Relaxing Instrument Exogeneity with Common Confounders

IV with Mismeasured Confounders

Christian Tien

EEA ESEM 2023

August 30, 2023

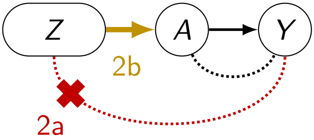
Outline

Idea

Quantity of interest: *Causal effect* of treatment A on outcome Y.	$\theta = \int Y(a)\pi(a)\,\mathrm{d}\mu_A(a)$
A is endogenous (simultaneity, unobserved confounders).	$Y(a) \not\perp\!\!\!\perp A$
We want to use relevant instruments Z for A.	$A(z) \neq A$
Instruments NOT unconditionally exogenous.	$Y(a) \not\perp\!\!\!\perp Z$
The unobserved *common confounders* U fully explain the association between Z and proxies W.	$Z \perp\!\!\!\perp W \mid U$
Instruments would be exogenous conditional on the common confounders U.	$Y(a) \perp\!\!\!\perp Z \mid U$

IV

Figure: DAG of an IV model



Assumption (IV Model)

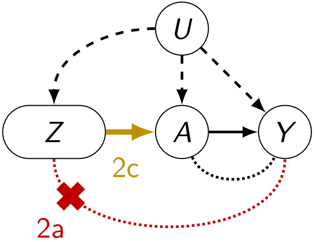
1. *SUTVA:* $Y = Y(A, Z)$

2a. *Instrument Exogeneity:*
$Y(a, z) = Y(a) \perp\!\!\!\perp Z$.

2b. *Instrument Relevance:*
For any $g(A) \in L_2(A)$,
$\mathbb{E}\left[g(A)|Z\right] = 0$ *only if* $g(A) = 0$.

Unobservable Confounders U

Figure: DAG with unobserved confounders



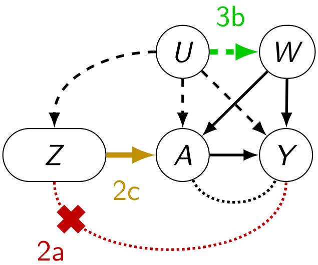
Assumption (Confounded IV)

2a. ***Cond.*** *Instrument Exogeneity:*
$Y(a,z) = Y(a) \perp\!\!\!\perp Z \mid \mathbf{U}$.

2c. ***Cond.*** *Instrument Relevance:*
For any $g(A,\mathbf{U}) \in L_2(A,\mathbf{U})$,
$\mathbb{E}\left[g(A,\mathbf{U})|Z\right] = 0$ *only if* $g(A,\mathbf{U}) = 0$.

Proxies W for Unobservables U

Figure: Introducing proxies W



Assumption (Confounded IV with relevant proxies)

2a. *Cond. Instrument Exogeneity:*
$Y(a, z) = Y(a) \perp\!\!\!\perp Z \mid U.$

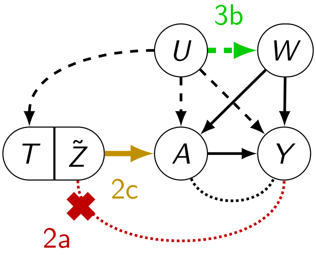
2c. *Cond. Instrument Relevance:*
For any $g(A, U) \in L_2(A, U)$,
$\mathbb{E}\left[g(A, \mathbf{U}) | Z\right] = 0$ *only if* $g(A, \mathbf{U}) = 0.$

3a. *Proxy Exogeneity:*
$W(z) = W \perp\!\!\!\perp Z \mid U.$

3b. *Proxy Relevance:*
For any $g(U) \in L_2(U)$,
$\mathbb{E}\left[g(U) | W\right] = 0$ *only if* $g(U) = 0.$

Index Sufficiency

Figure: Focus on Z



Assumption (Confounded IV with rel. proxies and index sufficiency)

2a. *Cond. Instrument Exogeneity:*
$Y(a,z) = Y(a) \perp\!\!\!\perp Z \mid U.$

2b. *Index sufficiency:* $U \perp\!\!\!\perp Z \mid T$ *for some* $T = \tau(Z)$.

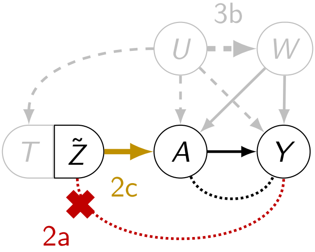
2c. *Cond. Instrument Relevance:*
For any $g(A,T) \in L_2(A,T)$,
$\mathbb{E}\left[g(A,T)|Z\right] = 0$ *only if* $g(A,T) = 0$.

3a. *Proxy Exogeneity:*
$W(z) = W \perp\!\!\!\perp Z \mid U.$

3b. *Proxy Relevance:*
For any $g(U) \in L_2(U)$,
$\mathbb{E}\left[g(U)|W\right] = 0$ *only if* $g(U) = 0$.

Block Backdoor Path

Figure: Retrieving standard IV model



Assumption 1.1

2a. *Cond. Instrument Exogeneity:* $Y(a,z) = Y(a) \perp\!\!\!\perp Z \mid U.$

2b. *Index sufficiency:* $U \perp\!\!\!\perp Z \mid T$ *for some* $T = \tau(Z)$.

2c. *Cond. Instrument Relevance:* *For any* $g(A,T) \in L_2(A,T)$, $\mathbb{E}\left[g(A,T)|Z\right] = 0$ *only if* $g(A,T) = 0.$.

3a. *Proxy Exogeneity:* $W(z) = W \perp\!\!\!\perp Z \mid U.$

3b. *Proxy Relevance:* *For any* $g(U) \in L_2(U)$, $\mathbb{E}\left[g(U)|W\right] = 0$ *only if* $g(U) = 0.$.

Related literature

- Instrumental Variables
  - with linear separability of unobservables in the outcome model [Newey and Powell, 2003]
  - for average structural function identification with strict monotonicity in the first stage reduced form [Imbens and Newey, 2009]
- Proximal learning
  - General proximal learning [Deaner, 2018, Tchetgen Tchetgen et al., 2020, Cui et al., 2020]
  - Control function approach [Nagasawa, 2018]
- Index sufficiency assumption
  - on unobserved heterogeneity for average effect identification in panel data [Liu et al., 2021]
- Semiparametric estimation
  - with nested nuisances [Chernozhukov et al., 2022]
  - with nuisances as solutions to (possibly ill-posed) inverse problems [Bennett et al., 2023]

Linear Example

Equation	Exogeneity	Relevance
$Y = A\beta + W\upsilon_Y + U\gamma_Y + \varepsilon_Y,$	$\mathbb{E}\left[\varepsilon_Y Z\right] = \mathbf{0},$	
$A = Z\zeta + W\upsilon_A + U\gamma_W + \varepsilon_A,$		$\text{rank}\left(\mathbb{E}\left[A^\intercal Z \mid T\right]\right) = d_A$
$Z = U\gamma_Z + \varepsilon_Z,$		$\text{rank}(\gamma_Z) = d_U < d_Z,$
$W = U\gamma_W + \varepsilon_W,$	$\mathbb{E}\left[\varepsilon_W^\intercal \varepsilon_Z\right] = \mathbf{0},$	$\text{rank}(\gamma_W) = d_U \leq d_W$

3SLS procedure I

1. Reduced form (rank-restricted) regression of W on Z:
   1.1 d_U known: With appropriate normalisation of δ_{WU},

$$\underset{N\times d_W}{W} = \underset{N\times d_Z}{Z}\,\underset{d_Z\times d_U}{\delta_{ZU}}\,\underset{d_U\times d_W}{\delta_{WU}^{\intercal}} + \epsilon_W \implies \underset{N\times d_U}{T} = \underset{N\times d_Z}{Z}\,\underset{d_Z\times d_U}{\hat{\delta}_{ZU}}$$

$$\left(\hat{\delta}_{ZU}, \hat{\delta}_{WU}\right) = \arg\min_{(\delta_{ZU},\delta_{WU})} \sum_{j=1}^{d_W}\sum_{i=1}^{N}\left(w_{i,j} - z_i^{\intercal}\delta_{ZU}\delta_{WU,j}^{\intercal}\right)^2$$

   If $d_W = d_U$, $\hat{\delta}_{ZU} = (Z^{\intercal}Z)^{-1} Z^{\intercal}W$ (OLS).

   1.2 d_U unknown:
   U are the set of unobserved variables explaining all correlation between W and Z.
   - If $d_U \geq \min\{d_W, d_Z\}$, then $\operatorname{rank}\left(\mathbb{E}\left[w_i z_i^{\intercal}\right]\right) = \min\{d_W, d_Z\}$.
   - If $d_U < \min\{d_W, d_Z\}$, then $d_U = \operatorname{rank}\left(\mathbb{E}\left[w_i z_i^{\intercal}\right]\right)$.

3SLS procedure II

A test for a sufficient (not necessary) condition of W's relevance for U (and necessary condition for Z's relevance for A given T) [Chen and Fang, 2019] is

$$H_0 : \text{rank}\left(\mathbb{E}\left[w_i z_i^{\intercal}\right]\right) \leq r_0 < \min\{d_W, d_Z\} \text{ vs}$$
$$H_1 : \text{rank}\left(\mathbb{E}\left[w_i z_i^{\intercal}\right]\right) > r_0.$$

Reject H_0 for $r_0 = \min\{d_W, d_Z\} - 1$:

- If $d_Z < d_W$: $d_U \geq d_Z$, so Z never relevant for A given T.
- If $d_W < d_Z$: $d_U \geq d_W$, so either W just-relevant for U ($d_W = d_U$), or W not relevant for U ($d_W < d_U$). Both are observationally equivalent.

If H_0 not rejected: Proceed with $d_U = r_0$ in step 1.1.

3SLS procedure III

2. Reduced form OLS regression of A on any subset of instruments Z_0 of dimension $N \times (d_Z - d_U)$ and T, W.

$$\underset{N\times d_A}{A} = \underset{N\times(d_Z-d_U)}{Z_0}\underset{(d_Z-d_U)\times d_A}{\delta_{Z_0A}} + \underset{N\times d_U}{T}\underset{d_U\times d_A}{\delta_{TA}} + \underset{N\times d_W}{W}\underset{d_W\times d_A}{\delta_{WA}} + \epsilon_A$$

Test of relevance of Z for A given T is an underidentification test [Windmeijer, 2021]:

$$H_0 : \text{rank}\left(\delta_{Z_0A}\right) < d_A \text{ vs } H_1 : \text{rank}\left(\delta_{Z_0A}\right) = d_A$$

Reject H_0: Z is relevant for A given T.

3SLS procedure IV

3. Outcome OLS regression of Y on exogenous variation in A $(Z_0\hat{\delta}_{Z_0A})$, and T, W.

$$\underset{N\times 1}{Y} = \underset{N\times d_A}{(Z_0\hat{\delta}_{Z_0A})}\underset{d_A\times d_1}{\delta_{AY}} + \underset{N\times d_U}{T}\underset{d_U\times 1}{\delta_{TY}} + \underset{N\times d_W}{W}\underset{d_W\times 1}{\delta_{WY}} + \epsilon_Y$$

Consistent estimator of β: $\hat{\delta}_{AY}$. Closed form expression:

$$\hat{\delta}_{AY} = \left(Z_0^\intercal M_T Z_0\right)^{-1}\left(Z_0^\intercal M_T Y\right)$$

$$M_T = I - Z\,(Z^\intercal Z)^{-1}\,[Z^\intercal W]_{\mathrm{rr}} \left([W^\intercal Z]_{\mathrm{rr}}\,(Z^\intercal Z)^{-1}\,[Z^\intercal W]_{\mathrm{rr}}\right)^{-1}[W^\intercal Z]_{\mathrm{rr}}\,(Z^\intercal Z)^{-1}\,Z^\intercal$$

$$[Z^\intercal W]_{\mathrm{rr}} := Z^\intercal Z\hat{\delta}_{ZU} \overset{\text{if } \{d_W = d_U\}}{=} Z^\intercal W$$

Intuition for relevance requirements

1. $\text{rank}(\gamma_W) = d_U \leq d_W$ ensures that $\mathbb{E}_{\mathrm{L}}\left[U|Z\right]$ and $\mathbb{E}_{\mathrm{L}}\left[W|Z\right]$ are proportional. Keep $\mathbb{E}_{\mathrm{L}}\left[U|Z\right]$ fixed by keeping $\mathbb{E}_{\mathrm{L}}\left[W|Z\right]$ fixed (via T).

Intuition for relevance requirements

1. $\operatorname{rank}(\gamma_W) = d_U \leq d_W$ ensures that $\mathbb{E}_{\mathrm{L}}[U|Z]$ and $\mathbb{E}_{\mathrm{L}}[W|Z]$ are proportional. Keep $\mathbb{E}_{\mathrm{L}}[U|Z]$ fixed by keeping $\mathbb{E}_{\mathrm{L}}[W|Z]$ fixed (via T).
2. $\mathbb{E}[A^{\intercal}Z|\, T] = d_A$: Use remaining variation in Z to instrument for A while keeping $\mathbb{E}_{\mathrm{L}}[U|Z]$ fixed (via T). Requires $(d_Z - d_U) \geq d_A$.

Obtaining a valid control function T

Lemma 3.0.1

Assume $W \perp\!\!\!\perp Z \mid U$ (3a). Take any $\tau \in L_2(Z)$, where $T := \tau(Z)$, such that $U \perp\!\!\!\perp Z \mid T$. Then, also $W \perp\!\!\!\perp Z \mid T$.

Obtaining a valid control function T

Lemma 3.0.1

Assume $W \perp\!\!\!\perp Z \mid U$ (3a). Take any $\tau \in L_2(Z)$, where $T := \tau(Z)$, such that $U \perp\!\!\!\perp Z \mid T$. Then, also $W \perp\!\!\!\perp Z \mid T$.

Lemma 3.0.2

Assume $W \perp\!\!\!\perp Z \mid U$ (3a), and for any $g(U) \in L_2(U)$, $\mathbb{E}\left[g(U)|W\right] = 0$ only when $g(U) = 0$ (3b). Take any $\tau \in L_2(Z)$, where $T := \tau(Z)$, such that $W \perp\!\!\!\perp Z \mid T$. Then, also $U \perp\!\!\!\perp Z \mid T$.

Linearly separable outcome model

Assumption 3.1

There exists some function $k_0 \in L_2(A)$ such that

$$Y = Y(A) = k_0(A) + \varepsilon, \qquad \mathbb{E}\left[\varepsilon | Z, U\right] = \mathbb{E}\left[\varepsilon | U\right]. \tag{1}$$

Theorem 3.1

Let assumptions (2b/2c/3a/3b) and 3.1 hold. Any $h \in \mathcal{L}_2(A, T)$ for which $\mathbb{E}\left[Y|Z\right] = \mathbb{E}\left[h(A,T)|Z\right]$, satisfies $h(A,T) = k_0(A) + \mathbb{E}\left[\varepsilon|T\right]$. Consequently, $\theta := \int_{\mathcal{A}} Y(a)\pi(a)\,\mathrm{d}\mu_A(a) = \mathbb{E}_{\mathcal{T}}\left[\int_{\mathcal{A}} h(a,T)\pi(a)\,\mathrm{d}\mu_A(a)\right]$.

A simple plug-in estimator would be the empirical equivalent $\hat{\theta} = \mathbb{E}_{\mathcal{T},n}\left[\int_{\mathcal{A}} \hat{h}(a,T)\pi(a)\,\mathrm{d}\mu_A(a)\right]$.

First stage monotonicity

Assumption 3.2 (Monotonicity)

$$A = h(Z, \eta) \tag{2}$$

1. *$h(Z, \eta)$ is strictly monotonic in η with probability 1.*
2. *η is a continuously distributed scalar with a strictly increasing conditional CDF $F_{\eta|U}$ on the conditional support of η.*
3. *$Z \perp\!\!\!\perp \eta \mid U$.*

Disturbance conditional on T

Lemma 3.1.1

$$F_{\eta|T} := \int_{\mathcal{U}} F_{A|Z,U}(A, Z, u) f_{U|T}(u, T)\,\mathrm{d}\mu_U(u)$$

is a strictly increasing CDF on the conditional support of η*, and* $Z \perp\!\!\!\perp \eta \mid T$.

Conditional unconfoundedness

Theorem 3.2

Let

$$V := F_{A|Z}(A, Z). \tag{3}$$

Under assumption 3.2, $V = F_{\eta|T}(\eta)$*, and*

$$A \perp\!\!\!\perp Y(a) \mid (V, T), \text{ for all } a \in \mathcal{A}. \tag{4}$$

Average structural function identification

Assumption 3.3 (Common Support)

For all $a \in \mathcal{A}$, the support of (V, T) equals the support of (V, T) conditional on A.

Theorem 3.3

Suppose (2a/2b/3a/3b) [relaxed IV model], 3.2 [monotonicity], and 3.3 [common support] hold. Then, $\theta := \int_{\mathcal{A}} Y(a)\pi(a)\,\mathrm{d}\mu_A(a)$ is identified by

$$\theta = \mathbb{E}_{\mathcal{V},\mathcal{T}}\left[\int_{\mathcal{A}} \mathbb{E}\left[Y|A = a, (V, T) = (v, t)\right]\pi(a)\,\mathrm{d}\mu_A(a)\right].$$

NLS97 Data

Y Household net worth at 35: continuous, in USD

A BA degree: 1 if BA degree obtained, 0 otherwise

NLS97 Data

- Y Household net worth at 35: continuous, in USD
- A BA degree: 1 if BA degree obtained, 0 otherwise
- Z Pre-college test results: subject GPA, ASVAB percentile
- W Risky behaviour dummies : dummies for whether i drank (etc) by 17

NLS97 Data

- Y Household net worth at 35: continuous, in USD
- A BA degree: 1 if BA degree obtained, 0 otherwise
- Z Pre-college test results: subject GPA, ASVAB percentile
- W Risky behaviour dummies : dummies for whether i drank (etc) by 17
- U Ability: Unmeasured intellectual capacity
- Other biases: Selection on unobservables into obtaining BA degree (at least partly result of individual optimisation)

NLS97 Data

- Y Household net worth at 35: continuous, in USD
- A BA degree: 1 if BA degree obtained, 0 otherwise
- Z Pre-college test results: subject GPA, ASVAB percentile
- W Risky behaviour dummies : dummies for whether i drank (etc) by 17
- U Ability: Unmeasured intellectual capacity
- ▶ Other biases: Selection on unobservables into obtaining BA degree (at least partly result of individual optimisation)
- X Covariates: sex, college GPA, parental education/net worth, siblings, region, etc

Assumption 4.1

1. *Linear model:* $Y = \alpha_Y + A\beta + U\gamma_Y + W\upsilon_Y + X\eta_Y + \varepsilon_Y.$

2a. *Cond. Instr. Exogeneity:* $\mathbb{E}\left[\varepsilon_Y^\intercal Z\right] = 0.$

2c. *Cond. Instr. Relevance:*
$$A = \alpha_A + Z\zeta + U\gamma_A + W\upsilon_A + X\eta_A + \varepsilon_A$$
$$rank\left(\mathbb{E}\left[(Z\zeta)\,A\middle|\,\mathbb{E}\left[W|Z,X\right],X\right]\right) = d_A.$$

3a. *Proxy Exogeneity:* $W = \alpha_W + U\gamma_W + X\eta_W + \varepsilon_W, \;\; \mathbb{E}\left[\varepsilon_W^\intercal Z\right] = 0.$

3b. *Proxy Relevance:* $rank(\gamma_W) = d_U \leq d_W$

$$\implies \beta = \frac{\mathbb{E}\left[(Z\zeta)\,Y \mid \mathbb{E}\left[W|Z,X\right],X\right]}{\mathbb{E}\left[(Z\zeta)\,A \mid \mathbb{E}\left[W|Z,X\right],X\right]}$$

1. Find T and test relevance of Z, W for U I

- Linear projection

$$\mathbb{E}\left[U|Z,X\right] = Z\gamma_Z + X\gamma_X$$
$$\mathbb{E}\left[W|Z,X\right] = \tilde{\alpha}_W + Z\gamma_Z\gamma_W + X\tilde{\eta}_W$$

- Hypothesis test: For some $r < \min\{d_Z, d_W\}$,

$$H_0 : \text{rank}(\gamma_Z\gamma_W) \leq r, \text{ vs } H_1 : \text{rank}(\gamma_Z\gamma_W) > r.$$

  If H_0 true, can hold anything correlating Z and W fixed.

- Normalisation by singular value decomposition:

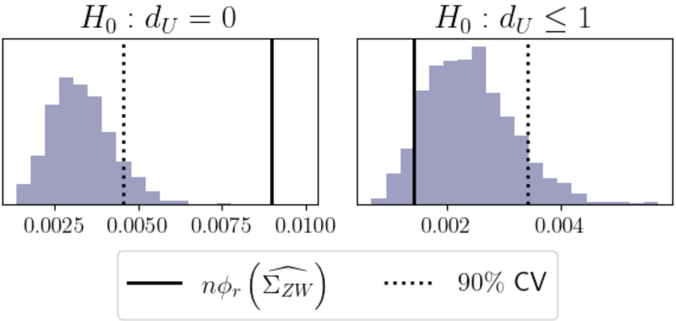
$$\gamma_Z\gamma_W = \underset{d_Z\times d_Z}{P_0}\,\underset{d_Z\times d_W}{\Pi_0}\,\underset{d_W\times d_W}{Q_0^\intercal} \overset{\{\text{if } H_0 \text{ true}\}}{=} \underset{d_Z\times r}{P_{0,1}}\,\underset{r\times r}{\Pi_{0,1}}\,\underset{r\times d_W}{Q_{0,1}^\intercal}$$
$$\implies \underset{n\times r}{T} = \underset{n\times d_Z}{Z}\,\underset{d_Z\times r}{P_{0,1}}\,\underset{r\times r}{\Pi_0}$$

1. Find T and test relevance of Z, W for U II

- Equivalent test: For some $r < \min\{d_Z, d_W\}$,

$$H_0 : \phi_r(\gamma_Z \gamma_W) = 0, \text{ vs } H_1 : \phi_r(\gamma_Z \gamma_W) > 0,$$

where $\phi_r(A) := \sum_{j>r} \pi_j^2(A)$ (sum of $(r+1)$-th to smallest singular value squared).



1. Find T and test relevance of Z, W for U III

- $\text{rank}(\gamma_Z \gamma_W) = 1$: one-dimensional U explains correlation between Z, W.

2. Test relevance of Z for A given T

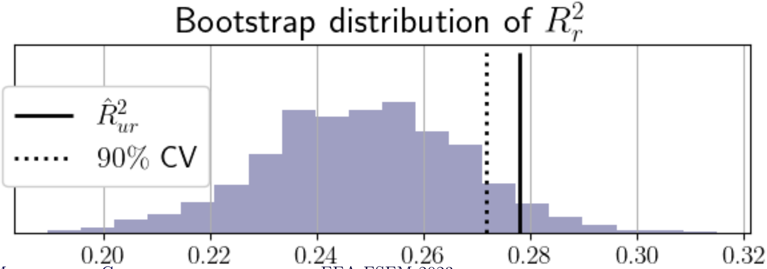
- Construct control function: $\underset{n\times 1}{T} = \underset{n\times d_Z}{Z}\underset{d_Z\times 1}{\hat{P}_{0,1}\hat{\pi}_1}$
- Compare restricted and unrestricted R^2:

$$A = \tilde{\alpha}_{A,ur} + Z\tilde{\zeta} + X\tilde{\eta}_{A,ur} + \tilde{\varepsilon}_{A,ur} \implies R^2_{ur} \quad (5)$$

$$A = \tilde{\alpha}_{A,r} + T\tilde{\gamma}_A + X\tilde{\eta}_{A,r} + \tilde{\varepsilon}_{A,r} \implies R^2_r \quad (6)$$

Hypothesis test:

$$H_0 : R^2_r = R^2_{ur}, \text{ vs } H_1 : R^2_r < R^2_{ur}.$$



3. Exogeneity of Z conditional on U I

$$Y = \alpha_Y + A\beta + U\gamma_Y + W\upsilon_Y + X\eta_Y + \varepsilon_Y.$$

2a. Cond. Instr. Exogeneity: $\mathbb{E}\left[\varepsilon_Y^\intercal (Z, X)\right] = 0.$

- Untestable in this just-identified model
- Does T reflect the hypothesised confounder ability?

Table: Construction of $\hat{T} = Z\hat{P}_{0,1}\hat{\Pi}_{0,1}$ (normalised to $\sigma(\hat{T}) = 1$)

	GPA English	GPA Math	GPA SocSci	GPA LifeSci	ASVAB percentile
$\hat{T}$	0.574	0.213	0.282	0.212	-0.278

- Weighted subject-GPA describes 94.4% of $\hat{T}$.
- Seems to reflect overall GPA and thus general ability.

3. Exogeneity of Z conditional on U II

Table: Effect of normalised $\hat{T}$ on W

	drink	smoke	try marijuana	run away	attack someone
$\hat{T}$	-0.071	-0.093	-0.101	-0.054	-0.093
Pr	0.653	0.468	0.296	0.106	0.189

		sell drugs	destroy property	steal $< 50\$$	steal $> 50\$$
$\hat{T}$		-0.052	-0.051	-0.051	-0.036
Pr		0.089	0.320	0.379	0.079

- One standard-deviation increase in $\hat{T}$ significantly reduces risky behaviour.
- Expected for confounder ability.

4. Estimation

$$\hat{\beta}_{\mathrm{OLS}} = \left(A^\intercal M_{W,X} A\right)^{-1}\left(A^\intercal M_{W,X} Y\right)$$

$$\hat{\beta}_{\mathrm{PL}} = \left(A^\intercal M_{\hat{T},X} A\right)^{-1}\left(A^\intercal M_{\hat{T},X} Y\right),$$

where $\hat{T}$ is constructed from the correlation of (A, Z) and W.

$$\hat{\beta}_{\mathrm{IV}} = \left(A^\intercal P_Z M_{W,X} A\right)^{-1}\left(A^\intercal P_Z M_{W,X} Y\right)$$

$$\hat{\beta}_{\mathrm{ICC}} = \left(A^\intercal P_Z M_{\hat{T},X} A\right)^{-1}\left(A^\intercal P_Z M_{\hat{T},X} Y\right),$$

where $\hat{T}$ is constructed from the correlation of Z and W.

Notation: Projection $P_X = X\left(X^\intercal X\right)^{-1} X^\intercal$ and annihilator $M_X = I - P_X$.

Note: $\underset{n\times r}{\hat{T}} = \underset{n\times d_Z}{Z}\ \underset{d_Z\times r}{\hat{P}_{0,1}}\ \underset{r\times r}{\hat{\Pi}_0}$ is a dimension-reduced projection of Z on W, because $\underset{d_Z\times d_Z}{\hat{P}_0}\ \underset{d_Z\times d_W}{\hat{\Pi}_0}\ \underset{d_W\times d_W}{\hat{Q}_0^\intercal} = \left(Z^\intercal M_X Z\right)^{-1}\left(Z^\intercal M_X W\right)$.

Results

Table: Estimates with different estimators (NLS97 data, $n = 1,890$)

	OLS	PL	IV	ICC
A	59,173	33,799	204,578	122,665
	(9,542)	(10,765)	(34,265)	(49,138)
$\hat{T}$		27,879		15,829
		(5,415)		(7,613)

- Positive ability bias (was ambiguous).
- Negative general selection bias (as expected).
- Standard error increases about 40% in ICC compared to IV.

Generalised setting

For all $k_0(\tau_0, g_0) \in \mathcal{K}_0(\tau_0, g_0)$, and $\tau_0(g_0) \in \mathcal{T}_{\text{valid}}(g_0) \subset \mathcal{T}_{\text{exog}}(g_0)$,

$$\begin{aligned}
\theta_0 &= \mathbb{E}\left[m(O; k_0(\tau_0, g_0))\right], \\
\mathcal{K}_0(\tau, g) &:= \left\{k \in \mathcal{K} : \ \mathbb{E}\left[k(A; \tau, g) | Z\right] = g(Z) - \tau(Z)\right\} \\
\mathcal{T}_{\text{exog}}(g) &:= \left\{\tau \in \mathcal{T} : \mathbb{E}\left[g(Z) | W\right] = \mathbb{E}\left[\tau(Z) | W\right]\right\} \\
g_0(Z) &:= \mathbb{E}\left[Y | Z\right]
\end{aligned}$$

Generalised setting

For all $k_0(\tau_0, g_0) \in \mathcal{K}_0(\tau_0, g_0)$, and $\tau_0(g_0) \in \mathcal{T}_{\text{valid}}(g_0) \subset \mathcal{T}_{\text{exog}}(g_0)$,

$$
\begin{aligned}
\theta_0 &= \mathbb{E}\left[m(O; k_0(\tau_0, g_0))\right], \\
\mathcal{K}_0(\tau, g) &\coloneqq \left\{k \in \mathcal{K} : \ \mathbb{E}\left[k(A; \tau, g) | Z\right] = g(Z) - \tau(Z)\right\} \\
\mathcal{T}_{\text{exog}}(g) &\coloneqq \left\{\tau \in \mathcal{T} : \mathbb{E}\left[g(Z) | W\right] = \mathbb{E}\left[\tau(Z) | W\right]\right\} \\
g_0(Z) &\coloneqq \mathbb{E}\left[Y | Z\right]
\end{aligned}
$$

Problems:

- Nested dependence of nuisance functions [Chernozhukov et al., 2022]
- Weak identification of nuisance functions: Non-uniqueness and ill-posedness [Bennett et al., 2023]

Continuous linear functionals

Assume that $k \mapsto \mathbb{E}\left[m(O;k)\right]$ for $k \in \mathcal{K}$ is a continuous linear functional over $\mathcal{K}$, such that by the Riesz representation theorem

$$\mathbb{E}\left[m_0(O;k)\right] = \mathbb{E}\left[\alpha_{k,0}(A)k(A)\right] \ \forall k \in \mathcal{K}. \tag{7}$$

Strong instrument relevance

Define the linear operator $P_{\mathcal{L}_2(Z)}^{A,\mathcal{K}}$ and its adjoint $P_{A,\mathcal{K}}^{\mathcal{L}_2(Z)}$ (where Π is the projection operator):

$$\left[P_{\mathcal{L}_2(Z)}^{A,\mathcal{K}} k\right](Z) := \mathbb{E}\left[k(A)|Z\right]$$

$$\left[P_{A,\mathcal{K}}^{\mathcal{L}_2(Z)} q_k\right](A) := \Pi_{\mathcal{K}}\,\mathbb{E}\left[q_k(Z)|A\right] = \Pi_{\mathcal{K}} q_k(Z)|A$$

Assumption 5.1 (Strong instrument relevance)

$\alpha_{k,0} \in \mathcal{N}^{\perp}\left(P_{A,\mathcal{K}}^{\mathcal{L}_2(Z)} P_{\mathcal{L}_2(Z)}^{A,\mathcal{K}}\right)$, *i.e.*

$$\Xi_{k,0} \neq \emptyset, \quad \textit{where} \quad \Xi_{k,0} := \arg\min_{\xi_k \in \mathcal{K}} \left(\frac{1}{2}\mathbb{E}\left[\mathbb{E}\left[\xi_k(A)|Z\right]^2\right] - \mathbb{E}\left[m_0(O;\xi_k)\right]\right) \tag{8}$$

$$= \left\{\xi_k \in \mathcal{K} : P_{A,\mathcal{K}}^{\mathcal{L}_2(Z)} P_{\mathcal{L}_2(Z)}^{A,\mathcal{K}} \xi_k = \alpha_{k,0}\right\}. \tag{9}$$

Implied restrictions on the function space $\mathcal{K}_0$

Strong instrument relevance as stated in 5.1 requires the remaining variation in Z to be relevant enough with respect to outcome-relevant variation in treatment A after integrating out any of the outcome-relevant variation in A associated with the endogenous variation $\tau(Z)$.

$$\mathcal{K} \times \mathcal{T} = \Big\{(k,\tau) : \Big(\mathbb{E}\left[k(A)|Z\right] = g(Z) - \tau(Z)\Big) \\ \wedge \Big(\mathbb{E}\left[g(Z)|W\right] = \mathbb{E}\left[\tau(Z)|W\right]\Big) \;\; \forall g \in \mathcal{G}\Big\}$$

$$\mathcal{K}_0 \times \mathcal{T}_{\text{exog}} = \Big\{(k,\tau) : \Big(\mathbb{E}\left[k(A)|Z\right] = g_0(Z) - \tau(Z)\Big) \\ \wedge \Big(\mathbb{E}\left[g_0(Z)|W\right] = \mathbb{E}\left[\tau(Z)|W\right]\Big)\Big\}$$

Debiasing step 1

Debiased moment wrt k:

$$m_1(O; k, \tau, g, q_k) = m_0(O; k) + q_k(Z)\left(g(Z) - \tau(Z) - k(A; \tau, g)\right)$$

$$\mathcal{Q}_k := \left\{q_k \in \mathcal{L}_2(Z) : q_k(Z) = \mathbb{E}\left[\xi_k(A)|Z\right] \ \forall \xi_k \in \mathcal{K}\right\}$$
$$\mathcal{Q}_{k,0} := \left\{q_k \in \mathcal{L}_2(Z) : q_k(Z) = \mathbb{E}\left[\xi_{k,0}(A)|Z\right] \ \forall \xi_{k,0} \in \Xi_{k,0}\right\}$$

Debiasing step 2

Debiased moment wrt k and τ:

$$m_2(O; k, \tau, g, q_k, q_\tau) = m_1(O; k, \tau, g, q_k) + q_\tau(W; q_k)\left(\tau(Z) - g(Z)\right)$$

$$\mathcal{Q}_\tau(q_k) := \left\{ q_\tau \in \mathcal{L}_2(W) : q_\tau = \mathbb{E}\left[\xi_\tau(Z; q_k) | W\right],\ \forall \xi_\tau(q_k) \in \mathcal{T}, q_k \in \mathcal{Q}_k \right\}$$

$$\mathcal{Q}_{\tau,0}(q_k) := \left\{ q_\tau \in \mathcal{L}_2(W) : q_\tau = \mathbb{E}\left[\xi_{\tau,0}(Z; q_k) | W\right],\ \forall \xi_{\tau,0}(q_{k,0}) \in \Xi_{\tau,0}(q_k), q_k \in \mathcal{Q}_k \right\}$$

$$\Xi_{\tau,0}(q_k) := \arg \min_{\xi_\tau(q_k) \in \mathcal{T}} \left(\frac{1}{2} \mathbb{E}\left[\mathbb{E}\left[\xi_\tau(Z; q_k) | W\right]^2\right] - \mathbb{E}\left[q_k(Z)\xi_\tau(Z; q_k)\right] \right).$$

For this step, use that a continuous linear functional in τ is always strongly identified in this setting (see paper on arxiv Theorem 4).

Debiasing step 3

Debiased moment wrt k, τ, and g:

$$m_3(O; k, \tau, g, q_k, q_\tau, \alpha_g) = m_2(O; k, \tau, g, q_k, q_\tau) + \alpha_g(Z; q_k, q_\tau)\left(Y - g(Z)\right)$$

$$\begin{aligned}\Xi_{g,0}(q_k, q_\tau) := \arg \min_{\xi_g(q_k, q_\tau) \in \mathcal{L}_2(Z)} \Bigg(& \frac{1}{2}\mathbb{E}\left[\xi_g(Z; q_k, q_\tau)^2\right] \\ & - \mathbb{E}\left[\left(q_k(Z) - q_\tau(W; q_k)\right)\xi_g(Z; q_k, q_\tau)\right]\Bigg).\end{aligned}$$

Each debiased moment identifies θ

Lemma 5.0.1

Assume $\theta_0 = \mathbb{E}\left[m_0(O; k_0(\tau_0, g_0))\right]$ *and 5.1. Then,*

$$\begin{aligned}\theta_0 = \mathbb{E}\left[m_0(O; k_0)\right] &= \mathbb{E}\left[m_1(O; k_0, \tau_0, g_0, q_{k,0})\right] \\ &= \mathbb{E}\left[m_2(O; k_0, \tau_0, g_0, q_{k,0}, q_{\tau,0})\right] \\ &= \mathbb{E}\left[m_3(O; k_0, \tau_0, g_0, q_{k,0}, q_{\tau,0}, \alpha_{g,0})\right].\end{aligned}$$

Each of m_j for $j \in \{1, 2, 3\}$ identify θ_0 because in each consecutive debiasing step a conditionally mean-zero term is added to the previous moment.

Robustness of final debiased moment

Theorem 5.1 (Robust error decomposition)

Suppose $\theta_0 = \mathbb{E}\left[m_0(O; k_0(\tau_0, g_0))\right]$, assumption 5.1, and the conditions for theorem **??** *hold for $q = q_\tau$. Then,*

$$
\begin{aligned}
&\mathbb{E}\left[m_3(O; k, \tau, g, q_k, q_\tau, \alpha_g)\right] - \theta_0 \\
&= \mathbb{E}\left[\Big(q_{k,0}(Z) - q_k(Z)\Big)\left(k(A; \tau, g) - k_0(A; \tau_0, g_0)\right)\right] \\
&\quad + \mathbb{E}\left[\left(q_{\tau,0}(W; q_k) - q_\tau(W; q_k)\right)\left(\tau_0(Z; g_0) - \tau(Z; g)\right)\right] \\
&\quad + \mathbb{E}\left[\left(\alpha_{g,0}(Z; q_k, q_\tau) - \alpha_g(Z; q_k, q_\tau)\right)\left(g(Z) - g_0(Z)\right)\right].
\end{aligned}
$$

Double robustness, Neyman orthogonality, and an error decomposition in terms of projections on more favourable subspaces (circumventing ill-posedness problems) follow.

Conclusion

- Identification approach between IV and proximal learning
- Allows some endogeneity in instruments, as long as relevant proxies exist for the unobserved causes of instrument endogeneity
- Motivated by traditional economic identification problems with self-selection into treatment (returns to education)
- Semiparametric estimation with $\sqrt{n}$-rates possible under assumptions on nuisance convergence rates in terms of more favourable projected errors (circumvents potential ill-posedness of inverse problems)

Thank You

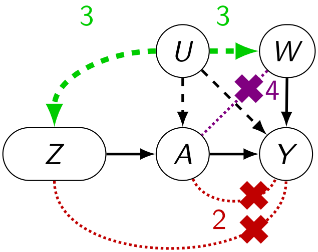
arXiv 2301.02052

ct493@cam.ac.uk

Appendix

Proximal Learning [Tchetgen Tchetgen et al., 2020]

Figure: Proximal learning

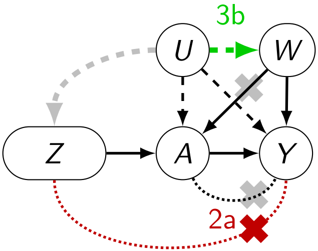


Assumption (Proximal learning)

2. *Cond. Exogeneity of action and its aligned proxy:* $Y(a, z) = Y(a) \perp\!\!\!\perp (A, Z) \mid U$.
3. *Relevance of both proxies:* For any $g(U) \in L_2(U)$, $\mathbb{E}\left[g(U)|Z\right] = 0$ *only if* $g(U) = 0$, $\mathbb{E}\left[g(U)|W\right] = 0$ *only if* $g(U) = 0$.
4. *Exogeneity of outcome-aligned proxies:* $W(a, z) = W \perp\!\!\!\perp (A, Z) \mid U$.

Relaxations compared to proximal learning

Figure: Proximal learning



Assumption

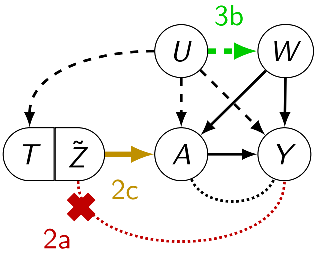
2a. *Cond. Exogeneity of action-aligned proxy:* $Y(a,z) = Y(a) \perp\!\!\!\perp Z \mid U$.

3a. *Proxy Exogeneity:* $W(z) = W \perp\!\!\!\perp Z \mid U$.

3b. *Relevance of outcome-aligned proxy:* For any $g(U) \in L_2(U)$, $\mathbb{E}\left[g(U)|W\right] = 0$ *only if* $g(U) = 0$.

Additional assumptions compared to proximal learning

Figure: Proximal learning



Assumption

2a. *Cond. Instrument Exogeneity:*
$Y(a,z) = Y(a) \perp\!\!\!\perp Z \mid U.$

2b. *Index sufficiency:* $U \perp\!\!\!\perp Z \mid T$ *for some* $T = \tau(Z)$.

2c. *Cond. Instrument Relevance:*
For any $g(A,T) \in L_2(A,T)$,
$\mathbb{E}\left[g(A,T)|Z\right] = 0$ *only if* $g(A,T) = 0$.

3a. *Proxy Exogeneity:*
$W(z) = W \perp\!\!\!\perp Z \mid U.$

3b. *Relevance of outcome-aligned proxy:*
For any $g(U) \in L_2(U)$,
$\mathbb{E}\left[g(U)|W\right] = 0$ *only if* $g(U) = 0$.

The obvious

Lemma 8.0.1

Assume $W \perp\!\!\!\perp Z \mid U$ *(3a). Take any* $\tau \in L_2(Z)$*, where* $T := \tau(Z)$*, such that* $U \perp\!\!\!\perp Z \mid T$*. Then, also* $W \perp\!\!\!\perp Z \mid T$*.*

The obvious

Lemma 8.0.1

Assume $W \perp\!\!\!\perp Z \mid U$ (3a). Take any $\tau \in L_2(Z)$, where $T := \tau(Z)$, such that $U \perp\!\!\!\perp Z \mid T$. Then, also $W \perp\!\!\!\perp Z \mid T$.

$$
\begin{aligned}
& f_{W,Z|T}(W, Z|T) \\
&\quad = \int_{\mathcal{U}} \underbrace{f_{W|Z,U,T}(W|Z, u, T)}_{=f_{W|U}(W|u)} \underbrace{f_{Z|U,T}(Z|u, T)}_{=f_{Z|T}(Z|T)} f_{U|T}(u, T) \, \mathrm{d}\mu_U(u) \\
&\quad = f_{Z|T}(Z|T) \underbrace{\int_{\mathcal{U}} f_{W|U}(W|u) f_{U|T}(u, T) \, \mathrm{d}\mu_U(u)}_{f_{W|T}(W|T)} \implies W \perp\!\!\!\perp Z \mid T
\end{aligned}
$$

The slightly less obvious I

Lemma 8.0.2

Assume $W \perp\!\!\!\perp Z \mid U$ (3a), and for any $g(U) \in L_2(U)$, $\mathbb{E}\left[g(U)|W\right] = 0$ only when $g(U) = 0$ (3b). Take any $\tau \in L_2(Z)$, where $T := \tau(Z)$, such that $W \perp\!\!\!\perp Z \mid T$. Then, also $U \perp\!\!\!\perp Z \mid T$.

The slightly less obvious II

Write $f_{W|Z}(W,Z)$ in two separate ways using T and relate them.

1. $f_{W|Z}(W,Z) = \int_{\mathcal{U}} f_{W|U}(W|u) f_{U|Z}(u|Z)\,\mathrm{d}\mu_U(u)$ by $W \perp\!\!\!\perp Z \mid U$
2. $f_{W|Z}(W,Z) = f_{W|T}(W,T) = \int_{\mathcal{U}} f_{W|U}(W,u) f_{U|T}(u,T)\,\mathrm{d}\mu_U(u)$ by construction of $T = \tau(Z)$ such that $W \perp\!\!\!\perp Z \mid T$.

$$\int_{\mathcal{U}} f_{W|U}(W,u) \left(f_{U|Z}(u|Z) - f_{U|T}(u|T) \right) \mathrm{d}\mu_U(u) = 0$$

$$\int_{\mathcal{U}} \left(f_{U|Z}(u|Z) - f_{U|T}(u|T) \right) \frac{f_W(W)}{f_U(u)} f_{U|W}(u,W)\,\mathrm{d}\mu_U(u) = 0$$

$$\mathbb{E}_U \left[\left. \frac{\left(f_{U|Z}(u|Z) - f_{U|T}(u|T) \right)}{f_U(u)} \right| W \right] f_W(W) = 0$$

The slightly less obvious III

Then, for any Z, let $g_Z(U) := \frac{\left(f_{U|Z}(u|Z) - f_{U|T}(u|T)\right)}{f_U(u)}$.

$$\mathbb{E}_U\left[g_Z(u)\middle|\, W\right] f_W(W) = 0$$

Completeness of W for U (3b) implies $g_Z(U) = 0$, and thus $f_{U|Z}(U|Z) = f_{U|T}(U|T)$, meaning $U \perp\!\!\!\perp Z \mid T$.

Valid control functions

Valid control functions $\tau \in L_2(Z)$ satisfy two conditions:

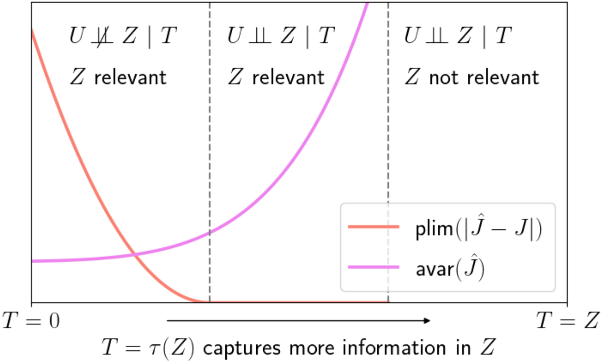
1. Conditional independence of W and Z:

$$W \perp\!\!\!\perp Z \mid \tau(Z).$$

2. Conditional relevance of Z for A:

$$\mathbb{E}\left[g(A, \tau(Z)|Z\right] = 0 \text{ only if } g(A, \tau(Z)) = 0.$$

Optimal control functions



- Optimal: minimum complexity $\tau(Z)$ subject to validity, e.g. $f(W|Z)$

Specification test

- Standard specification test for sufficient complexity of τ_1 vs τ_2.

$$H_0 : Z \perp\!\!\!\perp Y(a) \mid \tau_1(Z), \text{ vs } H_1 : Z \not\!\perp\!\!\!\perp Y(a) \mid \tau_1(Z).$$

- Test whether $\tau_1(.)$ valid, if $\tau_2(.)$ valid.
- Under H_0:
  - $\text{plim}\left(\hat{\theta}(\tau_1)\right) = \text{plim}\left(\hat{\theta}(\tau_2)\right)$,
  - $\text{avar}\left(\hat{\theta}(\tau_1)\right) < \text{avar}\left(\hat{\theta}(\tau_2)\right)$.
- Under H_1:
  - $\text{plim}\left(\hat{\theta}(\tau_1)\right) \neq \text{plim}\left(\hat{\theta}(\tau_2)\right)$.

Simple linear model

$$Y = \alpha_Y + A\beta + U \underset{d_U \times 1}{\gamma_Y} + W \underset{d_W \times 1}{\upsilon_Y} + X \underset{d_X \times 1}{\eta_Y} + \varepsilon_Y,$$

$U\gamma_Y$ Probably positive effect of ability U on net worth Y, by salary and non-salary mediation [Griliches, 1977].

ε_Y All variation in Y, which is jointly unexplained by (A, U, W, X). Individual-specific, heterogeneous characteristics.

Expected bias I

- i chooses whether to obtain a BA degree by maximising expected utility subject to information set $\mathcal{I}$:

$$A = \underset{a \in \{0,1\}}{\arg\max} \left(\mathbb{E}\left[u(Y(a)) - c(a) | A = a, \mathcal{I} \right] \right),$$

$u: \mathcal{Y} \to \mathbb{R}$ is a diminishing returns utility function
$c: \{0,1\} \to \mathbb{R}$ is a cost function for obtaining a BA degree.

- Optimal decision rule assuming full information:

$$\begin{aligned} A &= \underset{a \in \{0,1\}}{\arg\max} \left(u(Y(a)) - c(a) \right), \\ &= \mathbb{1}\left(u(Y(1)) - u(Y(0)) > c(1) - c(0) \right). \end{aligned}$$

Expected bias II

$U \Uparrow$ Ambiguous bias direction

$$\mathbb{1}\Big(\underbrace{u(Y(1)) - u(Y(0))}_{\Downarrow \text{ as } Y(a)\Uparrow \text{ and } \Delta(Y(1)-Y(0))=0} > \underbrace{c(1) - c(0)}_{\Downarrow \text{ when ability higher}} \Big)$$

$\varepsilon_Y \Uparrow$ Positive bias direction

$$\mathbb{1}\Big(\underbrace{u(Y(1)) - u(Y(0))}_{\Downarrow \text{ as } Y(a)\Uparrow \text{ and } \Delta(Y(1)-Y(0))=0} > c(1) - c(0)\Big)$$

- Byproduct: Separate ability bias from other biases.
- Theoretical models can produce differing bias directions.
- Empirical validation would be useful.

Andrew Bennett, Nathan Kallus, Xiaojie Mao, Whitney Newey, Vasilis Syrgkanis, and Masatoshi Uehara. Inference on strongly identified functionals of weakly identified functions. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 2265–2265. PMLR, 2023.

Qihui Chen and Zheng Fang. Improved inference on the rank of a matrix. *Quantitative Economics*, 10(4):1787–1824, 2019.

Victor Chernozhukov, Whitney Newey, Rahul Singh, and Vasilis Syrgkanis. Automatic debiased machine learning for dynamic treatment effects and general nested functionals. *arXiv preprint arXiv:2203.13887*, 2022.

Yifan Cui, Hongming Pu, Xu Shi, Wang Miao, and Eric Tchetgen Tchetgen. Semiparametric proximal causal inference. *arXiv preprint arXiv:2011.08411*, 2020.

Ben Deaner. Proxy controls and panel data. *arXiv preprint arXiv:1810.00283*, 2018.

Zvi Griliches. Estimating the returns to schooling: Some econometric problems. *Econometrica: Journal of the Econometric Society*, pages 1–22, 1977.

Guido W Imbens and Whitney K Newey. Identification and estimation of triangular simultaneous equations models without additivity. *Econometrica*, 77(5):1481–1512, 2009.

Laura Liu, Alexandre Poirier, and Ji-Liang Shiu. Identification and estimation of average partial effects in semiparametric binary response panel models. *arXiv preprint arXiv:2105.12891*, 2021.

Kenichi Nagasawa. Treatment effect estimation with noisy conditioning variables. *arXiv preprint arXiv:1811.00667*, 2018.

Whitney K Newey and James L Powell. Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578, 2003.

Eric J Tchetgen Tchetgen, Andrew Ying, Yifan Cui, Xu Shi, and Wang Miao. An introduction to proximal causal learning. *arXiv preprint arXiv:2009.10982*, 2020.

Frank Windmeijer. Testing underidentification in linear models, with applications to dynamic panel and asset pricing models. *Journal of Econometrics*, 2021.