Identification-robust inference for the LATE with high-dimensional covariates

Yukun Ma
Vanderbilt University

ESEM
August 27, 2023

Motivation

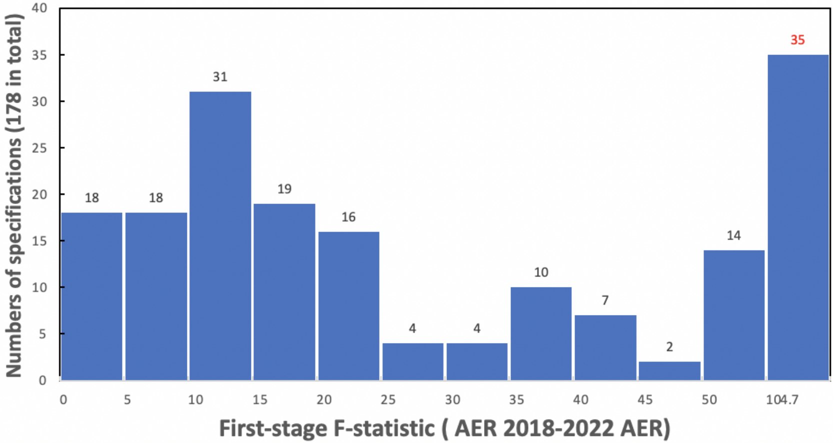


Figure: *American Economic Review* 2018-2022

heterscadesticity

LATE

- the effect of a treatment for subjects who comply with the experimental treatment assigned to their sample group (compliers).
- Assume we have N observations
  - Y_i : outcome of interest for unit i.
  - $D_i \in \{0, 1\}$: receipt of treatment.
  - $Z_i \in \{0, 1\}$: offer of the treatment.
  - X_i : p-dimensional controls
    (e.g. high-dimensional covariates $p \gg N$).
- Imbens and Angrist (1994) propose

$$\theta = \frac{\mathbf{E}_P[Y|Z=1] - \mathbf{E}_P[Y|Z=0]}{\mathbf{E}_P[D|Z=1] - \mathbf{E}_P[D|Z=0]} = \frac{ITT}{ITT_D} := \frac{\delta}{\pi}.$$

- Weak identification in LATE: $\pi \to 0$

Weak identification

- When instruments $\boldsymbol{Z}$ are weakly correlated with endogenous regressors $\boldsymbol{D}$, conventional methods for IV estimation and inference become unreliable.

  $$\boldsymbol{\theta} = \frac{\boldsymbol{\delta}}{\boldsymbol{\pi}},$$

  normal approximation of $\widehat{\boldsymbol{\theta}}$ can be derived using delta method by linearized $\widehat{\boldsymbol{\theta}}$ in $(\widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\pi}})$. However, $\widehat{\boldsymbol{\theta}}$ is highly nonlinear in $\widehat{\boldsymbol{\pi}}$ when $\widehat{\boldsymbol{\pi}}$ is close to zero.
- Solution: test inversion.
  Given $\boldsymbol{H_0} : \boldsymbol{\theta} = \boldsymbol{\theta_0}$, we have $\boldsymbol{\delta} - \boldsymbol{\theta_0}\boldsymbol{\pi} = \mathbf{0}$. Then the AR statistic

  $$AR(\theta) = (\delta - \theta\pi)'\Omega(\theta)^{-1}(\delta - \theta\pi)$$

  follows a χ^2 distribution under $\boldsymbol{H_0}$.
- A large literature in econometrics has developed methods for making inference with weak instruments,
  - Stock and Wright (2000) $\Rightarrow$ S test.
  - Kleibergen (2002) $\Rightarrow$ K test.
  - Andrews and Mikusheva (2016) $\Rightarrow$ QLR test and pQLR test.
- none of them considers the model with high-dimensional covariates.

Weak identification

- When instruments $\boldsymbol{Z}$ are weakly correlated with endogenous regressors $\boldsymbol{D}$, conventional methods for IV estimation and inference become unreliable.

$$\boldsymbol{\theta} = \frac{\boldsymbol{\delta}}{\boldsymbol{\pi}},$$

normal approximation of $\widehat{\boldsymbol{\theta}}$ can be derived using delta method by linearized $\widehat{\boldsymbol{\theta}}$ in $(\widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\pi}})$. However, $\widehat{\boldsymbol{\theta}}$ is highly nonlinear in $\widehat{\boldsymbol{\pi}}$ when $\widehat{\boldsymbol{\pi}}$ is close to zero.

- Solution: test inversion.
Given $\boldsymbol{H_0} : \boldsymbol{\theta} = \boldsymbol{\theta_0}$, we have $\boldsymbol{\delta} - \boldsymbol{\theta_0}\boldsymbol{\pi} = \mathbf{0}$. Then the AR statistic

$$\boldsymbol{AR}(\boldsymbol{\theta}) = (\boldsymbol{\delta} - \boldsymbol{\theta}\boldsymbol{\pi})'\boldsymbol{\Omega}(\boldsymbol{\theta})^{-1}(\boldsymbol{\delta} - \boldsymbol{\theta}\boldsymbol{\pi})$$

follows a $\boldsymbol{\chi}^2$ distribution under $\boldsymbol{H_0}$.

- A large literature in econometrics has developed methods for making inference with weak instruments,
  - Stock and Wright (2000) $\Rightarrow$ S test.
  - Kleibergen (2002) $\Rightarrow$ K test.
  - Andrews and Mikusheva (2016) $\Rightarrow$ QLR test and pQLR test.

- none of them considers the model with high-dimensional covariates.

Contributions

- Weak identification in an IV context:
  - S statistic by Stock and Wright (2000), K statistic by Kleibergen (2005), Conditional test by Moreira (2003,2009), Andrews and Mikusheva (2016).
  - An important complement to existing literature: $\boldsymbol{p \gg N}$
- ML based econometric methods:
  - Belloni, Chernozhukov, and Kato (2015), Chernozhukov et al. (2013,2016,2017).
  - An important complement to existing literature: weak identification.

Setup

- Model random vector $W = (Y, D, Z, X')'$ as follows,

$$\begin{aligned} E[D|Z,X] &= \Lambda(Z\beta_{11} + X'\beta_{12}) && \text{(First stage)} \\ E[Z|X] &= \Lambda(X\gamma) && \text{(Propensity score)} \\ E[Y|Z,X] &= Z\beta_{21} + X'\beta_{22} && \text{(Reduce form)} \end{aligned}$$

  - Y: the outcome of interest
  - $D \in \{0,1\}$: receipt of treatment
  - $Z \in \{0,1\}$: offer of treatment
  - X: p-dimensional controls
  - $\Lambda(t) = \frac{\exp(t)}{1+\exp(t)}$ for all $t \in \mathbb{R}$

Setup

- Model random vector $W = (Y, D, Z, X')'$ as follows,

$$E[D|Z,X] = \Lambda(Z\beta_{11} + X'\beta_{12}) := m(Z,X) \quad \text{(First stage)}$$
$$E[Z|X] = \Lambda(X\gamma) := p(X) \quad \text{(Propensity score)}$$
$$E[Y|Z,X] = Z\beta_{21} + X'\beta_{22} := g(Z,X) \quad \text{(Reduce form)}$$

  - Y: the outcome of interest
  - $D \in \{0,1\}$: receipt of treatment
  - $Z \in \{0,1\}$: offer of treatment
  - X: p-dimensional controls
  - $\Lambda(t) = \frac{\exp(t)}{1+\exp(t)}$ for all $t \in \mathbb{R}$

- The doubly robust LATE proposed by Tan (2006) is given by

$$\theta_0 = \frac{E[g(1,X)-g(0,X)+\frac{Z}{p(X)}(Y-g(1,X))-\frac{1-Z}{1-p(X)}(Y-g(0,X))]}{E[m(1,X)-m(0,X)+\frac{Z}{p(X)}(D-m(1,X))-\frac{1-Z}{1-p(X)}(D-m(0,X))]} := \frac{E[a]}{E[b]}$$

Setup

- Consider a score for LATE

$$\psi(W;\theta,\eta) = \overbrace{g(1,X) - g(0,X) + \frac{Z(Y-g(1,X))}{p(X)} - \frac{(1-Z)(Y-g(0,X))}{1-p(X)}}^{a}$$
$$- \theta \times \underbrace{\left(m(1,X) - m(0,X) + \frac{Z(D-m(1,X))}{p(X)} - \frac{(1-Z)(D-m(0,X))}{1-p(X)}\right)}_{b},$$

with

  - low-dimensional parameter vector $\theta \in \Theta$.
  - nuisance parameter $\eta = (g, m, p) \in T$ for a convex set T.
  - specifically, $\eta = (\beta_{11}, \beta_{12}, \beta_{21}, \beta_{22}, \gamma)$.

$$E[\psi(W_i;\theta_0,\eta_0)] = 0.$$

Properties of the score

- Moment condition:

$$\boldsymbol{E}[\psi(W_i;\theta_0,\eta_0)] = \mathbf{0}.$$

- Neyman orthogonality condition:

$$\partial_\eta \mathbf{E}_P \psi(W;\theta_0,\eta_0)[\eta-\eta_0] = \mathbf{0}.$$

High-dimensional QLR test statistic

Step 1: Randomly split the sample $\{1,\cdots,N\}$ into K folds $\{I_1,\cdots,I_K\}$.

Step 2: For each $k\in\{1,\cdots,K\}$, obtain $\widehat{\eta}_k$ by using only the subsample of those observations with indices $i\in\{1,\cdots,N\}\setminus I_k$

1. use lasso logistic regression to estimate $(\widehat{\beta}_{11,k},\widehat{\beta}_{12,k})$,

$$(\widehat{\beta}_{11,k},\widehat{\beta}_{12,k})\in\arg\min_{\beta_{11},\beta_{12}}\mathbb{E}_{I_k^c}[L_1(W_i;\beta_{11},\beta_{12})]+\frac{\lambda_1}{|I_k^c|}\|(\beta_{11},\beta_{12})\|_1,$$

$$L_1(W_i;\beta_{11},\beta_{12})=D_i(Z_i\beta_{11}+X_i'\beta_{12})-\log(1+\exp(Z_i\beta_{11}+X_i'\beta_{12})).$$

2. use lasso logistic regression to estimate $\widehat{\gamma}_k$.
3. use lasso OLS regression to estimate $(\widehat{\beta}_{21},\widehat{\beta}_{22})$,

$$(\widehat{\beta}_{21,k},\widehat{\beta}_{22,k})\in$$
$$\arg\min_{\beta_{21},\beta_{22}}\mathbb{E}_{I_k^c}[(Y_i-Z_i\beta_{21}-X_i'\beta_{22})^2]+\frac{\lambda_3}{|I_k^c|}\|(\beta_{21},\beta_{22})\|_1.$$

High-dimensional QLR test statistic

Step 3:

Compute $\widehat{q}_N(\theta)$ and $\widehat{\Omega}(\theta_1, \theta_2)$ for $\theta_1, \theta_2 \in \Theta$,

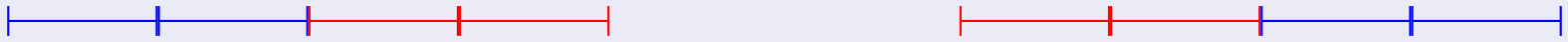
$$\widehat{q}_N(\theta) = \frac{1}{\sqrt{N}} \sum_{k=1}^{K} \sum_{i \in I_k} \psi(W_i; \theta, \widehat{\eta}_k)$$

$$\begin{aligned}\widehat{\Omega}(\theta_1, \theta_2) = & \frac{1}{N} \sum_{k=1}^{K} \sum_{i \in I_k} \psi(W_i; \theta_1, \widehat{\eta}_k) \psi(W_i; \theta_2, \widehat{\eta}_k) \\ & - \frac{1}{N^2} \sum_{k=1}^{K} \sum_{k'=1}^{K} \sum_{i \in I_k, i' \in I_{k'}} \psi(W_i; \theta_1, \widehat{\eta}_k) \psi(W_{i'}; \theta_2, \widehat{\eta}_{k'}).\end{aligned}$$

An illustration of K=2-fold cross-fitting.

I_1 Score I_2 Nuisance

I_1 Nuisance I_2 Score



$\sum_{i \in I_1} \psi(W_i; \theta, \widehat{\eta}_1)$

$\sum_{i \in I_2} \psi(W_i; \theta, \widehat{\eta}_2)$

High-dimensional QLR test statistic

Step 4: Take independent draws $\xi \sim N(0, \widehat{\Omega}(\theta_0, \theta_0))$ and calculate $R = R(\xi, h_N, \widehat{\Omega})$, where

$$R(\xi, h_N, \widehat{\Omega}) = \xi^2 \widehat{\Omega}(\theta_0, \theta_0)^{-1} - \inf_{\theta} (V(\theta)\xi + h_N)^2 \widehat{\Omega}(\theta, \theta)^{-1},$$

with $V(\theta) = \widehat{\Omega}(\theta, \theta_0)\widehat{\Omega}(\theta_0, \theta_0)^{-1}$ and $h_N(\theta) = \widehat{q}_N(\theta) - \widehat{\Omega}(\theta, \theta_0)\widehat{\Omega}(\theta_0, \theta_0)^{-1}\widehat{q}_N(\theta_0)$.

Step 5: Calculate the conditional critical value $c_\alpha(\widetilde{h})$ as

$$c_\alpha(\widetilde{h}) = \min\{c : P(R(\xi, h_N, \widehat{\Omega}) > c | h_N = \widetilde{h}) \leqslant \alpha\}.$$

Step 6: Reject the null hypothesis $H_0 : \theta = \theta_0$ when $R(\xi, h_N, \widehat{\Omega})$ exceeds the $(1 - \alpha)$ quantiles $c_\alpha(h_N)$ and report the $(1 - \alpha)$ confidence interval $CI_\alpha = \{\theta : R(\xi, h_N, \widehat{\Omega}) \leqslant c_\alpha(h_N)\}$.

Main result

The empirical process

$$\mathbb{G}_N(\cdot) = \underbrace{\frac{1}{\sqrt{N}} \sum_{i \in [N]} \big(\psi(W_i; \cdot, \eta_0)}_{q_N(\theta)} - \mathbf{E}_P[\psi(W; \cdot, \eta_0)]\big).$$

Propose an estimator of $\mathbb{G}_N(\cdot)$ as

$$\widehat{\mathbb{G}}_N(\theta) = \underbrace{\sqrt{N}\Big(\frac{1}{N} \sum_{k=1}^{K} \sum_{i \in I_k} \psi(W_i; \theta, \widehat{\eta}_k)}_{\widehat{q}_N(\theta)} - \mathbf{E}_P\left[\psi(W_i; \theta, \widehat{\eta}_k)\right]\Big).$$

Theorem

Suppose some regularity assumptions hold. Under the null, we have

$$\widehat{\mathbb{G}}_N(\theta) = \mathbb{G}_N(\theta) + O_P(N^{-1/2} r'_N).$$

The process $\widehat{\mathbb{G}}_N(\cdot)$ *weakly converges to a centered Gaussian process* $\mathbb{G}(\cdot)$ *over* $\theta \in \Theta$ *with covariance function* $\Omega(\theta_1, \theta_2) = \mathbf{E}_P[(\psi(W; \theta_1, \eta_0) - \mathbf{E}_P[\psi(W; \theta_1, \eta_0)])\,(\psi(W; \theta_2, \eta_0) - \mathbf{E}_P[\psi(W; \theta_2, \eta_0)])]$ *as* $N \to \infty$.

Variance estimation

The variance $\boldsymbol{\Omega(\theta_1, \theta_2)}$ can be consistently estimated uniformly under $\boldsymbol{H_0}$ by

$$\boldsymbol{\widehat{\Omega}(\theta_1, \theta_2) = \frac{1}{N} \sum_{k \in [K]} \sum_{i \in I_k} \psi(W_i; \theta_1, \widehat{\eta}_k) \psi(W_i; \theta_2, \widehat{\eta}_k)}$$
$$\boldsymbol{- \frac{1}{N^2} \sum_{k, k' \in [K]} \sum_{i \in I_k, i' \in I_{k'}} \psi(W_i; \theta_1, \widehat{\eta}_k) \psi(W_{i'}; \theta_2, \widehat{\eta}_{k'})}$$

and $\boldsymbol{\widehat{\Omega}(\theta_1, \theta_2) = \Omega(\theta_1, \theta_2) + O_P(\rho_N)}$ with $\boldsymbol{\rho_N \lesssim \delta_N}$.

Simulation designs

- $X \sim N(0, \Sigma)$ with $\Sigma_{jk} = 0.5^{|j-k|}$
- $N = 500$, $\dim(X) = 5, 100, 300,$ and 500
- compliance class $Q := \begin{cases} 0 & \text{never-taker} \\ 1 & \text{always-taker} \\ 2 & \text{compliers} \end{cases}$
- $P[Q = 2|X] = \frac{\exp(\beta_0+\beta_1 x)}{1+\exp(\beta_0+\beta_1 x)} = \begin{cases} 0.1 & \text{weakly identified case} \\ 0.5 & \text{strongly identified case} \end{cases}$
- $Z = \frac{\exp(\gamma_0+\gamma_1 x)}{1+\exp(\gamma_0+\gamma_1 x)} + v$ with $v \sim N(0, 1) \xRightarrow{s.t.} P(Z = 1) = 0.5$
- $D = Z * \mathbb{1}\{Q = 2\} + Q * \mathbb{1}\{Q \neq 2\}$
- $Y = D + X + \varepsilon$ with $\quad \varepsilon \sim N(0, 1) \quad \Longrightarrow \theta_0 = 1$.

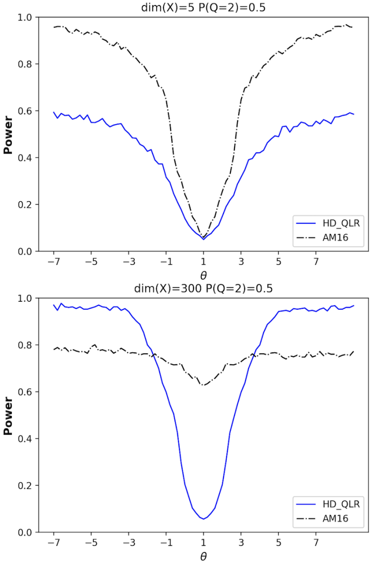
Simulation designs

I compare the proposed method HD-QLR (this paper) with

- the conditional QLR test (AM16) : robust against weak identification but not against high-dimensional setting
- ML methods (CCDDHNR18 and BCFH17): robust against high-dimensional setting but not against weak identification.

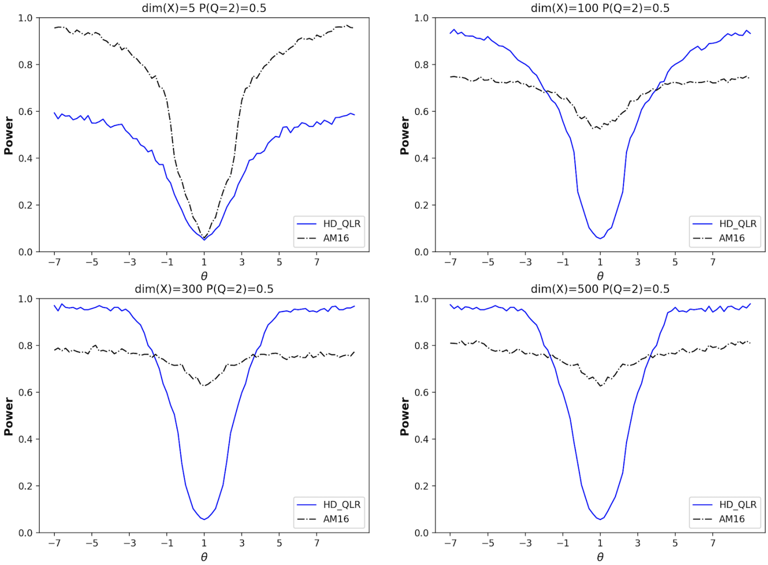
Comparisons: strong identification

- AM16 with HD-QLR (this paper)



Comparisons: weak identification

- AM16 with HD-QLR (this paper)



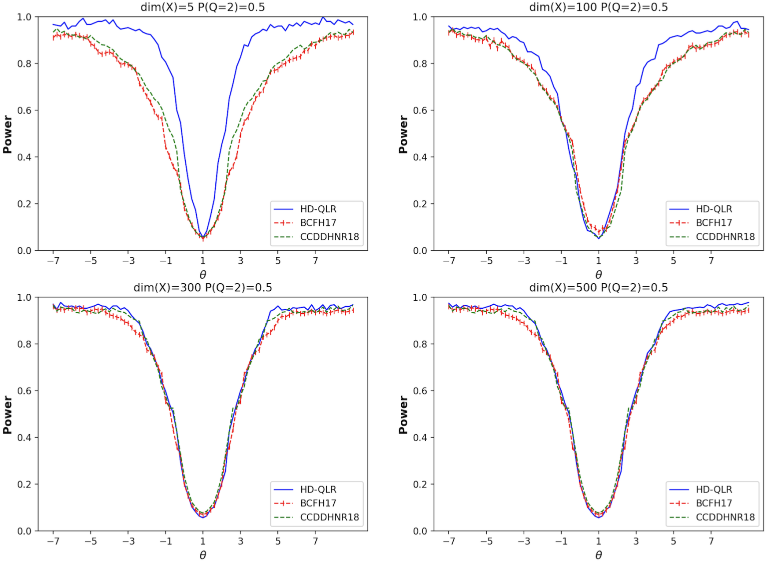
Simulation designs

I compare the proposed method HD-QLR (this paper) with

- the conditional QLR test (AM16) : robust against weak identification but not against high-dimensional setting
- ML methods (CCDDHNR18 and BCFH17): robust against high-dimensional setting but not against weak identification.

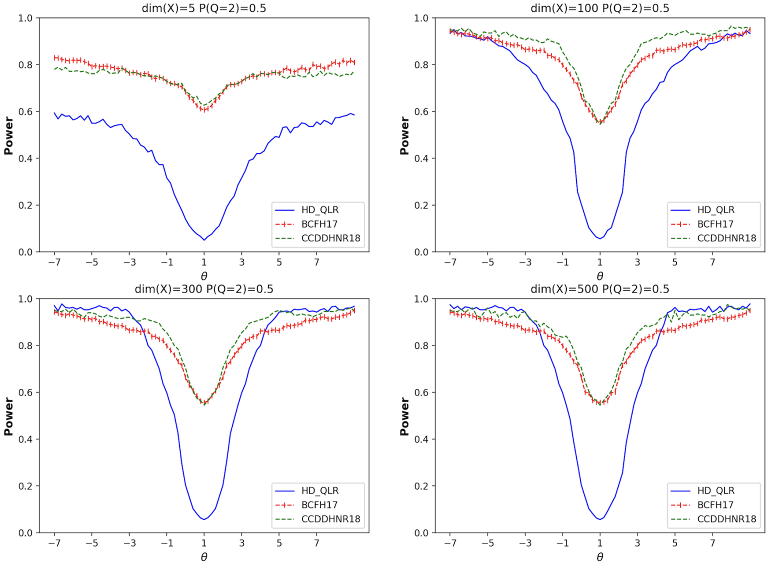
Comparisons: strong identification

- CCDDHNR18 and BCFH17 with HD-QLR (this paper)



Comparisons: weak identification

- CCDDHNR18 and BCFH17 with HD-QLR (this paper)



Revisit Erik Hornung (2015) "Railroads and growth in Prussia"

- Data: highly detailed city-level data from the historical German state of Prussia.
- $\boldsymbol{Y_i}$ **:** urban population growth rate.
- $\boldsymbol{D_i}$ **:** whether the city was connected to the railroad in a given year.
- $\boldsymbol{Z_i}$ **:** whether the city was located within a straight-line corridor between two important cities (nodes).
- $\boldsymbol{X_i}$ **:** whether the city has street access, whether the city has waterway access, military population, age composition, school enrollment rate, etc.

Results

$\boldsymbol{Y_{it}}$: population growth rate	periods						
	49-52	52-55	55-58	58-61	61-64	64-67	67-71
Panel A: AM16							
LATE	0.010	0.020	0.063	0.030	0.037	0.056	0.044
CI	[-0.017, 0.05]	[0.004, 0.039]	[0.030, 0.063]	[0.011, 0.050]	[0.019, 0.050]	[0.012, 0.420]	[0.018, 0.155]
length of CI	0.067	0.035	0.033	0.039	0.031	0.408	0.137
Panel B: CCDDHNR18							
LATE	0.012	0.011	0.007	0.000	0.020	0.012	0.011
CI	[-0.019, 0.039]	[-0.014, 0.035]	[-0.009, 0.044]	[-0.016, 0.030]	[-0.019, 0.052]	[-0.014, 0.039]	[-0.016, 0.036]
length of CI	0.058	0.048	0.053	0.046	0.070	0.052	0.052
Panel C:BCFH17							
LATE	0.009	0.009	0.012	0.006	0.015	0.012	0.013
CI	[-0.009, 0.026]	[-0.006, 0.023]	[-0.007, 0.031]	[-0.008, 0.020]	[-0.009, 0.040]	[-0.018, 0.041]	[-0.008, 0.034]
length of CI	0.035	0.029	0.038	0.028	0.049	0.059	0.042
Panel D: HD-QLR (this paper)							
LATE	0.010	0.011	0.014	0.004	0.018	0.014	0.011
CI	[0.000, 0.021]	[0.002, 0.018]	[0.003, 0.027]	[-0.001, 0.016]	[0.003, 0.029]	[-0.004, 0.032]	[-0.002, 0.023]
length of CI	0.021	0.016	0.024	0.017	0.026	0.033	0.024
Size N	929	924	914	926	924	919	919
dim(X)	212	212	212	212	212	212	212

Takeaways

- I develop a test statistic to make inference for the high-dimensional LATE, independent of the strength of identification.

	Low-dimensional model	High-dimensional model
Strong Identification	t-test,⋯	CCDDHNR18, BDFH17
Weak Identification	AR, S, K, AM16	HD-QLR

- The test has uniformly correct asymptotic size.
- Simulation results indicate that the proposed test is robust against weak identification and high-dimensional settings, outperforming other conventional tests.
- Empirical illustrations show that conventional tests exhibit a positive bias in the length of confidence intervals and lose significance when high-dimensional covariates are taken into account.

Takeaways

- I develop a test statistic to make inference for the high-dimensional LATE, independent of the strength of identification.

	Low-dimensional model	High-dimensional model
Strong Identification	t-test,⋯	CCDDHNR18, BDFH17
Weak Identification	AR, S, K, AM16	HD-QLR

- The test has uniformly correct asymptotic size.
- Simulation results indicate that the proposed test is robust against weak identification and high-dimensional settings, outperforming other conventional tests.
- Empirical illustrations show that conventional tests exhibit a positive bias in the length of confidence intervals and lose significance when high-dimensional covariates are taken into account.

Future work

- Not limited to LATE, extend to general IV estimation.

Thank you!

feel free to email me any comments

yukun.ma@vanderbilt.edu

Motivation: Lee et al. (2022)

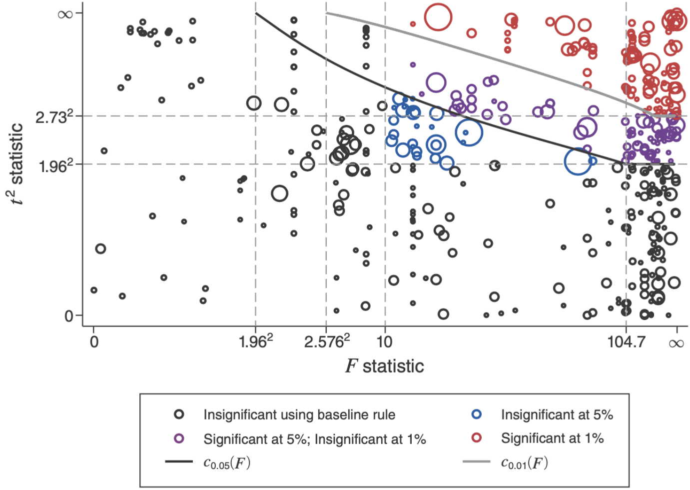


Figure: *American Economic Review* 2013-2019

back