_{D,U}(d,k)\}_{d,k}, \{p_{D,Z}(d,j)\}_{d,j}\Big).$$

$\tilde{\lambda}$ is matrix inverses of (Λ_0, Λ_1).

p collects joint probabilities of (D_i, U_i) and (D_i, Z_i); also a function of (Λ_0, Λ_1).

Implementation, step 0: feasible GMM

Proposition 1.
Assumptions 1-3 hold. Suppose that a parameter of interest θ is identified by an infeasible moment condition

$$\mathbf{E}\left[m\big(Y_i(1), Y_i(0), D_i, X_i; \theta\big)\right] = 0.$$

Then, there is a moment function $\tilde{m}$ such that θ is identified by a feasible quadratic moment condition

$$\mathbf{E}\left[\tilde{m}\left(\left(Y_i, D_i, X_i, Z_i\right), \left(Y_{i'}, D_{i'}, X_{i'}, Z_{i'}\right); \theta, \tilde{\lambda}, p\right)\right] = 0$$

where $i \neq i'$.

The proxy Z_i is used to shift U_i;
for DTE parameters that involve Z_i, a similar result with higher-order moments.

Implementation, step 1: nonnegative matrix factorization

For first-step nuisance parameter estimation,
use **nonnegative matrix factorization** to estimate (Λ_0, Λ_1).

Recall $\mathbf{H}_d = \Gamma_d \cdot \Lambda_d$.
Given discretized Y_i, X_i and Z_i, estimate $\mathbf{H}_d$ with sample analogue:

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

- $\mathbb{H}_d$ is a $\sqrt{n}$-consistent estimator of $\mathbf{H}_d$.

Implementation, step 1: nonnegative matrix factorization

Solve the following nonnegative matrix factorization problem:

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

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 satisfy $D_i \perp\!\!\!\perp X_i \mid U_i$ $\cdots$ *(linear constraints)*

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

(1) is solved iteratively; higher computational burden. algorithm

Implementation, step 1: nonnegative matrix factorization

The first-step nuisance parameter estimators are:

$$\hat{\lambda} = \text{vec}\left(\left(\widehat{\Lambda}_0\right)^{-1}, \left(\widehat{\Lambda}_1\right)^{-1}\right) \qquad \text{and} \qquad \hat{p} = \text{vec}\left(\{\hat{p}_{D,U}(d,k)\}_{d,k}, \{\hat{p}_{D,Z}(d,j)\}_{d,j}\right)$$

where

$$\hat{p}_{D,Z}(d,j) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{D_i = d, Z_i = z^j\} \qquad \forall d = 0,1 \text{ and } j = 1,\cdots K$$

$$\begin{pmatrix} \hat{p}_{D,U}(d,1) \\ \vdots \\ \hat{p}_{D,U}(d,K) \end{pmatrix} = \widehat{\Lambda}_d \begin{pmatrix} \hat{p}_{D,Z}(d,1) \\ \vdots \\ \hat{p}_{D,Z}(d,K) \end{pmatrix} \qquad \forall d = 0,1$$

since $p_{D,U}(d,k) = \sum_{j=1}^{K} \Pr\left\{D_i = d, U_i = u^k, Z_i = z^j\right\}$

$$= \sum_{j=1}^{K} \underbrace{\Pr\left\{U_i = u^k | D_i = d, Z_i = z^j\right\}}_{=\text{components of the } k\text{-th row of } \Lambda_d} \cdot \Pr\left\{D_i = d, Z_i = z^j\right\}.$$

Implementation, step 1: nonnegative matrix factorization

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 $n \to \infty$.

A direct corollary is that $\hat{\lambda}$ and $\hat{p}$ are consistent for $\tilde{\lambda}$ and p at the same rate.

The same rate can be established for $\widehat{\Gamma}_0$ and $\widehat{\Gamma}_1$.

Implementation: comparison to diagonalization

Exisiting estimators for nonparametric finite mixture rely on (joint) diagonalization.
The same $n^{-\frac{1}{2}}$ rate and asymptotic normality.

Hu (2008); Kasahara and Shimotsu (2009); Bonhomme et al. (2016) and more.

$$\mathbb{H}_d(y)\left(\sum_{y'}\mathbb{H}_d(y')\right)^{-1} = \widehat{\Gamma}_X \cdot \widehat{\Delta}_d(y) \cdot \left(\widehat{\Gamma}_X\right)^{-1}.$$

Additional regularization in the NMF estimator:

- Nonnegativity constraints for $\hat{f}_{X|U}$. *(eigenvectors)*
- Sum-to-one constraints for $\hat{f}_{Y(d)|U}$. *(eigenvalues)*

DTE estimation depends on nuisance parameter estimation: $\left(\Lambda_0\right)^{-1}$ and $\left(\Lambda_1\right)^{-1}$.
Additional regularization helps.

Implementation, step 2: Neyman orthogonality

Nuisance parameter estimation has first-order impact on DTE estimation.
Thus, I orthogonalize the feasible moment function $\tilde{m}$. Delta method

$\hat{\lambda}$ and $\hat{p}$ are highly nonlinear functions of $\mathbb{H}_0$ and $\mathbb{H}_1$.
No usual "first-order condition"-type moments available.

1. For $\tilde{\lambda}$, use the quadratic moments of conditional independence: for $d = 0, 1$,

$$\Pr\{Y_i(d) = y, X_i = x | U_i = u\} = \Pr\{Y_i(d) = y | U_i = u\} \cdot \Pr\{X_i = x | U_i = u\}.$$

more

2. For $p_{D,U}$, use the moments of law of iterated expectation:

$$\Pr\{D_i = d, X_i = x\} = \sum_{k=1}^{k} p_{D,U}(d,k) \Pr\{X_i = x \middle| U_i = u^k\}.$$

Implementation, step 2: Neyman orthogonality

Let ϕ be the moment function for the additional moments.

Lemma 1. Assumptions 1-3 hold. Then, $\begin{pmatrix} \mathbf{E}\left[\frac{\partial}{\partial \tilde{\lambda}}\phi\right] \\ \mathbf{E}\left[\frac{\partial}{\partial p}\phi\right] \end{pmatrix}$ has full row rank.

With $\mu = \begin{pmatrix} \mathbf{E}\left[\frac{\partial}{\partial \tilde{\lambda}}\phi\right] \\ \mathbf{E}\left[\frac{\partial}{\partial p}\phi\right] \end{pmatrix}^{+} \begin{pmatrix} \mathbf{E}\left[\frac{\partial}{\partial \tilde{\lambda}}\tilde{m}\right] \\ \mathbf{E}\left[\frac{\partial}{\partial p}\tilde{m}\right] \end{pmatrix}$, the **orthogonalized moment function** is

$$\begin{aligned} &\psi\left((Y_i, D_i, X_i, Z_i), (Y_{i'}, D_{i'}, X_{i'}, Z_{i'}); \theta, \tilde{\lambda}, p, \mu\right) \\ &= \tilde{m}\left((Y_i, D_i, X_i, Z_i), (Y_{i'}, D_{i'}, X_{i'}, Z_{i'}); \theta, \tilde{\lambda}, p\right) - \mu^{\intercal}\phi\left((Y_i, D_i, X_i, Z_i), (Y_{i'}, D_{i'}, X_{i'}, Z_{i'}); \tilde{\lambda}, p\right) \end{aligned}$$

This applies to any moment-identified parameter in a finite mixture model.

Implementation, step 2: plug-in GMM

Using Neyman orthogonality, asympotic normality is established.

Theorem 3.
Assumptions 1-3 hold. $\hat{\theta}$ is the plug-in GMM estimator of θ, using ψ. Then,

$$\sqrt{n}\left(\hat{\theta}-\theta\right) \xrightarrow{d} \mathcal{N}\left(0, \Sigma\right)$$

as $n \to \infty$ with some consistently estimable Σ.

Inference is computationally easy;
asymptotic standard error is computed from $\psi\left(\left(Y_i, D_i, X_i, Z_i\right), \left(Y_{i'}, D_{i'}, X_{i'}, Z_{i'}\right); \hat{\theta}, \hat{\lambda}, \hat{p}, \hat{\mu}\right)$.

Wide applicability beyond DTE estimation:
random coefficient model, dynamic discrete choice and more.

Implementation: falsification test

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

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

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 $n \to \infty$. degrees of freedom

Simulation

Simulation

Monte Carlo simulations ($B = 1000$) with two DGPs where $X_i, Z_i, U_i \in \{1, 2, 3\}$.

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

and

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

$$\left(\Pr\{Z_i = z | U_i = k\}\right)_{z,k} = \begin{pmatrix} 0.689 & 0.175 & 0.078 \\ 0.233 & 0.650 & 0.233 \\ 0.078 & 0.175 & 0.689 \end{pmatrix}, \quad \begin{pmatrix} 0.911 & 0.050 & 0.022 \\ 0.067 & 0.900 & 0.067 \\ 0.022 & 0.050 & 0.911 \end{pmatrix}.$$

The smallest singular values for Λ is 0.337 and 0.806. specifics

Estimate marginal distribution of treatment effect $F_{Y(1)-Y(0)}(\delta)$.

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.

Simulation

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

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

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.

The variables in the dataset are:

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

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

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

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

"This year's underlying health status U_i only depends on last year's health status."

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 be 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 5: Eigenvalue ratios and growth ratios

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

Table 6: Kleibergen-Paap rank test statistics for H_0 : 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 7: Falsification test statistics (T_n^1, T_n^2) and their p-values

Empirical illustration: conditional ATE

	(1)	(2)	(3)	(4)	(5)
CATE($\mathcal{Y}$) (\$)	-20.43	164.17	297.49	-157.24	-732.94*
	(166.31)	(389.05)	(350.17)	(515.51)	(443.00)
$\mathcal{Y}$	$[0, 42.7]$	$(42.7, 132.2]$	$(132.2, 286.8]$	$(286.8, 671.1]$	$(671.1, \infty]$

Table 8: Conditional average treatment effect $\mathbf{E}[Y_i(1) - Y_i(0) | Y_i(0) \in \mathcal{Y}]$.

Conditional ATE across five quintiles of $Y_i(0)$:
$\mathcal{Y} = \big(F^{-1}_{Y(0)}(0), F^{-1}_{Y(0)}(1/5)\big], \cdots, \big(F^{-1}_{Y(0)}(4/5), F^{-1}_{Y(0)}(1)\big]$.

In Jones et al. (2019), p-values for ATE are 0.86-0.94. On page 1890 of Jones et al. (2019),

> *"There may exist subpopulations who did benefit from the intervention or who would have benefited had they participated."*

Empirical illustration: treatment effect distribution

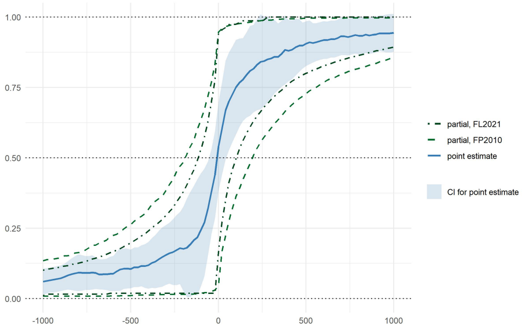


Figure 2: Marginal distribution of $Y_i(1) - Y_i(0)$.

Fan and Park (2010): $\big(Y_i(1), Y_i(0)\big) \perp\!\!\!\perp D_i$.

Frandsen and Lefgren (2021): $\big(Y_i(1), Y_i(0)\big) \perp\!\!\!\perp D_i$,

$F_{Y(1)|Y(0)}(y|y')$ is decreasing in y' for all y,

$F_{Y(0)|Y(1)}(y|y')$ is decreasing in y' for all y.

Empirical illustration: treatment effect distribution

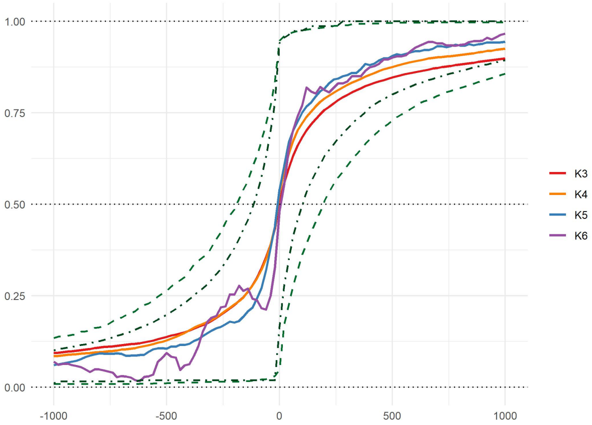
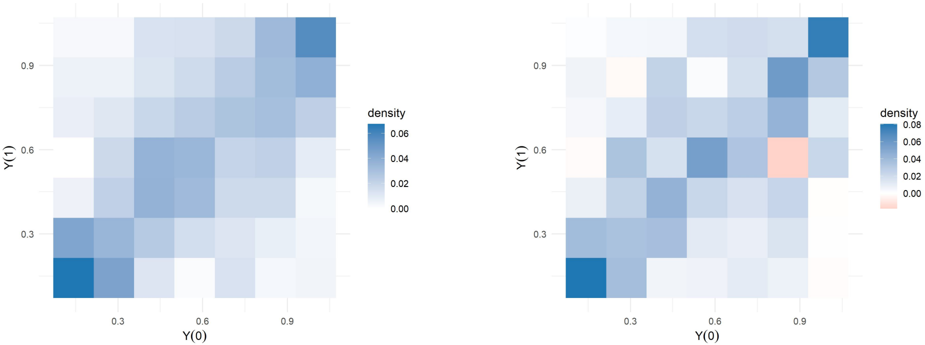


Figure 3: Marginal distribution of $Y_i(1) - Y_i(0)$, across K.

For 37% of the support, $\widehat{F}_{Y(1)-Y(0)}$ with $K = 6$ was decreasing.
Possible misspecification/discretization bias for on the right tail.

Empirical illustration: joint density of $Y_i(1)$ and $Y_i(0)$



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

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

High correlation on the two ends of the spectrum.

Conclusion

- Assume a latent variable U such that

$$Y_i(1) \perp\!\!\!\perp Y_i(0) \;\Big|\; U_i.$$

  This assumption could be thought of as a 'latent rank invariance' condition.

- Use two proxy variables X_i and Z_i to identify the distribution of $Y_i(d)|U_i$.
  Measurement error model and hidden Markov model motivate the use of proxies.

- Nonnegative matrix factorization estimates finite mixture.
  Better finite sample performance due to additional regularization.

- An asymptotic distribution is derived for the DTE estimator.

Conditional independence framework

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

- 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.

- Distribution of $(Y_i(d), X_i) \mid U_i$ is set identified: Henry et al. (2014).

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

- Nonclassical measurement error literature: $Y_i, X_i, Z_i \mid U_i \sim$ ind.
  Hu (2008); Hu and Schennach (2008) and more.

- Conditional distributions are point identified.

- Extended to a potential outcome setup.
  Additionally, $Y_i(1) \perp\!\!\!\perp Y_i(0) \mid U_i$ recovers the joint dist. of $Y_i(1)$ and $Y_i(0)$.

back

Conditional independence: example 1 *(repeated measurement)*

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.

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

Conditional independence: example 1 *(repeated measurement)*

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 $\left(X_i, Z_i\right)$ 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.

back

Conditional independence: example 2 *(past and future outcomes)*

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

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

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

Conditional independence: example 2 *(past and future outcomes)*

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: $\quad V_{i1} \perp\!\!\!\perp V_{i3} \,\big|\, (V_{i2}, D_i)$.
- D_i is randomly assigned at time $t = 2$: $\{V_{it}\}_{t=1}^{2} \perp\!\!\!\perp D_i$.

back

Regime-changing treatment effect mechanism

In both measurement error model and hidden Markov model,

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

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

The regime-specific random shocks are purely random, satistfying $\varepsilon_i^1 \perp\!\!\!\perp \varepsilon_i^0 \mid U_i$.
In contrast to *input-changing* treatment mechanism. more

back

Examples of 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): $\leftarrow$ 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.

back

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 $V_i \perp\!\!\!\perp \varepsilon_i \,\big|\, U_i$,

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

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

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

Assumptions 1-3

Assumptions 1-2.
$Y_i(1), Y_i(0), X_i, \left(D_i, Z_i\right)$ are mutually independent of each other given U_i.

Assumption 3.

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

b. *(full rank)* Γ_X, Λ_0 and Λ_1 have rank K.

c. *(no repeated eigenvalue)* For any $k \neq k'$, there exist some $d \in \{0, 1\}$ and y such that

$$\Pr\left\{Y_i(d) = y \middle| U_i = u^k\right\} \neq \Pr\left\{Y_i(d) = y \middle| U_i = u^{k'}\right\}.$$

A3 P1 T2

Assumption 4

Assumption 4.

- **a.** *(continuous U_i)* $U_i \in [0,1]$.
- **b.** *(bounded density)* All marginal and conditional densities of $\left(Y_i(1), Y_i(0), X_i, Z_i, U_i\right)$ 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

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

Why do we need Assumption 5?

Under Assumptions 1,2 and 3/4, we have identified

$$\{f_{Y(1)|U}(\cdot|u), f_{Y(0)|U)}(\cdot|u), f_{X|U}(\cdot|u), f_{U|D=1,Z}(u|\cdot), f_{U|D=0,Z}(u|\cdot)\}_u.$$

Each group of five is not yet connected to a value of u; unordered collection.

When the collection is finite, not having an ordering is okay.
Probability mass function of U_i is still identified.

When the collection is infinite, not having an ordering matters.
Why? Now we are in density territory.
$\{f_{U|D=1,Z}(u|\cdot), f_{U|D=0,Z}(u|\cdot)\}_u$ gives us $\{f_U(u)\}_u$ but not f_U.
Assumption 5 orders $\{f_U(u)\}_u$ using $\tilde{u} = \mathbf{E}[Y_i(d)|U_i = u]$. back

Multidimensional U_i

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

- $U_i = (U_i^C, U_i^N)$: cognitive and noncognitive ability of a child.
- $X_i = (X_i^C, X_i^N)$: cognitive ability test scores and noncognitive ability test scores.
- $Z_i = (Z_i^C, Z_i^N)$: 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.

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 estimator.

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

back

Derivation of feasible moment function

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

back

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

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

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

back

Nonnegative matrix factorization

Γ_d can be further decomposed into Γ_X and $\Gamma_{Y(d)}$, using $Y_i(d) \perp\!\!\!\perp X_i \mid U_i$.

The minimization problem

$$\min \|\mathbb{H}_d - \Gamma_d \cdot \Lambda_d\|_F$$

becomes a quadratic program with linear constraints,
once we fix two out of the three matrices $\Gamma_X, \Gamma_{Y(d)}, \Lambda_d$.

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

1. Given $\left(\Gamma_0^{(s)}, \Gamma_1^{(s)}\right)$, update (Λ_0, Λ_1).
2. Given $\left(\Lambda_0^{(s+1)}, \Lambda_1^{(s+1)}, {\Gamma_{Y(0)}}^{(s)}, {\Gamma_{Y(1)}}^{(s)}\right)$, update Γ_X.
3. Given $\left(\Lambda_0^{(s+1)}, \Lambda_1^{(s+1)}, {\Gamma_X}^{(s+1)}\right)$, update $\left(\Gamma_{Y(0)}, \Gamma_{Y(1)}\right)$.
4. Iterate **1**-**3** until convergence.

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

Implementation: comparison to diagonalization

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.

back

Asymptotic normality + Delta method

The first-step NMF can be thought of as a (loosely defined) GMM or MLE estimator.

1. Nonnegativity constraints often bind.

   The estimators may not be asymptotically normal.
   e.g. Bonhomme et al. (2016) derives asymptotic normality, while not imposing nonnegativity.

2. DTE parameters are highly nonlinear in Λ_0 and Λ_1.

   Asymptotic normality using Delta method may converge slowly.
   Bootstrapping for standard error $\Rightarrow$ higher computation burden, given NMF.

3. Orthogonalization may help with discretization bias when U_i is continuous.

   We would need $\widehat{\Lambda}_d$ and $\left(\widehat{\Lambda}_d\right)^{-1}$ to converge to true bivariate functions at $n^{-\frac{1}{4}}$ rate.

Hence, $\sqrt{n}$-consistency and orthogonalization. back

Additional moments in orthogonalization

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

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

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

Degrees of freedom

If testing $X_i \perp\!\!\!\perp D_i \mid Z_i$ with

$$\hat{f}_{X|D=d,Z}(x|z) = \frac{\sum_{i=1}^{n} \mathbf{1}\{X_i = x, D_i = d, Z_i = z\}}{\sum_{i=1}^{n} \mathbf{1}\{D_i = d, Z_i = z\}}$$

$T_n \xrightarrow{d} \chi^2(K \cdot (M_X - 1))$ due to linear relationship $\sum_{m=1}^{M_X} \hat{f}_{X|D=d,Z}(x^m|z) = 1$.

$\hat{f}_{X|D=1,U}(x|u)$ and $\hat{f}_{X|D=0,U}(x|u)$ uses additional moments ϕ for orthogonalization.
Not necessarily $\sum_{m=1}^{M_X} \hat{f}_{X|D=d,U}(x^m|u) = 1 \; \forall d, u$. 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:

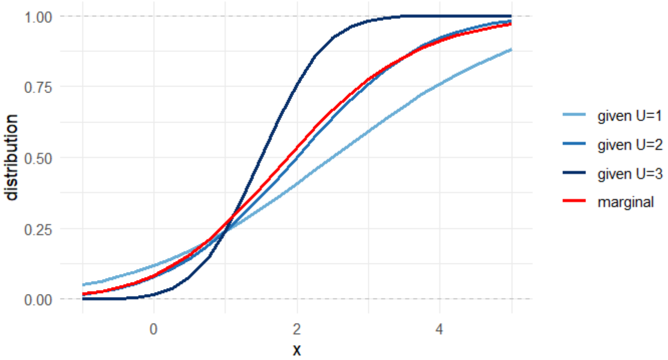


Figure 5: Marginal and conditional distributions of $Y_i(1) - Y_i(0)$.

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.

Chernozhukov, Victor, Sokbae Lee, Adam M Rosen, and Liyang Sun, "Policy Learning with Confidence," *arXiv preprint arXiv:2502.10653*, 2025.

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.

Epstein, Larry G and Uzi Segal, "Quadratic social welfare functions," *Journal of Political Economy*, 1992, *100* (4), 691–712.

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.

References III

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 and Edward Vytlacil, "Structural equations, treatment effects, and econometric policy evaluation 1," *Econometrica*, 2005, *73* (3), 669–738.

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, "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.

References IV

Kasahara, Hiroyuki and Katsumi Shimotsu, “Nonparametric identification of finite mixture models of dynamic discrete choices,” *Econometrica*, 2009, *77* (1), 135–175.

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

Levy, H and HM Markowitz, “Approximating Expected Utility by a Function of Mean and Variance,” *The American Economic Review*, 1979, *69* (3), 308–317.

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.

Mogstad, Magne, Andres Santos, and Alexander Torgovitsky, “Using instrumental variables for inference about policy relevant treatment parameters,” *Econometrica*, 2018, *86* (5), 1589–1619.

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.

Reeve, Henry WJ, Timothy I Cannings, and Richard J Samworth, “Optimal subgroup selection,” *The Annals of Statistics*, 2023, *51* (6), 2342–2365.

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.