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

Yukun Ma†‡

August 18, 2023

Abstract

This paper investigates the local average treatment effect (LATE) with high-dimensional covariates, irrespective of the strength of identification. We propose a novel test statistic for the high-dimensional LATE, and demonstrate that our test has uniformly correct asymptotic size. By employing the double/debiased machine learning (DML) method to estimate nuisance parameters, we develop easy-to-implement algorithms for inference and confidence interval calculation of the high-dimensional LATE. Simulations indicate that our test is robust against both weak identification and high-dimensional setting concerning size control and power performance, outperforming other conventional tests. Applying the proposed method to railroad and population data to study the effect of railroad access on urban population growth, we observe shorter length of confidence intervals and smaller point estimates for the railroad access coefficients compared to the conventional results.

Keywords: Weak identification, local average treatment effect, double/debiased machine learning, high-dimensional covariates.

* First arXiv date: February 20, 2023

† Yukun Ma: yukun.ma@vanderbilt.edu. Department of Economics, Vanderbilt University, VU Station B #351819, 2301 Vanderbilt Place, Nashville, TN 37235-1819, USA

‡ For their useful comments, I wish to thanks Harold Chiang, Jean-Marie Dufour, Atsushi Inoue, Edward Kennedy, Yuya Sasaki, Luis Carvajal-Osorio, Tucker Smith, and seminar participants at NY Camp Econometrics XVII, 1st CIREQ Interdisciplinary PhD Student Conference on Big Data and Artificial Intelligence, 2023 Asia Meeting of the Econometric Society. All reminding errors are my own.

1 Introduction

In models where certain explanatory variables are correlated with the error term, least squares estimators yield inconsistent coefficient estimates. Instrumental variables (IV) are often employed as a solution, as they are uncorrelated with the error term but correlated with the endogenous explanatory variables. However, when the correlation between the instruments and endogenous variables is weak, IV estimation becomes imprecise, leading to unreliable tests and confidence intervals. This is the weak-instrument problem which remains a significant concern in empirical practice.

Empirical researchers often seek to make inferences about the coefficients of endogenous variables in IV regression. An example is the influential study by Angrist and Krueger (1991), using quarter of birth as an IV to estimate returns from schooling. However, Bound et al. (1995) argue that Angrist and Krueger's results may be unreliable due to the weak correlation between one's quarter of birth and their education attainment. Moreover, the common practice of pretesting, with a rule-of-thumb F-statistic threshold of 10 proposed by Staiger and Stock (1994), is challenged by Lee et al. (2022). In their paper, they introduce the tF critical value function and reveal that achieving a true 5 percent test with critical value of 1.96 instead requires an F exceeding 104.7. Applying this criterion to their sample of 61 *American Economic Review* papers published between 2013 and 2019, they find that one-quarter of the initially presumed statistically significant specifications turn out to be insignificant.

Angrist and Imbens (1995a) develop a framework for estimating the local average treatment effect (LATE). This type of estimate represents the treatment effect for a group of compliers who decide to take the treatment if and only if assigned to the treatment group. Using the IV method to estimate LATE has garnered considerable attention in the literature. In the LATE framework, weak identification arises when instruments are only weakly correlated with endogenous regressors or when the share of compliers is relatively small. Our interest lies in studying and addressing the issue of weak identification in the LATE framework.

The issue of weak instruments has been extensively studied in the literature, leading to the development of various econometric techniques for estimating and making inference about a structure parameter θ based on moment equailities. In particular, many models imply that certain function of the data and model parameters has mean zero when evalueted at the true parameter value θ_0.

Our focus is on testing the hypothesis that the mean function is equal to zero at θ_0. The existing literature proposes numerous tests for this hypothesis, such as Stock and Wright (2000), Kleibergen (2002), and Andrews and Mikusheva (2016). While these papers develop methods aimed at making inference about target parameters in the presence of weak identification, they do not consider models with high-dimensional covariates, which have become increasingly prevalent in today's big-data environment. We differ from the previous work by developing an identification-robust test that employs machine-learning methods, which enable us to explore a wider range of controls than what has been previously considered.

Based on our simulation results, we find that our proposed method outperforms the conventional identification-robust test. While the conventional test shows robustness against weak identification, it suffers from severe size distortion in high-dimensional scenarios. In contrast, our proposed method demonstrates robustness to both weak identification and high dimensionality. Additionally, our proposed method outperforms existing machine-learning methods. Although the conventional machine-learning methods exhibit robustness in high-dimensional settings, they encounter significant size distortion and power loss under weak identification scenarios. Overall, our proposed method exhibits robustness to both weak identification and high-dimensionality in terms of size control and power performance. We demonstrate that our proposed test maintains uniform size control across a broad range of data-generating processes, accommodating both low-dimensional and high-dimensional scenarios, as well as weakly and strongly identified cases. Additionally, in situations where the dimensionality of the covariates in the LATE framework is small, our proposed method coincides with the conventional identification-robust test in terms of size control, but with a slight trade-off only under weakly identified scenarios.

In our empirical illustrations, we employ our proposed method to examine two impacts: (1) the effect of railroad access on city growth, and (2) the effect of cholera-related deaths on rental price. For the first application, our findings demonstrate that, when compared with the conventional identification-robust test, our proposed method tends to yield smaller estimates in magnitude. Moreover, the initial significant results obtained through the conventional identification-robust test tend to lose significance after applying our proposed method. Additionally, when compared with the conventional machine-learning methods, our proposed method yields narrower confidence intervals. In the second application,

we observe identical results, further reinforcing the reliability and consistency of our proposed method.

This paper contributes to the rapidly growing literature on weak identification by providing procedures for inference and the estimation of confidence intervals of LATE parameters in high-dimensional models. To the best of our knowledge, this is the first paper to make inferences about the high-dimensional LATE model, irrespective of the strength of identification. We construct a high-dimensional conditional test statistic with uniformly correct asymptotic size. Furthermore, we provide a practical guideline containing step-by-step algorithms for inference and confidence interval of the high-dimensional LATE using machine learning methods, specifically based on the lasso technique.

1.1 Relations to the Literature

This paper contributes to the literature on weak identification and high-dimensional models by providing a test that can be used to make inferences for the LATE with high-dimensional covariates.

Since the 1990s, weak identification in the IV context has received considerable attention in the literature.[1] To test the mean function is equal to zero at the true parameter value θ_0, Stock and Wright (2000) propose the concepts of weakly identified Generalized Mothod of Moments (GMM) and introduce the S statistic in the quadratic form of the objective function, which is a generalized form of the Anderson-Rubin test statistic and follows a χ^2 asymptotic distribution under the null hypothesis. Kleibergen (2005) proposes the K statistic, using the asymptotic independence between the Jacobian estimator of the objective function and the sample average of the moment. However, these tests have low power under weak identification settings, as they only study the process local to the point θ_0 and ignore a significant amount of information.

To address this issue, Moreira (2003) proposes the conditional likelihood ratio test for weakly identified linear IV models based on the conditional distribution of nonpivotal statistics. This test of structural coefficients has improved power relative to previous tests when identification is weak. More recently, Andrews and Mikusheva (2016) have developed conditional test statistics to test the hypoth-

[1] See Staiger and Stock (1994), Bound et al. (1995), Stock and Wright (2000), Kleibergen (2002), Stock and Yogo (2002), Moreira (2003), Kleibergen (2005), Andrews et al. (2006), Moreira (2009), Andrews and Mikusheva (2016), Andrews and Guggenberger (2019), Moreira and Moreira (2019), and Mikusheva and Sun (2022). See Stock et al. (2002), Andrews and Stock (2005), and Andrews et al. (2019) for surveys of weak identification literature.

esis that θ_0 satisfies the moment condition without making any assumptions about point identification or the strength of identification. Their approach has desirable power properties since the test depends on the full path of the observed process without losing information. However, none of these papers considers models with high-dimensional covariates typically found in today's big-data environment.

In the past decade, there has been a surge in the literature on machine learning-based econometric methods for high-dimensional models, in which the dimensionality of parameters is potentially much larger than the sample size of available data ($p \gg N$). Belloni et al. (2015) propose a Neyman orthogonal score for a class of Z-estimation framework in the presence of high-dimensional nuisance parameters. Belloni et al. (2018) construct a confidence interval using the Neyman orthogonality condition in the high-dimensional setting. Chernozhukov et al. (2013,2016,2017) derive the Central Limit Theorem (CLT) for the high-dimensional model using the Gaussian approximation approach. Belloni et al. (2014) provide an overview of the methods to estimate and make inferences for high-dimensional data. Chernozhukov et al. (2018) introduces the double/debiased machine learning (DML) method under the i.i.d setting. They combine the Neyman orthogonality condition [2] and cross-fitting methods. More recently, Chernozhukov et al. (2022) provide a general construction of doubly robust moment function with robustness to nonparametric or high-dimensional first steps. However, none of these papers on high-dimensional models consider weak identification issues.

This paper also relates to the literature on IV estimation of LATE. Angrist and Imbens (1995a) first introduce the simple IV estimand for the average treatment effect for compliers. Motivated by Angrist and Krueger (1991), Angrist and Imbens (1995b) extend LATE to ordered treatments, such as years of schooling. Subsequent researchers start to incorporate covariates for estimating LATE, including Angrist et al. (2000), Hirano et al. (2000), Yau and Little (2001), and Abadie (2003), employing either parametric or semiparametric estimation approaches. Tan (2006) proposes an LATE estimator with robustness against the misspecification of either propensity score model or the outcome regression model. Frölich (2007) provides a fully nonparametric $\sqrt{N}$-consistent and efficient estimator for the LATE with confounding covariates. More recently, Belloni et al. (2017) introduces an efficient estimator and reliable confidence bands for the LATE with nonparametric/high-dimensional

[2] We refer readers to Pfanzagl and Wefelmeyer (1985), Bickel et al. (1993), Newey (1994), and Tsiatis (2006) for the development of the Neyman orthogonal score.

components using the orthogonal moment condition and machine-learning method. Angrist (2022) employs empirical examples to illustrate the importance of the LATE framework for causal inference.[3]

In this paper, we use the doubly robust estimand of LATE as our target parameter. To the best of our knowledge, this paper is the first to develop a method for the LATE with high-dimensional covariates without any assumption about identification.

1.2 Outline

The rest of the paper is structured as follows. In Section 2, a practice guideline of the proposed method and algorithm is given. In Section 3, we present the theoretical framework, including a justification explanation of our proposed method in Section 3.1, followed by the presentation of the general weak convergence result in Section 3.2. Section 3.3 contains the low-level sufficient conditions for the LATE framework. In Section 4, we showcase our Monte Carlo simulation results. In Section 5, two empirical illustrations are given. We conclude in Section 6. The appendix includes all proofs of the theorems and lemmas.

2 Overview

In this section, we provide a brief overview of our proposed method without theories. This overview serves as a concise guideline in practice. In Section 3, we will formally introduce the theoretical rationale for our method.

2.1 Notation

Consider the standard IV setup, the researcher observes a dataset of N iid observations, denoted as $\{W_i = (Y_i, D_i, Z_i, X_i')\}_{i=1}^N$. The outcome of interest for unit i is denoted by Y_i. Let $D_i \in \{0, 1\}$ be a binary indicator of the receipt of treatment. We define X_i as a set of p-dimensional controls with the dimensionality p potentially much larger than the available sample size, N. Additionally, there

[3]See, e.g., Card (1993), Kane and Rouse (1993), Acemoglu and Angrist (2000) Kling (2006), Oreopoulos (2006), Angrist et al. (2010), Galiani et al. (2011), Maestas et al. (2013), French and Song (2014), Dahl et al. (2014), Moser et al. (2014), Aizer and Doyle Jr (2015), Bisbee et al. (2017), Benzell and Cooke (2021) for empirical literature that employ IV for LATE estimation.

exists an instrument variable Z_i, which is also a binary indicator, such as the offer of treatment. This instrument is randomly assigned conditional on the covariates.

Let $\{\mathcal{P}_N\}_N$ be a sequence of sets of probability law of $\{W_i\}_i$. The analysis allows for an increasing dimensionality of W_i as the sample size N grows. We denote $P = P_N \in \mathcal{P}_N$ as the law with respect to sample size N, and E_P represents the expectation for law P. For any set B, its complement set is denoted as $B^c = \{1, \cdots, N\} \setminus B$, and $|B|$ represents the cardinality of the set B. Finally, we introduce the subsample expectation operator by $\mathbb{E}_B[\cdot] := \frac{1}{|B|}\sum_{i\in B}[\cdot]$.

2.2 Anderson-Rubin-Type Neyman Orthogonal Score for the LATE

We model the random vector $W = (Y, D, Z, X')'$ as follows,

$$D = m_0(Z, X) + v, \quad \mathrm{E}_P[v|Z, X] = 0, \quad \text{(First stage)} \tag{2.1}$$

$$Y = g_0(Z, X) + u, \quad \mathrm{E}_P[u|Z, X] = 0, \quad \text{(Reduced form)} \tag{2.2}$$

$$Z = p_0(X) + e, \quad \mathrm{E}_P[e|X] = 0, \quad \text{(Propensity score)} \tag{2.3}$$

where the function m_0 maps the support of (Z, X) to $(\varepsilon, 1-\varepsilon)$, the function g_0 maps the support of (Z, X) to $\mathbb{R}$, and the function p_0 maps the support of X to $(\varepsilon, 1-\varepsilon)$ for some $\varepsilon \in (0, 1/2)$.

The LATE proposed by Tan (2006)[4] is given by

$$\text{LATE} = \frac{\mathrm{E}_P[g_0(1,X) - g_0(0,X)] + \mathrm{E}_P\left[\frac{Z(Y-g_0(1,X))}{p_0(X)}\right] - \mathrm{E}_P\left[\frac{(1-Z)(Y-g_0(0,X))}{1-p_0(X)}\right]}{\mathrm{E}_P[m_0(1,X) - m_0(0,X)] + \mathrm{E}_P\left[\frac{Z(D-m_0(1,X))}{p_0(X)}\right] - \mathrm{E}_P\left[\frac{(1-Z)(D-m_0(0,X))}{1-p_0(X)}\right]} := \frac{\alpha_{01}}{\alpha_{02}},$$

where α_{01} and α_{02} correspond to the numerator and denominator, respectively. The numerator is the intent-to-treat (ITT) effect, while the denominator is the compliance probability or the share of compliers. The usual normal distribution of the LATE estimator can be obtained using the delta method, which linearizes the LATE estimator with respect to the estimators $(\widehat{\alpha}_{01}, \widehat{\alpha}_{02})$. Following the weak IV literature, we model weak identification by allowing the denominator α_{02} to be close to zero, which corresponds to the case where the share of the compliers is small. Notably, in section

[4]This LATE estimand, termed doubly robust LATE in Tan (2006), is robust against the misspecification of either propensity score or the outcome regression model. The Newman orthogonal score ψ in (2.4) coincides with the double robust score in the context of LATE. However, in this paper, we only focus on Newman orthogonality, excluding double robustness exploration.

3, we also accommodate the denominator α_{02} being exactly zero, which corresponds to completely unidentified case. Then, the normal approximation fails in the weak identification setting because the LATE estimator is highly nonlinear in $\widehat{\alpha}_{02}$ when $\widehat{\alpha}_{02}$ is very close to zero. In order to construct valid hypothesis tests and confidence sets for LATE without considering the strength of identification, we construct the function ψ by

$$\begin{aligned}\psi(W;\theta,\eta) = g(1,X) - g(0,X) + \frac{Z(Y-g(1,X))}{p(X)} - \frac{(1-Z)(Y-g(0,X))}{1-p(X)} \\ - \theta \times \left(m(1,X) - m(0,X) + \frac{Z(D-m(1,X))}{p(X)} - \frac{(1-Z)(D-m(0,X))}{1-p(X)} \right),\end{aligned} \tag{2.4}$$

where $W = (Y, D, Z, X')'$, $\theta \in \Theta$ is the LATE, and $\eta = (g, m, p) \in \mathcal{T} \subset \mathcal{R}^{d_\eta}$ are the nuisance parameters. Note that this function ψ is an Anderson-Rubin-type Neyman orthogonal (beriefly mention why is orthogonal, not mention target parameter here) score function for the model (2.1)-(2.3). Note that the score ψ satisfies the moment condition $\mathrm{E}_P[\psi(W;\theta_0,\eta_0)] = 0$, where θ_0 and η_0 represent the true values of θ and η, respectively.

2.3 Inference Procedure

We next introduce how to make inferences about the make sure target have been mentionedtarget parameter θ in practice. We are interested in testing the null hypothesis $\theta = \theta_0$. We first estimate the first-stage nuisance parameters η by some machine learning methods. With a fixed positive integer $K > 1$, we randomly partition $\{1, \cdots, N\}$ into K parts $\{I_k\}_{k=1}^K$. For each $k \in \{1, \cdots, K\}$, the nuisance parameter estimate $\widehat{\eta}_k$ is computed using the subsample of those observations with index $i \in I_k^c$. After that, we apply the cross-fitting (data-splitting) method proposed by Chernozhukov et al. (2018) to calculate the covariance estimator of the process $\sqrt{N}\psi(W_i;\cdot,\eta_0)$, which is expressed as

$$\widehat{\Omega}(\theta_1,\theta_2) = \frac{1}{N}\sum_{k=1}^{K}\sum_{i\in I_k}\psi(W_i;\theta_1,\widehat{\eta}_k)\psi(W_i;\theta_2,\widehat{\eta}_k) - \frac{1}{N^2}\sum_{k=1}^{K}\sum_{k'=1}^{K}\sum_{i\in I_k, i'\in I_{k'}}\psi(W_i;\theta_1,\widehat{\eta}_k)\psi(W_{i'};\theta_2,\widehat{\eta}_{k'}), \tag{2.5}$$

for $\theta_1, \theta_2 \in \Theta$. Note that $\widehat{\Omega}(\theta_1,\theta_2)$ is computed using the sample of those observations with index $i \in I_k$. This computation is repeated K times. Following this procedure, we take random draws $\xi \sim N(0, \widehat{\Omega}(\theta_0,\theta_0))$ under the null. Then under the null, we calculate a conditional test statistic

$R(\xi, h, \widehat{\Omega})$ given $h = h_N(\theta)$, where

$$R(\xi, h, \Omega) = \xi^2 \Omega(\theta_0, \theta_0)^{-1} - \inf_{\theta} (V(\theta)\xi + h)^2 \Omega(\theta, \theta)^{-1}, \text{ and} \tag{2.6}$$

$$h_N(\theta) = \widehat{q}_N(\theta) - \widehat{\Omega}(\theta, \theta_0)\widehat{\Omega}(\theta_0, \theta_0)^{-1}\widehat{q}_N(\theta_0), \tag{2.7}$$

with $V(\theta) = \Omega(\theta, \theta_0)\Omega(\theta_0, \theta_0)^{-1}$ and $\widehat{q}_N(\theta) = \frac{1}{\sqrt{N}} \sum_{k=1}^{K} \sum_{i \in I_k} \psi(W_i; \theta, \widehat{\eta}_k)$. After that, the conditional critical value $c_\alpha(\widetilde{h})$ is defined as

$$c_\alpha(\widetilde{h}) = \min\{c : P(R(\xi, h_N, \widehat{\Omega}) > c | h_N = \widetilde{h}) \leq \alpha\}. \tag{2.8}$$

Note that given any realization of $h_N(\cdot)$, the critical value $c_\alpha(\cdot)$ can be easily calculated.

We specifically examine a logit model class in which a binary outcome D_i, denoting an individual i's receipt of treatment, is determined by the treatment offer, Z_i, and a set of p-dimensional covariates, X_i. Additionally, we employ the logit model to estimate the propensity score and estimate the outcome regression through linear regression analysis. We present the models as follows,

$$\begin{aligned}
&\mathrm{E}_P[D_i | Z_i, X_i] = \Lambda(Z_i \beta_{11}^0 + X_i' \beta_{12}^0), \\
&\mathrm{E}_P[Z_i | X_i] = \Lambda(X_i' \gamma^0), \\
&\mathrm{E}_P[Y_i | Z_i, X_i] = Z_i \beta_{21}^0 + X_i' \beta_{22}^0,
\end{aligned}$$

where Λ denotes the logistic CDF defined by $\Lambda(t) = \exp(t)/(1 + \exp(t))$ for all $t \in \mathbb{R}$, and the true nuisance parameters vector $\eta_0 = (\beta_{11}^0, \beta_{12}^0, \beta_{21}^0, \beta_{22}^0, \gamma^0)$. The log-likelihood functions of the logit model are $L_1(\beta_{11}, \beta_{12}) = \mathbb{E}_N[L_1(W_i; \beta_{11}, \beta_{12})]$ and $L_2(\gamma) = \mathbb{E}_N[L_2(W_i; \gamma)]$, where $L_1(W_i; \beta_{11}, \beta_{12}) = D_i(Z_i \beta_{11} + X_i' \beta_{12}) - \log(1 + \exp(Z_i \beta_{11} + X_i' \beta_{12}))$ and $L_2(W_i; \gamma) = Z_i X_i' \gamma - \log(1 + \exp(X_i' \gamma))$. The score for LATE is then specified as

$$\begin{aligned}
\psi(W_i; \theta, \eta) = \beta_{21} &+ \frac{Z_i(Y_i - \beta_{21} - X_i' \beta_{22})}{\Lambda(X_i' \gamma)} - \frac{(1 - Z_i)(Y_i - X_i' \beta_{22})}{1 - \Lambda(X_i' \gamma)} \\
&- \theta \times \left[\Lambda(\beta_{11} + X_i' \beta_{12}) - \Lambda(X_i' \beta_{12}) + \frac{Z_i(D_i - \Lambda(\beta_{11} + X_i' \beta_{12}))}{\Lambda(X_i' \gamma)} - \frac{(1 - Z_i)(D_i - \Lambda(X_i' \beta_{12}))}{1 - \Lambda(X_i' \gamma)} \right].
\end{aligned} \tag{2.9}$$

It is essential to highlight that in the score, the logit model can be easily substituted with other models like probit model or linear regression. We present a concrete inference procedure as the following algorithm. While we specifically base our algorithm on lasso, for the sake of clarity, it is

worth noting that other machine learning methods can also be used as a substitute for lasso. Suppose that we have some generic penalty tuning parameter λ_1, λ_2, and λ_3. Formal and theoretical justified choice of these items are delayed to Lemma 2 and 3 in Appendix A.

Algorithm 1. *(K-fold DML for high-dimensional LATE with Lasso)*

Step 1. Randomly split the sample with size N into K folds $(I_k)_{k=1}^K$.

Step 2. For each $k \in \{1, \cdots, K\}$, obtain the nuisance parameter estimate by lasso:

(a) obtain an lasso logistic estimate $(\widehat{\beta}_{11}, \widehat{\beta}_{12})$ of the nuisance parameter by using only the subsample of those observations with indices $i \in \{1, \cdots, N\} \setminus I_k$,

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

(b) obtain an lasso logistic estimate $\widehat{\gamma}$ of the nuisance parameter by using only the subsample of those observations with indices $i \in \{1, \cdots, N\} \setminus I_k$,

$$\widehat{\gamma}_k \in \arg\min_{\gamma} \mathbb{E}_{I_k^c}[L_2(W_i; \gamma)] + \frac{\lambda_2}{|I_k^c|}\|\gamma\|_1.$$

(c) obtain an lasso OLS estimate $(\widehat{\beta}_{21}, \widehat{\beta}_{22})$ of the nuisance parameter by using only the subsample of those observations with indices $i \in \{1, \cdots, N\} \setminus I_k$,

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

Step 3. Compute $\widehat{\Omega}(\theta_0, \theta_0)$ where $\widehat{\Omega}$ is defined in equation (2.5) with $\widehat{\eta}_k = (\widehat{\beta}_{11,k}, \widehat{\beta}_{12,k}, \widehat{\beta}_{21,k}, \widehat{\beta}_{22,k}, \widehat{\gamma}_k)$ and $\psi(W; \theta, \eta)$ is defined in equation (2.9).

Step 4. We take independent draws $\xi^* \sim N(0, \widehat{\Omega}(\theta_0, \theta_0))$ and calculate $R^* = R(\xi^*, h_N, \widehat{\Omega})$ by the definition in equation (2.6), which represents a random draw from the conditional distribution of R given h_N under the null.

Step 5. Given the critical value defined in equation (2.8), we 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 $CI_\alpha = \{\theta : R(\xi^*, h_N, \widehat{\Omega}) \leq c_\alpha(h_N)\}$.

Remark 1. *Andrews and Mikusheva (2016) develop a conditional inference approach for moment condition models that does not reply on any assumptions about identification. Their proposed conditional quasi-likelihood ratio (QLR) tests possess uniformly correct size for a wide range of models. Nonetheless, their test statistic may not be applied to certain high-dimensional research designs, such as the LATE model with high-dimenisonal covariates. To address this limitation, we employ machine learning methods to handle the high-dimensional covariates in the possible models, and specify the score in Andrews and Mikusheva (2016) as our score for LATE. To the best of our knowledge, our method is the first to provide inference for the LATE model with high-dimensional covariates, without imposing any assumptions regarding the strength of identification. Furthermore, our method can be easily extended to enable inferences for other high-dimensional models, without replying on any point identification assumption.*

3 Theory

3.1 Definition of the High-dimensional QLR test

In this section, we provide a precise definition of our proposed method, termed the high-dimensional QLR test. To begin with, we formulate the score function $\psi(W;\theta,\eta)$ that satisfies the moment restriction,

$$\mathrm{E}_P[\psi(W;\theta_0,\eta_0)] = 0, \tag{3.1}$$

where θ_0 and η_0 denote the true values of the target parameter θ and the nuisance parameter η, respectively. The nuisance parameter η may be finite-, high-, or infinite-dimensional.

Let us define the Gateaux derivative by $D_r[\eta-\eta_0] := \partial_r\{E_p[\psi(W;\theta_0,\eta_0+r(\eta-\eta_0))]\}$ for $r \in [0,1)$. We say that the score ψ satisfies the Newman orthogonality condition if the pathwise derivative $D_r[\eta-\eta_0]$ exists for all $r \in [0,1)$ and $\eta \in \mathcal{T}_N$, where $\mathcal{T}_N$ is a nuisance realization set with $\mathcal{T}_N \subset \mathcal{T}$, and the Gateaux derivative operator with respect to η vanishes when evaluated at the true parameter values:

$$\partial_\eta \mathrm{E}_P\psi(W;\theta_0,\eta_0)[\eta-\eta_0] = 0, \tag{3.2}$$

for all $\eta \in \mathcal{T}_N$. The Newman orthogonality condition in (3.2) implies that the moment condition $\mathrm{E}_P[\psi(W;\theta_0,\eta_0)] = 0$ remains insensitive to local perturbations of η in a neighborhood of η_0. It is worth noting that the score for LATE, as given in equation (2.4), satisfies both the moment condition (3.1) and the Neyman orthogonality condition (3.2). Newman orthogonality condition has a long history in statistics and econometrics. Newey (1990,1994), Andrews (1994), Robins and Rotnitzky (1995), Linton (1996) make use of this condition in semiparametric models.

Let us define $q_N(\theta) = N^{-1/2}\sum_{i=1}^{N}\psi(W_i;\theta,\eta_0)$ and $S_N(\cdot) = \mathrm{E}_P[q_N(\cdot)]$. Under no assumption of identification for the parameter θ, the null hypothesis $H_0 : \theta = \theta_0$ can be reformulated as testing $S_N(\theta) = 0$. Here, $S_N(\theta)$ represents an infinite-dimensional nuisance function for $\theta \neq \theta_0$. Let $\mathcal{S}_N$ be the set of functions $S_N(\cdot)$ that may arise in our model, and let $\mathcal{S}_0$ be the subset of $\mathcal{S}_N$ that contains functions satisfying $S_N(\theta_0) = 0$. hence, $H_0 : \theta = \theta_0$ implies our new null hypothesis $H_0' : S_N \in \mathcal{S}_0$, which we refer to from now on as our null hypothesis. With these notations, let us construct an empirical process $\mathbb{G}_N(\cdot)$ as

$$\mathbb{G}_N(\cdot) = q_N(\cdot) - S_N(\cdot) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left\{\psi(W_i;\cdot,\eta_0) - \mathrm{E}_P[\psi(W;\cdot,\eta_0)]\right\}. \tag{3.3}$$

In Section 3, we will show that under mild conditions, the process $\mathbb{G}_N(\cdot)$ weakly converges to $\mathbb{G}(\cdot)$ as $N \to \infty$ under the null, where $\mathbb{G}(\cdot)$ is a mean-zero Gaussian process with covariance function $\Omega(\theta_1,\theta_2) = \mathrm{E}_P[\mathbb{G}(\theta_1)\mathbb{G}(\theta_2)]$. Consider the process

$$h_N(\theta) = q_N(\theta) - \widehat{\Omega}(\theta,\theta_0)\widehat{\Omega}(\theta_0,\theta_0)^{-1}q_N(\theta_0), \tag{3.4}$$

where $\widehat{\Omega}(\cdot,\cdot)$ is a consistent estimator of $\Omega(\cdot,\cdot)$. By rearranging equation (3.4), we obtain

$$q_N(\theta) = h_N(\theta) + \widehat{\Omega}(\theta,\theta_0)\widehat{\Omega}(\theta_0,\theta_0)^{-1}q_N(\theta_0). \tag{3.5}$$

Note that the process $q_N(\cdot)$ can be decomposed into two independent random components, the process $h_N(\cdot)$ and $q_N(\theta_0)$. As the distribution of $q_N(\theta_0)$ follows $N(0,\Omega(\theta_0,\theta_0))$ and does not depend on the nuisance function S_N, the conditional distribution of any functions of $q_N(\cdot)$ given $h_N(\cdot)$, under the null hypothesis, remains independent of S_N. To test the null hypothesis $S_N \in \mathcal{S}_0$, a statistic $R = R(q_N(\theta),\Omega)$ can be employed. Importantly, the conditional distribution of the statistic $R(q_N(\cdot),\Omega)$ given $h_N(\cdot)$ does not depend on $S_N(\cdot)$. Therefore, this approach is applicable to both strongly and

weakly identified cases, as it does not require any assumption about their behavior. The test statistic R is akin to the conditional QLR test statistic proposed by Andrews and Mikusheva (2016). However, it is worth noting that their test cannot to be applied to high-dimensional models. In our paper, we specify the score in Andrews and Mikusheva (2016) as the score for LATE in equation (2.9).

In light of the non-applicability of the conditional QLR test under the high-dimensional model, we now propose a novel test that utilizes the DML method. After obtaining the nuisance parameter estimates $\widehat{\eta}_k$ using lasso with observations indexed by $i \in \{1, \cdots, N\} \setminus I_k$, we compute certain transformations of the score using observations indexed by $i \in I_k$. In the rest of this section, we will specify several estimators and the confidence interval specifically designed for the high-dimensional LATE.

We propose an estimator of $\mathbb{G}_N(\theta)$ as

$$\widehat{\mathbb{G}}_N(\theta) = \sqrt{N} \left\{ \frac{1}{N} \sum_{k=1}^{K} \sum_{i \in I_k} \psi(W_i; \theta, \widehat{\eta}_k) - \mathrm{E}_P \left[\psi(W_i; \theta, \widehat{\eta}_k) \right] \right\}. \tag{3.6}$$

Note that $\widehat{\mathbb{G}}_N(\theta)$ is computed using the sample of those observations with index $i \in I_k$. This computation is repeated K times. An estimator of $q_N(\theta)$ is given by $\widehat{q}_N(\theta) = N^{-1/2} \sum_{k=1}^{K} \sum_{i \in I_k} \psi(W_i; \theta, \widehat{\eta}_k)$. We propose a uniformly consistent estimator of $\Omega(\theta_1, \theta_2)$ for any $\theta_1, \theta_2 \in \Theta$ as

$$\widehat{\Omega}(\theta_1, \theta_2) = \frac{1}{N} \sum_{k=1}^{K} \sum_{i \in I_k} \psi(W_i; \theta_1, \widehat{\eta}_k) \psi(W_i; \theta_2, \widehat{\eta}_k) - \frac{1}{N^2} \sum_{k=1}^{K} \sum_{k'=1}^{K} \sum_{i \in I_k, i' \in I_{k'}} \psi(W_i; \theta_1, \widehat{\eta}_k) \psi(W_{i'}; \theta_2, \widehat{\eta}_{k'}). \tag{3.7}$$

Subsequently, we propose a test statistic $R(\widehat{q}_N(\theta_0), h_N, \widehat{\Omega})$, where

$$R(\xi, h, \Omega) = \xi^2 \Omega(\theta_0, \theta_0)^{-1} - \inf_{\theta} (V(\theta)\xi + h)^2 \Omega(\theta, \theta)^{-1}, \tag{3.8}$$

with $V(\theta) = \Omega(\theta, \theta_0)\Omega(\theta_0, \theta_0)^{-1}$. Then we define the conditional critical value $c_\alpha(\widetilde{h})$ by

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

Finally, the $(1 - \alpha)$ confidence interval is

$$CI_\alpha = \{\theta : R(\widehat{q}_N(\theta_0), h_N, \widehat{\Omega}) \leq c_\alpha(h)\}.$$

3.2 General Weak Convergence Result

In this section, we present formal theory supporting the process in Section 2 works. We show that our test has uniformly correct asymptotic size.

To simplify the notation, let us first fix some terms. For any finite-dimensional vector δ, we define the l_1-norm by $\|\delta\|_1$, l_2-norm by $\|\delta\|$, l_∞-norm by $\|\delta\|_\infty$, and l_0-seminorm by $\|\delta\|_0$, which represents the number of non-zero components of δ. The sample expectation operator is written as $\mathbb{E}_N[\cdot] = \frac{1}{N}\sum_{i=1}^N[\cdot]$. We use $\|x_{ij}'\delta\|_{2,N}$ to indicate the prediction norm of δ, namely, $\|x_{ij}'\delta\|_{2,N} = \sqrt{\mathbb{E}_N[(x_{ij}'\delta)^2]}$. For any matrix A, $\|A\|$ denotes the ℓ_2-norm of the matrix. Let $c_0 > 0, c_1 > 0, q \geq 4$ be some finite constants with $c_0 \leq c_1$. Let $K \geq 2$ be a fixed integer. Let $\{\delta_N\}_{N=1}^\infty$ be a sequence of positive constants approaching 0, such that $\delta_N \geq N^{-1/2}$. Let $\{a_N\}_{N\geq 1}$, $\{v_N\}_{N\geq 1}$, and $\{K_N\}_{N\geq 1}$ be some sequences of positive constants, possibly growing to infinity, where $v_N \geq 1$ for all $N \geq 1$. We use $a \lesssim b$ to mean $a \leq cb$ for some $c > 0$ that does not depend on N.

We focus on the cases with linear Neyman orthogonal score ψ of the form

$$\psi(w;\theta,\eta) = \psi^a(w;\eta)\theta + \psi^b(w;\eta), \text{ for all } w \in \text{supp}(W). \tag{3.9}$$

Assume that $\theta \in \Theta$, where Θ is a compact set in $\mathcal{R}$, and that $\eta \in \mathcal{T}$, a convex set in the norm space equipped with a norm $\|\cdot\|_e$. With these notations, we consider the following two assumptions.

Assumption 1. *For $N \geq 3$ and $P \in \mathcal{P}_N$, the following conditions hold.*

(i) The true parameter θ_0 satisfies equation (3.1).

(ii) The map $\eta \mapsto \mathrm{E}_P[\psi(W;\theta,\eta)]$ is twice continuously Gateaux-differentiable on the realization set $\mathcal{T}_N$.

(iii) ψ satisfies the Newman orthogonality condition (3.2).

(iv) The score ψ is linear in the sense of (3.9).

(v) Θ is a compact set.

Assumption 2. *For all $N \geq 3$ and $P \in \mathcal{P}_N$, the following conditions hold.*

(i) Given a random subset I of $\{1,\cdots,N\}$ with size $n = N/K$, the nuisance parameter estimator $\widehat{\eta} = \widehat{\eta}\,((W_i)_{i\in I^c})$ belongs to the realization set $\mathcal{T}_N$ with probability at least Δ_N, where $\mathcal{T}_N$ contains η_0 and satisfies the following conditions.

(ii) The following conditions on the rates m_N, m'_N, r'_N hold:

$$(a) \quad m_N := \sup_{\eta \in \mathcal{T}_N, \theta \in \Theta} (\mathrm{E}_P[(\psi(W;\theta,\eta))^q])^{1/q} \leq c_1,$$

$$(b) \quad m'_N := \sup_{\eta \in \mathcal{T}_N} (\mathrm{E}_P[(\psi^a(W;\eta))^q])^{1/q} \leq c_1,$$

$$(c) \quad r'_N := \sup_{\eta \in \mathcal{T}_N, \theta \in \Theta} (\mathrm{E}_P[(\psi(W;\theta,\eta) - \psi(W;\theta,\eta_0))^2])^{1/2} \leq \delta_N.$$

(iii) $\sup_{\theta \in \Theta} \mathrm{E}_P[\psi(W;\theta,\eta_0)^2] \geq c_0$.

Remark 2. *Assumptions 1, 2 are related to Assumptions 3.1, 3.2 in Chernozhukov et al. (2018). We emphasize that Assumption 3.1 (e) in Chernozhukov et al. (2018) serves as the identification condition in their paper, ensuring that the denominator in the LATE is always greater than zero. However, in our paper, we intentionally remove this assumption to accommodate the weak identification issue, allowing for the possibility of a zero denominator, corresponding to a completely unidentified case. In order to derive the uniform convergence of the Gaussian process $\widehat{\mathbb{G}}_N$, we impose restrictions over $\theta \in \Theta$ in Assumption 2 (ii)-(iii).*

Assumption 1 stipulates that the score satisfies the moment condition, Neyman orthogonality condition, and a mild smoothness condition. Assumption 2 is a mild regularity condition. Assumption 2(i)(ii) assert that the estimator of the nuisance parameter $\widehat{\eta}$ belongs to a shrinking neighbourhood of the true nuisance parameter η_0 and contracts around η_0 at a rate of r'_N for all $\theta \in \Theta$. Assumption 2 (iii) ensures a non-degenerate limit distribution. While these conditions are high-level, we will provide more specific low-level conditions in the context of LATE in section 3.3.

Theorem 1. *Suppose Assumptions 1 and 2 hold. For $\theta \in \Theta$, we have*

$$\widehat{\mathbb{G}}_N(\theta) = \mathbb{G}_N(\theta) + O_P(N^{-1/2} r'_N), \tag{3.10}$$

where recall that

$$\mathbb{G}_N(\theta) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \{\psi(W_i;\theta,\eta_0) - \mathrm{E}_P[\psi(W;\theta,\eta_0)]\}, \quad \textit{and}$$

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

The process $\widehat{\mathbb{G}}_N(\cdot)$ weakly converges to a centered Gaussian process $\mathbb{G}(\cdot)$ over Θ with covariance function $\Omega(\theta_1, \theta_2) = \mathrm{E}_P[(\psi(W;\theta_1,\eta_0) - \mathrm{E}_P[\psi(W;\theta_1,\eta_0)])(\psi(W;\theta_2,\eta_0) - \mathrm{E}_P[\psi(W;\theta_2,\eta_0)])]$ as N goes to infinity. Moreover, $\Omega(\theta_1, \theta_2)$ can be consistently estimated by

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

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

Remark 3. *Theorem 1 serves as an extension of Chernozhukov et al. (2018). They propose the pointwise convergence of the target parameter estimator $\widehat{\theta}$ and variance estimator for the DML estimator. We extend their result and show that our proposed empirical process uniformly converges to a Gaussian process over Θ, and our variance estimator $\widehat{\Omega}(\theta_1,\theta_2)$ is a uniformly consistent estimator of Ω over Θ. The weak convergence result enables us to handle the weak identification issues effectively.*

3.3 Lower-level Sufficient Conditions in the LATE framework

In this subsection, we provide lower-level sufficient conditions that guarantee Theorem 1 in the application to the LATE. Let us define $\mathcal{T}_{N(i)}$ as the parameter space of the i-th parameter in $\eta = (\eta_1, \eta_2, \eta_3)$ with $i \in \{1, 2, 3\}$. Let $s_N \geq 1$ be a sequence of integers. Let q, c, C_1 be some finite and positive constants with $q > 4$. Let $a_N = p \vee N$. Let M_N be a sequence of positive constants such that $M_N \geq (\mathrm{E}_P[(Z_i \vee \|X_i\|_\infty)^{2q}])^{1/2q}$. Let $\{\Delta_N\}_{N\geq 1}$ be sequences of positive constants that converges to zero. For any $T \subset [p+1]$, $\delta = (\delta_1, \cdots, \delta_{p+1})' \in \mathbb{R}^{p+1}$ with $\delta_{T,j} = \delta_j$ if $j \in T$ and $\delta_{T,j} = 0$ if $j \notin T$. Define the minimum and maximum sparse eigenvalue by

$$\phi_{\min}(m) = \inf_{\|\delta\|_0 \leq m} \frac{\|(Z_i, X_i')\delta\|_{2,N}}{\|\delta_T\|_1}, \quad \phi_{\max}(m) = \sup_{\|\delta\|_0 \leq m} \frac{\|(Z_i, X_i')\delta\|_{2,N}}{\|\delta_T\|_1}.$$

Assumption 3. *(Regularity conditions for LATE) For $P \in \mathcal{P}_N$, the following conditions hold.*

(i) Equations (2.1)-(2.3) are satisfied with binary D and Z.

(ii) $\|Y\|_{P,q} \leq c_1$.

(iii) For some $\varepsilon > 0$, $\varepsilon \leq P(Z = 1|X) \leq 1 - \varepsilon$ almost surely.

(iv) $\|u\|_{P,2} \geq c_0$.

(v) $\|\mathrm{E}_P[u^2|X]\|_{P,\infty} \leq c_1$.

(vi) Θ *is compact.*

(vii) $\mathrm{E}_P[D|Z=1,X] \geq \mathrm{E}_P[D|Z=0,X]$.

Remark 4. *Chernozhukov et al. (2018) also specializes their results for the LATE framework and provide regularity conditions for LATE estimation. However, their Assumption 5.2 (d) imposes that the denominator is bounded from below by a positive number, which restricts their method from handling weakly identified or unidentified cases. In contrast, our approach does not require a strictly positive denominator for LATE. Our Assumption 3 (vii) allows for the possibility of* $\mathrm{E}_P[D|Z=1,X]-\mathrm{E}_P[D|Z=0,X]$ *approaching zero or even equating zero, thus encompassing weakly identified or unidentified cases.*

Assumption 3 introduces some low-level conditions specifically for the LATE. Assumption 3 (i) emphasizes that both the treatment and instrument are binary. Equations (2.1) and (2.2) play a pivotal role in establishing the Instrument Independence condition, ensuring that given the covariates X, the joint distribution of the outcome Y and the endogenous variable D remains independent of Z. This implies that the instrument Z is "as good as randomly assigned" once we condition on X. Equation (2.3) enforces the Exclusion Restriction condition, ensuring that variations in the instrument Z exclusively affect potential outcomes through its impact on D. Assumption 3 (ii) requires the l_q-norm of the outcome variable is bounded. Assumption 3 (iii) is a standard overlap condition that states for every value of the covariates X, there is at least a small probability that the unit is not treated and at least a small fraction of the population is treated. Assumption 3 (iv) (v) impose constrains on the error term u of the reduce form. Assumption 3 (iv) imposes a lower bound on the l_2-norm of u and (v) restricts the upper bound of the uniform norm of u. Importantly, (vii) allows for the application of the proposed method to weakly or unidentified cases.

Next, we impose the following conditions to guarantee the convergence rate of the nuisance parameter estimators. Recall that the true nuisance parameters vector are denoted as $\eta_0 = (\beta_{11}^0, \beta_{12}^0, \beta_{21}^0, \beta_{22}^0, \gamma^0)$.

Assumption 4. *(Sparse eigenvalue conditions) The sparse eigenvalue conditions hold with probability* $1-o(1)$*, namely, for some* $l_N \to \infty$ *slow enough, we have*

$$1 \lesssim \phi_{\min}(l_N s_N) \leq \phi_{\max}(l_N s_N) \lesssim 1.$$

Assumption 5. *(Sparsity)* $\|\beta_{12}^0\|_0 + \|\beta_{22}^0\|_0 + \|\gamma^0\|_0 \leq s_N$.

Assumption 6. *(Parameters)* $\|\beta_{12}^0\| + \|\beta_{22}^0\| + \|\gamma^0\| \leq C_1$.

Assumption 7. *(Covariates) For $q > 4$, the following condition hold:*

(i) $\inf_{\|\xi\|=1} \mathrm{E}_P[((Z_i, X_i')\xi)^2] \geq c$.

(ii) $\sup_{\|\xi\|=1} \mathrm{E}_P[((Z_i, X_i')\xi)^2] \leq C_1$.

(iii) $N^{-1/2+2/q} M_N^2 s_N \log^2 a_N \leq \Delta_N$.

Assumption 4 is the sparse eigenvalue condition which is similar to Assumption RE in Bickel et al. (2009). In Assumption 5, we impose the number of non-zero components in the high-dimensional nuisance parameter vector by s_N, which in introduced in Assumption 7 (iii). Assumption 6 requires the l_2-norms of the true nuisance parameter vector β_{12}^2, β_{22}^0, and γ^0 are bounded, which is a standard condition. Assumption 7 (i) imposes a lower bound on the second moments of the covariates. Assumption 7 (ii) imposes the second moments of the covariates to be bounded in a uniform manner. Assumption 7 (iii) impose some restriction the rate that the sparsity index s_N, the bound of $2q$-th moments of the covariates, and the dimensionality can grow.

These conditions are sufficient for the high-level conditions invoked in Theorem 1, as formally stated in the following lemma.

Lemma 1. *Suppose Assumptions 3-7 hold. Then Assumptions 1 and 2 hold for the score function $\psi(W; \theta, \eta)$ in equation* (2.4) *in the LATE framework.*

By lemma 1 and Theorem 1, we obtain the following Theorem 2 about the weak convergence of the Gaussian process defined by the score function of the LATE with high-dimensional covariates.

Theorem 2. *Suppose Assumption 3-7 hold. With $\psi(W; \theta, \eta)$ defined as equation (2.4), the process $\widehat{\mathbb{G}}_N(\cdot)$ weakly converges to a centered Gaussian process $\mathbb{G}(\cdot)$ under the null with covariance function $\Omega(\theta_1, \theta_2) = \mathrm{E}_P[(\psi(W; \theta_1, \eta_0) - \mathrm{E}_P[\psi(W; \theta_1, \eta_0)])(\psi(W; \theta_2, \eta_0) - \mathrm{E}_P[\psi(W; \theta_2, \eta)]]$ as N goes to infinity. The covariance function estimator $\widehat{\Omega}(\theta_1, \theta_2)$ defined in* (2.5) *concentrates around the covariance function $\Omega(\theta_1, \theta_2)$ uniformly over Θ,*

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

Remark 5. *We show that our variance estimator is a uniformly consistent estimator of* $\Omega(\theta_1, \theta_2)$ *over* $\theta_1, \theta_2 \in \Theta$ *in Theorem 2. Therefore, our proposed high-dimensional quasi-likelihood test has uniformly correct asymptotic size in the context of LATE, as formally stated in the following theorem.*

Theorem 3. *Suppose that Assumptions 3-7 hold. The test that rejects the null hypothesis* $H_0 : \theta = \theta_0$ *when* $R(q_N(\theta_0), h_N, \Omega)$ *exceeds the* $(1-\alpha)$ *quantile* $c_\alpha(h_N)$ *of its conditional distribution given* $h_N(\cdot)$ *has correct size. Under the null, we have*

$$\lim_{N\to\infty} P(R(\widehat{q}_N(\theta_0), h_N, \widehat{\Omega}) > c_\alpha(h_N)) = \alpha.$$

4 Simulation Studies

4.1 Simulation Setup

We generate data with sample size $N = 500$. We construct the high-dimensional covariates X by

$$X_i \sim N\left(0, \begin{pmatrix} U^0 & U^1 & \dots & U^{\dim(X)-2} & U^{\dim(X)-1} \\ U^1 & U^0 & \dots & U^{\dim(X)-3} & U^{\dim(X)-2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ U^{\dim(X)-2} & U^{\dim(X)-3} & \dots & U^0 & U^1 \\ U^{\dim(X)-1} & U^{\dim(X)-2} & \dots & U^1 & U^0 \end{pmatrix}\right),$$

where $U = 0.25$, and $\dim(X) = 5, 100, 300$, and 500, respectively. Let Q represent the latent compliance class which took values 0 for never-taker, 1 for always-taker, and 2 for compliers. The compliance score is generated by the logistic function $\delta(x) = P(Q = 2|X = x) = \Lambda(\beta_0 + \beta_1 x)$ and the probability of being a never-taker or always-taker is $(1 - \delta(x))/2$, where $\beta_1 = (0.5, 0.5^2, \cdots, 0.5^{\dim(X)})'$ and β_0 is chosen such that $P(Q = 2) = 0.1$ and 0.5, respectively. The instrument variable Z is constructed by $P(Z = 1|X = x) = \Lambda(\gamma_0 + \gamma_1 x)$ with $\gamma_1 = 1$ and γ_0 is set such that $P(Z = 1) = 0.5$. Then the treatment D is constructed by $D = Z \cdot \mathbb{1}\{Q = 2\} + Q \cdot \mathbb{1}\{Q \neq 2\}$, where $\mathbb{1}\{\cdot\}$ represents the indicator function. We generate the outcome variable by $Y_i = D_i + X_i + \varepsilon_i$ where ε_i draws from a standard normal distribution. In this setup, the average treatment effect is equal to 1 for all individuals. Thus we have $\theta_0 = 1$. It is worth noting that $P(Q = 2) = 0.1$ and 0.5 represent the weakly identified and strongly identified cases, respectively. The number of folds in cross fitting is set to be $K = 3$.

4.2 Results

We provide results for four different approaches in our study. These include: (1) the conditional QLR test (AM16) proposed in Andrews and Mikusheva (2016), which is robust against weak identification but not robust to high dimensionality, (2) two conventional machine-learning (ML) methods, which are not robust against weak identifications, and (3) the proposed high-dimensional QLR test (HD-QLR), which is robust against both weak identification and high dimensionality. The first ML method (CCDDHNR18) we employ follows the DML algorithm proposed in Chernozhukov et al. (2018), employing the same Neyman orthogonal score $\psi(W;\theta,\eta)$ and the cross-fitting technique as our proposed inference procedure. However, they employ a different variance estimator and inference method compared to our proposed test, rendering it non-robust to weak identification. The second ML method (BCFH17) we consider is derived from Belloni et al. (2017). It removes the cross-fitting technique used in the algorithm of Chernozhukov et al. (2018) while keeping the other steps the same. For notation simplicity, we refer to these four approaches as AM16, CCDDHNR18, BCFH17, and HD-QLR.

Figures 1 and 2 plot the power curves for nominal 5% tests in the strongly identified and weakly identified simulation designs, respectively. We conduct 2500 iterations of Monte Carlo simulation for both strongly identified and weakly identified settings, comparing the power curves of the four approaches. We consider four scenarios with different numbers of covariates, namely $\dim(X)$= 5, 100, 300, and 500.

Figure 1 represents strongly identified design, with $P(Q=2)=0.5$. The upper left figure represents the power curves for the "low-dimensional" LATE framework with $\dim(X)=5$. We can easily see that all tests maintain satisfactory size control, while our proposed method exhibits better power performance. In the remaining three figures with $\dim(X)$ =100, 300, and 500, our proposed method performs quite competitively with CCDDHNR18 and BCFH17, as all three methods are specifically designed for high-dimensional settings. While CCDDHNR18 and BCFH17 have good power close to the null, their power performance is slightly weaker than that of the proposed method as more distant alternatives. On the other hand, AM16 suffers from significant size distortion and substantial power loss, as the conditional QLR method is not robust against high-dimensional setting.

Figure 2 depicts the power curves in the context of weak identification, with $P(Q=2)=0.1$. In

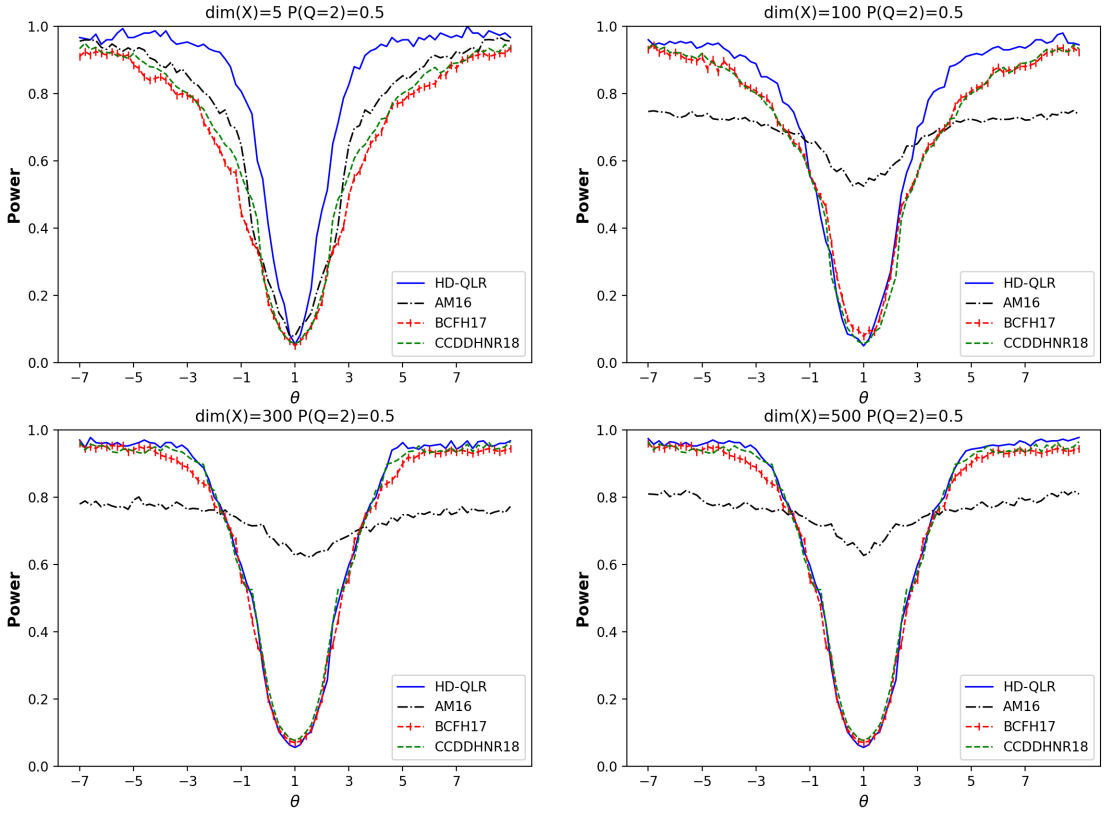

Figure 1: Power curves for HD-QLR (solid blue), AM16 (dash-dot black), BCFH17 (vertical marker red), and CCDDHNR18 (dashed green). Power of nominal 5% tests in strongly identified LATE design with $P(Q = 2) = 0.5$, and $N = 500$ observations. Based on 2,500 replications, and for HD-QLR and AM16, 1,000 draws of conditional critical values were conducted. Upper left panel, number of covariates $\dim(X) = 5$; upper right panel, $\dim(X) = 100$; lower left panel, $\dim(X) = 300$; lower right panel, $\dim(X) = 500$.

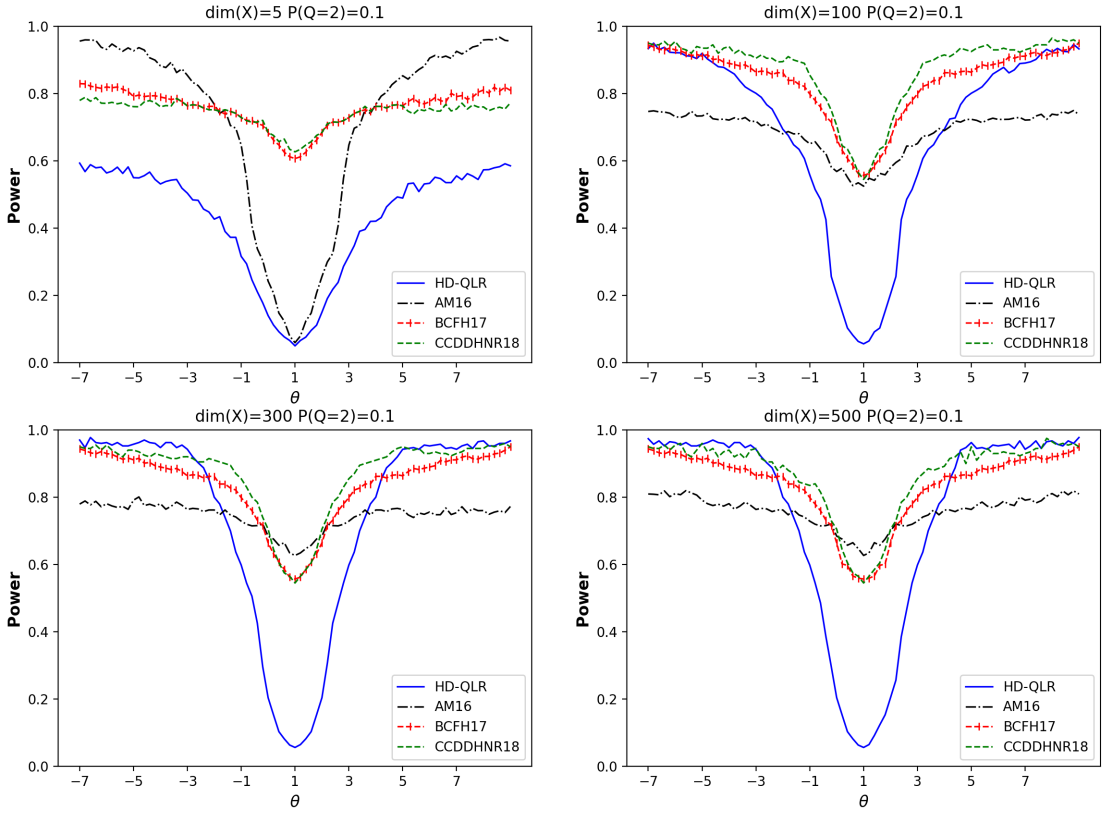

Figure 2: Power curves for HD-QLR (solid blue), AM16 (dash-dot black), BCFH17 (vertical marker red), and CCDDHNR18 (dashed green). Power of nominal 5% tests in weakly identified LATE design with $P(Q = 2) = 0.1$, and $N = 500$ observations. Based on 2,500 replications, and for HD-QLR and AM16, 1,000 draws of conditional critical values were conducted. Upper left panel, number of covariates $\dim(X) = 5$; upper right panel, $\dim(X) = 100$; lower left panel, $\dim(X) = 300$; lower right panel, $\dim(X) = 500$.

the "low-dimensional" LATE model, both our proposed method and AM16 control size well, while AM16 exhibits significantly better power performance compared to the proposed method, since AM16 is specifically designed for weakly identified models with low-dimensional nuisance parameters. In contrast, in the remaining designs, our proposed method outperforms AM16 in terms of both size control and power performance. This is due to the non-robust nature of AM16 in high-dimensional settings. Moreover, both CCDDHNR18 and BCFH17 suffer from severe size distort and significant power loss across all settings, as both methods are not robust against weak identification. Furthermore, upon observing the two upper-left plots in Figure 1 and 2, it becomes evident that our proposed method not only effectively controls size under high-dimensional settings, but also demonstrates excellent size control under low-dimensional scenarios.

Based on our findings, we confirm that the proposed method demonstrates robustness to both weak identification and high dimensionality. On the other hand, CCDDHNR18 and BCFH17 exhibit robustness to high-dimensional settings but are not robust against weak identification. Similarly, AM16 shows robustness to weak identification but lacks robustness to high dimensionality.

5 Empirical Illustrations

5.1 The Impact of Railroad Access on City Growth

To demonstrate the methods outlined in the preceding sections, we re-examine the IV estimation by Hornung (2015) of the impact of railroad access on city growth in 19th-century Prussia. In this study, straight-line corridors between important cities (nodes) are constructed and whether the city is located in this line is used as an instrument for the analysis. We compare our proposed method of high-dimensional QLR test to the other three conventional tests, that is, AM16, CCDDHNR18, and BCFH17, for analyzing the effect of railroad access. Our objective is to enhance our comprehension of the conclusions presented in the literature by conducting a new empirical analysis with the following econometric considerations in mind: 1. we incorporate high-dimensional covariates in order to mitigate unobserved confoundedness; and 2. we account for the weak identification issue in the data and, for the first time, report confidence intervals robust to this source of misleading inference.

Consider the empirical model

$$Y_{it} = D_i\theta_0 + X_i'\beta_1 + \epsilon_{it},$$
$$D_i = \frac{\exp(Z_i\eta_0 + X_i'\beta_2)}{1+\exp(Z_i\eta_0 + X_i'\beta_2)} + v_i,$$

for estimation of θ_0, where Y_{it} denotes the urban population growth rate in city i at time period t, D_i is a dummy variable if there is a railroad access by 1848 in city i, Z_i denotes whether the city i was located within a straight-line corridor between junction stations (nodes) in 1848, the explanatory variables X_i includes a lagged dependent variable, as well as the distance to the closest node of railroad lines, age composition, the primary education of the urban population, county-level concentration of large landholdings, access to the main roads, rivers, and ports, pre-railroad city growth 1831-1837, and the size of the civilian and military population in 1849.

In the context of the study, it is important to note that the adoption of railroad technology by cities located on a straight line between two important cities is based on random assignment. This random assignment occurs because the placement of these cities along the straight line is not intentionally controlled by any entity. In 19th-century Prussia, the decision to construct railroads was not made by the government due to financial limitations, but rather through negotiations between each city council and private railroad enterprises. Hence, each city had the autonomy to determine whether or not to proceed with railroad construction. In this study, the "compliers" refer to (1) cities that are on the straight line between two important cities AND eventually got a railroad station, and (2) cities that are NOT on the straight line between two important cities AND didn't get a train station.

We apply the proposed method to the city-level railroad data of Hornung (2015). According to Table 5 of Hornung (2015), the first-stage F-statistics range from 26.46 to 38.29, indicating instrument weakness based on the tF critical value function proposed by Lee et al. (2022). We conducted a re-analysis by incorporating the polynomial and interaction terms of the original covariates and present the results in Table 1. The sample sizes range from $N = 898$ cities with the dimensionality of covariates $\dim(X) = 204$ to $N = 926$ cities with $\dim(X) = 212$.

Table 1 summarizes the results. In order to highlight the efficiency of the proposed method in the high-dimensional context, we report LATE estimates along with their corresponding confidence intervals and the lengths of the confidence intervals. In Panel A, we report the results obtained

Y_{it}: population	Main periods		Subperiods						
growth rate	1831-37	49-71	49-52	52-55	55-58	58-61	61-64	64-67	67-71
Panel A: AM16									
LATE	0.007	0.039	0.010	0.020	0.063	0.030	0.037	0.056	0.044
CI	[-0.002,	[0.022,	[-0.017,	[0.004,	[0.030,	[0.011,	[0.019,	[0.012,	[0.018,
	0.020]	0.05]	0.05]	0.039]	0.063]	0.050]	0.050]	0.420]	0.155]
length of CI	0.022	0.028	0.067	0.035	0.033	0.039	0.031	0.408	0.137
Panel B: CCDDHNR18									
LATE	0.001	0.014	0.012	0.011	0.007	0.000	0.020	0.012	0.011
CI	[-0.026,	[-0.004,	[-0.019,	[-0.014,	[-0.009,	[-0.016,	[-0.019,	[-0.014,	[-0.016,
	0.021]	0.033]	0.039]	0.035]	0.044]	0.030]	0.052]	0.039]	0.036]
length of CI	0.047	0.037	0.058	0.048	0.053	0.046	0.070	0.052	0.052
Panel C: BCFH17									
LATE	0.000	0.012	0.009	0.009	0.012	0.006	0.015	0.012	0.013
CI	[-0.017,	[-0.003,	[-0.009,	[-0.006,	[-0.007,	[-0.008,	[-0.009,	[-0.018,	[-0.008,
	0.017]	0.026]	0.026]	0.023]	0.031]	0.020]	0.040]	0.041]	0.034]
length of CI	0.034	0.029	0.035	0.029	0.038	0.028	0.049	0.059	0.042
Panel D: HD-QLR with the number of folds K=4									
LATE	0.001	0.013	0.010	0.011	0.014	0.004	0.018	0.014	0.011
CI	[-0.012,	[0.005,	[0.000,	[0.002,	[0.003,	[-0.001,	[0.003,	[-0.004,	[-0.002,
	0.012]	0.020]	0.021]	0.018]	0.027]	0.016]	0.029]	0.032]	0.023]
length of CI	0.024	0.015	0.021	0.016	0.024	0.017	0.026	0.033	0.024
Size N	898	906	929	924	914	926	924	919	919
dim(X)	204	212	212	212	212	212	212	212	212

Table 1: Displayed are estimates, confidence intervals (CI) and the length of confidence intervals for the coefficient of the railroad access. Panel A displays the results of AM16. Panel B presents the results obtained from CCDDHNR18. Panel C shows the results derived from BCFH17. Panel D showcases the results of the proposed HD-QLR test with the number $K = 4$ folds for cross fitting. Estimation and inference results in panel B and D are based on 10 iterations of resampled cross fitting.

Y_{it}: population	Main periods		Subperiods						
growth rate	1831-37	49-71	49-52	52-55	55-58	58-61	61-64	64-67	67-71
Panel A: HD-QLR with the number of folds K=3									
LATE	0.008	0.012	0.010	0.011	0.014	0.004	0.018	0.014	0.011
CI	[-0.013,	[0.005,	[0.001,	[0.002,	[0.003,	[-0.002,	[0.003,	[-0.005,	[-0.003,
	0.011]	0.021]	0.023]	0.030]	0.026]	0.016]	0.030]	0.032]	0.022]
length of CI	0.024	0.016	0.023	0.018	0.023	0.018	0.027	0.037	0.025
Panel B: HD-QLR with the number of folds K=4									
LATE	0.001	0.013	0.010	0.011	0.014	0.004	0.018	0.014	0.011
CI	[-0.012,	[0.005,	[0.000,	[0.002,	[0.003,	[-0.001,	[0.003,	[-0.004,	[-0.002,
	0.012]	0.020]	0.021]	0.018]	0.027]	0.016]	0.029]	0.032]	0.023]
length of CI	0.024	0.015	0.021	0.016	0.024	0.017	0.026	0.033	0.024
Size N	898	906	929	924	914	926	924	919	919
dim(X)	204	212	212	212	212	212	212	212	212

Table 2: Displayed are estimates, CI and the length of CI for the coefficient of the railroad access. Panel A showcases the results of the proposed HD-QLR test with the number $K = 3$ folds for cross fitting. Panel B demonstrates the results of the proposed test with the number $K = 4$ folds for cross fitting. Estimation and inference are based on 10 iterations of resampled cross fitting.

from AM16, which is not robust in high-dimensional setting. Additionally, we present results by CCDDHNR18 in Panel B and results by BCFH17 in panel C. Furthermore, Panel D displays results from the proposed method with $K = 4$ folds of cross-fitting. Different columns report the results for several dependent variables across various time periods. Specifically, Columns (II) and (III) report results in two main periods: 1831-37 , and 1849-1871. Columns (IV)-(X) represent results in seven subperiods. To mitigate the uncertainty induced by sample splitting, we compute estimates and confidence intervals based on the average of ten randomized DML following Chernozhukov et al. (2018).

Upon comparing the results from Panel A and Panel D, we observe that the point estimates from the conventional test AM16, which does not account for the high-dimensional controls, are consistently larger than those obtained by the proposed method. Additionally, the confidence intervals derived from AM16 are consistently wider than those obtained from the proposed test. Another notable finding is that several effects that were deemed significant in AM16, without considering high-dimensional covariates, become insignificant after accounting for these covariates in the proposed method. Specifically, the incorporation of high-dimensional covariates results in the loss of statistical significance for three out of the seven coefficients.

Comparing the results from Panel B, Panel C, and Panel D, we observe a similarity in the LATE estimates across these panels. However, it is important to note that the lengths of confidence intervals derived from CCDDHNR18 and BCFH17, which lack robustness against weak identification, are significantly larger than that obtained from our proposed method.

To examine the impact of the number of folds on the results, we present the outcomes obtained from our proposed method with varying number of folds in Table 2. Upon observation, we find that the point estimates and confidence intervals exhibit similarity across Panel A and B, despite the variation in the number of folds used.

To robustly account for the weak identification issue in the high-dimensional context, we recommend researcher employ our proposed high-dimensional conditional QLR test.

5.2 The Boundary Effects on Rental Prices

In this subsection, we reexamine the IV estimation performed by Ambrus et al. (2020) concerning the impact on housing prices of a cholera epidemic. The authors present estimates of the epidemic's effect on property values a decade later, following the unexpected cholera outbreak of 1954, using a fuzzy regression discontinuity (RD) design. In their design, Y_i is defined as the log rental price of house i in 1864. The variable D_i is an indicator equal to 1 if house i experiences at least one cholera death. The variable Z_i is an indicator equal to 1 if property i falls inside the Broad Street pump (BSP) catchment areas, which are the contaminated areas affected by the cholera outbreak. The control variable X_i comprises all house characteristic variables listed in Table 1 of Ambrus et al. (2020), such as distance to the closest pump, distance to the fire station, distance to the urinal, sewer access, and a total of 23 variables.

As discussed in Hahn et al. (2001) regarding RD design, when D_i is binary and certain alternative conditions are met, allowing for dependence between D_i and θ, the RD effect captures the LATE for compliers at the cutoff point. Compliers refer to observations where the variable Z_i switches from zero to one. In this context, the estimated effect of cholera-related deaths on housing prices can be considered as a LATE within the fuzzy RD framework. The "compliers" refer to (1) houses located within the contaminated areas affected by the cholera AND experience at least one cholera death, and (2) houses located outside the contaminated areas affected by the cholera AND do not experience any cholera death.

Table B2 in Ambrus et al. (2020) presents fuzzy RD and IV estimates of the effect of cholera-related death on rental prices in 1864. Notably, the first-stage F-statistics in the IV estimation are around 10, indicating weak identification according to the tF critical value function proposed by Lee et al. (2022). In light of this observation, we preceed with a reanalysis by incorporating all the house characteristic variables listed in Table 1 in Ambrus et al. (2020). The sample size for this reanalysis is $N = 467$ and the dimensionality of the covariates $\dim(X) = 23$.

Table 3 presents the results of LATE estimates, CIs, and the lengths of CI by comparing the four approaches: AM16, CCDDHNR18, BCFH17, and the proposed HD-QLR. When comparing AM16, which does not consider high-dimensional covariates, with the other three high-dimensional approaches, we

	AM16	CCDDHNR18	BCFH17	HD-QLR
LATE	-1.205	-0.413	-0.357	-0.421
CI	[-2.230, -0.650]	[-3.565, 2.670]	[-1.291, 0.576]	[-1.080, 0.035
length of CI	1.580	6.235	1.866	1.115

Table 3: Displayed are LATE estimates, CIs and the length of CI for the coefficient of the cholera-related deaths using four different approaches: AM16, CCDDHNR18, BCFH17, and the proposed HD-QLR. Estimation and inference results in CCDDHNR18 and HD-QLR are based on 10 iterations of resampled cross fitting with $K = 4$ folds for cross fitting.

observe that AM16 yields the smallest point estimate. This suggests that AM16 tends to overestimate the effect under high-dimensional scenarios. However, the significant results obtained in AM16 become insignificant after accounting for these high-dimensional covariates in the other three methods. Furthermore, among CCDDHNR18, BCFH17, and the proposed HD-QLR test, although the point estimates are similar, our proposed method achieves the smallest length of CI. On the other hand, the other two conventional ML methods, which do not account for weak identification, yield relatively large lengths of CIs.

We strongly recommend researchers to utilize the proposed HD-QLR test when dealing with high-dimensional models that may potentially encounter weak identification issues.

6 Conclusion

In this paper, we study the issue of weak identification in the LATE framework with high-dimensional covariates. Our primary contribution is the development of an identification-robust test, accompanied by an easily implementable algorithm for inference and confidence interval construction for the LATE estimate. We demonstrate that our proposed method maintains uniformly correct asymptotic size. Simulation studies demonstrate that, in finite sample, our proposed method outperforms the convention identification-robust test and the conventional ML tests in terms of size control and power performance under the high-dimensional LATE with weak identification. Furthermore, we apply the

proposed method to revisit the study conducted by Hornung (2015) concerning the effect of railroad access on urban population growth. Our method yields estimates that are generally smaller than those obtained using conventional identification-robust test, while the confidence intervals are substantially shorter compared to ML-based approaches. Similarly, when revisiting the study by Ambrus et al. (2020) on boundary effects in rental price, we obtain exactly the same results as previously reported. Overall, our approach provides robustness against both weak identification and high-dimensional setting, showcasing its potential applicability in various empirical studies.

Appendix

A Useful Lemmas

Here, for the convenience of readers, we provide the convergence rate for Lasso with logistoc model and OLS model in Lemma 2 and 3. Lemma 2 relies on Lemma 1 in Belloni et al. (2016). Lemma 3 relies on Theorem 1 in Belloni and Chernozhukov (2013).

Lemma 2. *(Convergence rate for Lasso with logistic model) Suppose that Assumption 4-7 hold. In addition, suppose that the penalty choice* $\lambda_1 = K_1\sqrt{N\log(pN)}$ *and* $\lambda_2 = K_2\sqrt{N\log(pN)}$ *for* $K_1, K_2 > 0$*. Then with probability* $1 - o(1)$*,*

$$\|(\widehat{\beta}_{11}, \widehat{\beta}_{12}) - (\beta_{11}^0, \beta_{12}^0)\| \vee \|\widehat{\gamma} - \gamma^0\| \lesssim \sqrt{\frac{s_N \log(pN)}{N}}.$$

Lemma 3. *(Convergence rate for Lasso with OLS) Suppose that Assumption 4-7 hold. Moreover, suppose that the penalty choice* $\lambda_3 = K_3\sqrt{N\log(pN)}$ *for* $K_3 > 0$*. Then with probability* $1 - o(1)$*,*

$$\|(\widehat{\beta}_{21}, \widehat{\beta}_{22}) - (\beta_{21}^0, \beta_{22}^0)\| \lesssim \sqrt{\frac{s_N \log(pN)}{N}}.$$

B Proofs of the Main Results

B.1 Proof of Theorem 1

Proof. Without loss of generality, we define the size of each fold I_k as $n = N/K$. For notation simplicity, we introduce the notation $[r] = \{1, \cdots, r\}$ for any $r \in \mathbb{N}$. We divide the proof into three steps. In Step 1, we prove the equation (3.10) and the asymptotic normality of $\widehat{\mathbb{G}}_N(\theta)$ over Θ, that is, the asymptotic normality of $(\widehat{\mathbb{G}}_N(\theta_1), \cdots, \widehat{\mathbb{G}}_N(\theta_L))$ for any $(\theta_1, \cdots, \theta_L) \in \Theta \times \cdots \times \Theta$. In step 2, we establish the asymptotic equicontinuity of $\widehat{\mathbb{G}}_N$ over Θ. Roughly speaking, this means that whenever $\theta_1 \in \Theta$ and $\theta_2 \in \Theta$ are close to each other, $\widehat{\mathbb{G}}_N(\theta_1) - \widehat{\mathbb{G}}_N(\theta_2)$ is close to zero. Since Θ is a compact set, the proof of the weak convergence result is done. In Step 3, we prove that $\widehat{\Omega}(\theta_1, \theta_2)$ is a uniformly consistent estimator of the covariance function $\Omega(\theta_1, \theta_2)$ over Θ.

Step 1. In this step, we first establish equation (3.10). Becasue K is a fixed integer and independent

of N, it suffices to show that for any $k \in [K]$,

$$\mathbb{E}_{n,k}[\psi(W;\theta,\widehat{\eta}_k)] - \mathrm{E}_P[\psi(W;\theta,\widehat{\eta}_k)] - (\mathbb{E}_{n,k}[\psi(W_i,\theta,\eta_0)] - \mathrm{E}_P[\psi(W;\theta,\eta_0)]) = O_p(N^{-1/2} r'_N). \quad \text{(B.1)}$$

For notation simplicity, we define $\mathbb{E}_{n,k}[f(W)] = n^{-1}\sum_{i\in I_k} f(W_i)$. In order to show this, let us fix $k \in [K]$ and introduce an empirical process notation,

$$\mathbb{G}_{n,k}[\phi(W)] = \frac{1}{\sqrt{n}} \sum_{i\in I_k} (\phi(W_i) - \mathrm{E}_P[\phi(W)]),$$

where ϕ is any P_N-integrable function of W. Then by triangle inequality, we have

$$\|\mathbb{E}_{n,k}[\psi(W;\theta,\widehat{\eta}_k)] - \mathrm{E}_P[\psi(W;\theta,\widehat{\eta}_k)] - (\mathbb{E}_{n,k}[\psi(W_i,\theta,\eta_0)] - \mathrm{E}_P[\psi(W;\theta,\eta_0)])\| \quad \text{(B.2)}$$

$$= n^{-1/2}\|\mathbb{G}_{n,k}[\psi(W;\theta,\widehat{\eta}_k)] - \mathbb{G}_{n,k}[\psi(W;\theta,\eta_0)]\| := n^{-1/2}\mathcal{I}_{k3}. \quad \text{(B.3)}$$

Notice that, conditional on $(W_i)_{i\in I_k^c}$, the estimator $\widehat{\eta}_k$ is non-stochastic. Then we have,

$$\begin{aligned}
\mathrm{E}_P[\mathcal{I}_{k3}^2|(W_i)_{i\in I_k^c}] &= \mathrm{E}_P\left[(\psi(W;\theta,\widehat{\eta}_k) - \psi(W;\theta_0,\eta_0))^2|(W_i)_{i\in I_k^c}\right] \\
&\leq \sup_{\eta\in\mathcal{T}_N} \mathrm{E}_P\left[(\psi(W;\theta,\eta) - \psi(W;\theta_0,\eta_0))^2|(W_i)_{i\in I_k^c}\right] \\
&\leq \sup_{\eta\in\mathcal{T}_N} \mathrm{E}_P[(\psi(W;\theta,\eta) - \psi(W;\theta_0,\eta_0))^2] \leq (r'_N)^2.
\end{aligned}$$

This completes the proof of equation (3.10). Combining (3.10) with the Lindeberg-Feller central limit theorem (CLT) and the Cramer-Wold device yields the asymptotic normality of $\widehat{\mathbb{G}}_N(\theta)$ for any $\theta \in \Theta$.

Step 2. In this step, we prove the asymptotic equicontinuity of $\widehat{\mathbb{G}}_N$ on Θ. The asymptotic equicontinuity of $\widehat{\mathbb{G}}_N$ can be stated as, for any $\epsilon_1 > 0$, and any $\theta_1, \theta_2 \in \Theta$ such that $|\theta_1 - \theta_2| \leq \delta$,

$$\lim_{\delta\to 0} \limsup_{N\to\infty} P\left(|\widehat{\mathbb{G}}_N(\theta_1) - \widehat{\mathbb{G}}_N(\theta_2)| > \epsilon_1\right) = 0. \quad \text{(B.4)}$$

By Markov's inequity, for any $\epsilon_1 > 0$,

$$P\left(\left|\widehat{\mathbb{G}}_N(\theta_1) - \widehat{\mathbb{G}}_N(\theta_2)\right| > \epsilon_1\right) \leq \frac{1}{\epsilon_1}\mathrm{E}_P\left[\left|\widehat{\mathbb{G}}_N(\theta_1) - \widehat{\mathbb{G}}_N(\theta_2)\right|\right].$$

Thus, it suffices to show that for each $k \in [K]$.

$$\lim_{\delta\to 0} \limsup_{N\to\infty} \sqrt{N}\mathrm{E}_P\left[|\mathbb{E}_{n,k}[(\theta_1 - \theta_2)\psi^a(W;\widehat{\eta}_k)] - \mathrm{E}_P[(\theta_1 - \theta_2)\psi^a(W;\widehat{\eta}_k)]|\right] = 0. \quad \text{(B.5)}$$

Note that

$$\mathrm{E}_P\left[\left|\mathbb{E}_{n,k}[(\theta_1-\theta_2)\psi^a(W;\widehat{\eta}_k)]-\mathrm{E}_P[(\theta_1-\theta_2)\psi^a(W;\widehat{\eta}_k)]\right|^2\right]\leq n^{-1}\delta^2\mathrm{E}_P\left[\psi^a(W;\widehat{\eta}_k)^2\right]\leq n^{-1}\delta^2c_1^2,$$

which implies the equation (B.5). Thus, we complete the proof of the asymptotic equicontinuity of $\widehat{\mathbb{G}}_N$ over Θ.

Step 3. In this step, we first show $\widehat{\Omega}(\theta_1,\theta_2)=\Omega(\theta_1,\theta_2)+O_P(\rho_N)$, and then we show $\widehat{\Omega}$ is a uniformly consistent estimator for Ω over Θ. To prove the first part, it suffices to show that for any pair $(\theta_1,\theta_2)\in\Theta$ and each $k\in[K]$,

$$\mathcal{I}_k=|\mathbb{E}_{n,k}[\psi(W;\theta_1,\widehat{\eta}_k)\psi(W;\theta_2,\widehat{\eta}_k)]-\mathrm{E}_P[\psi(W;\theta_1,\eta_0)\psi(W;\theta_2,\eta_0)]|=O_p(\rho_N),\quad \textit{and}$$
$$\mathcal{I}'_k=|\mathbb{E}_{n,k}[\psi(W;\theta,\widehat{\eta}_k)]-\mathrm{E}_P[\psi(W;\theta,\eta_0)]|=O_p(\rho_N).$$

Note that by triangle inequality, we have $\mathcal{I}_k\leq\mathcal{I}_{k1}+\mathcal{I}_{k2}$, and $\mathcal{I}'_k\leq\mathcal{I}_{k4}+\mathcal{I}_{k5}$, where

$$\mathcal{I}_{k1}=|\mathbb{E}_{n,k}[\psi(W;\theta_1,\widehat{\eta}_k)\psi(W;\theta_2,\widehat{\eta}_k)]-\mathbb{E}_{n,k}[\psi(W;\theta_1,\eta_0)\psi(W;\theta_2,\eta_0)]|,$$
$$\mathcal{I}_{k2}=|\mathbb{E}_{n,k}[\psi(W;\theta_1,\eta_0)\psi(W;\theta_2,\eta_0)]-\mathrm{E}_P[\psi(W;\theta_1,\eta_0)\psi(W;\theta_2,\eta_0)]|,$$
$$\mathcal{I}_{k4}=|\mathbb{E}_{n,k}[\psi(W;\theta,\widehat{\eta}_k)]-\mathbb{E}_{n,k}[\psi(W;\theta,\eta_0)]|,\quad \mathcal{I}_{k5}=|\mathbb{E}_{n,k}[\psi(W;\theta,\eta_0)]-\mathrm{E}_P[\psi(W;\theta,\eta_0)]|.$$

First, we bound $\mathcal{I}_{k2}$ and $\mathcal{I}_{k5}$. Note that for $q\geq4$, we have

$$\mathrm{E}_P[\mathcal{I}_{k2}^2]\leq\sup_{\theta\in\Theta}n^{-1}\mathrm{E}_P[\psi(W;\theta,\eta_0)^4]\leq n^{-1}c_1^4,$$
$$\mathrm{E}_P[\mathcal{I}_{k5}^2]\leq\sup_{\theta\in\Theta}n^{-1}\mathrm{E}_P[\psi(W;\theta,\eta_0)^2]\leq n^{-1}c_1^2,$$

where the last inequality follows from Assumption 2 (ii) and Jensen's inequality. Next, we try to bound $\mathcal{I}_{k1}$.

$$\begin{aligned}
\mathcal{I}_{k1}&=\left|\frac{1}{n}\sum_{i\in I_k}[\psi(W_i;\theta_1,\widehat{\eta}_k)\psi(W_i;\theta_2,\widehat{\eta}_k)-\psi(W_i;\theta_1,\eta_0)\psi(W_i;\theta_2,\eta_0)]\right|\\
&\leq\frac{1}{n}\sum_{i\in I_k}\left|\psi(W_i;\theta_1,\widehat{\eta}_k)\psi(W_i;\theta_2,\widehat{\eta}_k)-\psi(W_i;\theta_1,\eta_0)\psi(W_i;\theta_2,\eta_0)\right|\\
&\leq\frac{2}{n}\sum_{i\in I_k}\sup_{\theta\in\Theta}\sup_{\eta\in\mathcal{T}_N}\Big(|\psi(W_i;\theta,\widehat{\eta}_k)-\psi(W_i;\theta,\eta_0)|\times|\psi(W_i;\theta,\eta)|\Big)\\
&\leq\frac{2}{n}\sum_{i\in I_k}\Big(\sup_{\theta\in\Theta}(\psi(W_i;\theta,\widehat{\eta}_k)-\psi(W_i;\theta,\eta_0))^2\Big)^{1/2}\times\Big(\sup_{\theta\in\Theta}\sup_{\eta\in\mathcal{T}_N}\frac{2}{n}\sum_{i\in I_k}\psi(W_i;\theta,\eta)^2\Big)^{1/2}
\end{aligned}$$

and the conditional expectation of the first term given $(W_i)_{i\in I_k^c}$ on the event that $\widehat{\eta}_k \in \mathcal{T}_N$ is equal to

$$\sup_{\theta\in\Theta} \mathrm{E}_P\left[\|\psi(W;\theta,\widehat{\eta}_k)-\psi(W;\theta,\eta_0)\|^2|(W_i)_{i\in I_k^c}\right] \le \sup_{\eta\in\mathcal{T}_N,\theta\in\Theta} \mathrm{E}_P\left[\|\psi(W;\theta,\eta)-\psi(W;\theta,\eta_0)\|^2|(W_i)_{i\in I_k^c}\right] = r_N'^2,$$

Because the event that $\widehat{\eta}_k \in \mathcal{T}_N$ holds with probability $1-\Delta_N = 1-o(1)$, it follows that $\mathcal{I}_{k1} = O_P(r'_N) = O_P(\delta_N)$. Since $\mathcal{I}_{k2} = O_P(N^{-1/2})$ and $\delta_N \ge N^{-1/2}$, we have $\mathcal{I}_k = O_p(\rho_N)$ with $\rho_N \lesssim \delta_N$. Then we try to bound $\mathcal{I}_{k4}$.

$$\begin{aligned}\mathcal{I}_{k4} &= \left|\frac{1}{n}\sum_{i\in I_k}[\psi(W_i;\theta,\widehat{\eta}_k)-\psi(W_i;\theta,\eta_0)]\right| \le \frac{1}{n}\sum_{i\in I_k}|\psi(W_i;\theta,\widehat{\eta}_k)-\psi(W_i;\theta,\eta_0)| \\ &\le \sup_{\theta\in\Theta}\left(\frac{1}{n}\sum_{i\in I_k}\|\psi(W_i;\theta,\widehat{\eta}_k)-\psi(W_i;\theta,\eta_0)\|^2\right)^{1/2}.\end{aligned}$$

By using the similar argument that we use to bound $\mathcal{I}_{k1}$, we obtain $\mathcal{I}_{k4} = O_P(r'_N)$. Therefore, we have $\mathcal{I}'_k = O_P(\rho_N)$ with $\rho_N \lesssim \delta_N$. This completes the proof of $\widehat{\Omega}(\theta_1,\theta_2) = \Omega(\theta_1,\theta_2) + O_P(\rho_N)$. To prove $\widehat{\Omega}$ is a uniformly consistent estimator of Ω over Θ, we need to show that for any $\varepsilon_2 > 0$, and any $\theta_1,\theta_2,\theta_1',\theta_2' \in \Theta$ such that $|\theta_1-\theta_1'| \le \delta_1$ and $|\theta_2-\theta_2'| \le \delta_2$, we have

$$\lim_{\delta_1,\delta_2\to 0}\limsup_{N\to\infty} P\left(|\widehat{\Omega}(\theta_1,\theta_2)-\widehat{\Omega}(\theta_1',\theta_2')| > \epsilon_2\right) = 0.$$

By Markov's inequality, for any $\varepsilon_2 > 0$,

$$P\left(|\widehat{\Omega}(\theta_1,\theta_2)-\widehat{\Omega}(\theta_1',\theta_2')| > \epsilon_2\right) \le \frac{1}{\varepsilon_2}\mathrm{E}_P\left[\left|\widehat{\Omega}(\theta_1,\theta_2)-\widehat{\Omega}(\theta_1',\theta_2')\right|\right].$$

Thus, it suffices to show that for each $k \in [K]$,

$$\mathcal{I}_{k6} = \mathrm{E}_P\left[\left|\mathbb{E}_{n,k}[\psi(W;\theta_1,\widehat{\eta}_k)\psi(W;\theta_2,\widehat{\eta}_k)] - \mathbb{E}_{n,k}[\psi(W;\theta_1',\widehat{\eta}_k)\psi(W;\theta_2',\widehat{\eta}_k)]\right|\right] = 0,$$

as $n\to\infty, \delta_1,\delta_2 \to 0$. Note that

$$\begin{aligned}\mathcal{I}_{k6} &\le \mathrm{E}_P\left[\mathbb{E}_n\left[|\psi(W;\theta_1,\widehat{\eta}_k)-\psi(W;\theta_1',\widehat{\eta}_k)|\cdot|\psi(W;\theta_2,\widehat{\eta}_k)|\right] + \mathbb{E}_n\left[|\psi(W;\theta_2,\widehat{\eta}_k)-\psi(W;\theta_2',\widehat{\eta}_k)|\cdot|\psi(W;\theta_1',\widehat{\eta}_k)|\right]\right] \\ &= \mathrm{E}_P\left[\mathbb{E}_n\left[|\psi^a(W;\widehat{\eta}_k)\cdot(\theta_1-\theta_1')|\cdot|\psi(W;\theta_2,\widehat{\eta}_k)|\right]\right] + \mathrm{E}_P\left[\mathbb{E}_n\left[|\psi^a(W;\widehat{\eta}_k)\cdot(\theta_2-\theta_2')|\cdot|\psi(W;\theta_1',\widehat{\eta}_k)|\right]\right] \\ &\le (\mathrm{E}_P\left[\psi^a(W;\widehat{\eta}_k)^2\cdot(\theta_1-\theta_1')^2\right])^{1/2}\cdot\left(\mathrm{E}_P\left[\psi(W;\theta_2,\widehat{\eta}_k)^2\right]\right)^{1/2} + (\mathrm{E}_P\left[\psi^a(W;\widehat{\eta}_k)^2\cdot(\theta_2-\theta_2')^2\right])^{1/2}\cdot\left(\mathrm{E}_P\left[\psi(W;\theta_1',\widehat{\eta}_k)^2\right]\right)^{1/2} \\ &\le \delta_1\cdot(\mathrm{E}_P\left[\psi^a(W;\widehat{\eta}_k)^4\right])^{1/4}\cdot\left(\mathrm{E}_P\left[\psi(W;\theta_2,\widehat{\eta}_k)^4\right]\right)^{1/4} + \delta_2\cdot(\mathrm{E}_P\left[\psi^a(W;\widehat{\eta}_k)^4\right])^{1/4}\cdot\left(\mathrm{E}_P\left[\psi(W;\theta_1',\widehat{\eta}_k)^4\right]\right)^{1/4} \\ &\le (\delta_1+\delta_2)c_1^2,\end{aligned}$$

where the second inequality follows from Cauchy-Schwarz inequality, the third inequality follows from Jensen's inequality, and the last one is from Assumption 2 (ii). It is obvious that $\lim_{\delta_1,\delta_2\to 0}\mathcal{I}_{k6}=0$. Therefore, $\widehat{\Omega}$ is a uniformly consistent estimator of Ω over Θ. This completes the whole proof of Theorem 1.

□

B.2 Proof of Theorem 2

Proof. As long as we show Lemma 1 holds, the proof of Theorem 2 is done. Let us define $\mathcal{T}_N$ as the set of all $\eta=(g,m,p)$ consisting of P-square-integrable function g,m and p such that

$$\|\eta-\eta_0\|_{P,q}\leq c_1,\quad \|\eta-\eta_0\|_{P,2}\leq\delta_N.$$

We proceed in four steps.

Step 1. We first verify the Assumption 1 that the score of the LATE in (2.4) satisfies the moment condition (3.1) and the Newman orthogonality condition (3.2). It can be easily verified that moment condition is satisfied by the construction of the score. The Gateaux derivative in the direction $\eta-\eta_0=(g-g_0,m-m_0,p-p_0)$ is given by

$$\begin{aligned}
&\partial_\eta \mathrm{E}_P\left[\psi(W;\theta_0,\eta)\right]\Big|_{\eta=\eta_0}(\eta-\eta_0)\\
&=\mathrm{E}_P\left[\left(1-\frac{Z}{p_0(X)}\right)(g(1,X)-g_0(1,X))\right]-\mathrm{E}_P\left[\left(1-\frac{1-Z}{1-p_0(X)}\right)(g(0,X)-g_0(0,X))\right]\\
&-\theta_0\mathrm{E}_P\left[\left(1-\frac{Z}{p_0(X)}\right)(m(1,X)-m_0(1,X))\right]+\theta_0\mathrm{E}_P\left[\left(1-\frac{1-Z}{1-p_0(X)}\right)(m(0,X)-m_0(0,X))\right]\\
&+\mathrm{E}_P\left[\left(\frac{\theta_0 Z(D-m(1,X))-Z(Y-g_0(1,X))}{p_0(X)^2}+\frac{\theta_0(1-Z)(D-m(0,X))-(1-Z)(Y-g_0(0,X))}{(1-p_0(X))^2}\right)\times(p(X)-p_0(X))\right]\\
&=0,
\end{aligned}$$

where the last equality follows from the law of iterated expectations and

$$\begin{aligned}
&\mathrm{E}_P[Z|X]=p_0(X),\quad \mathrm{E}_P[Z(Y-g_0(1,X))|X,Z]=0,\quad \mathrm{E}_P[Z(D-m_0(1,X))|X,Z]=0,\qquad \text{(B.6)}\\
&\mathrm{E}_P[1-Z|X]=1-p_0(X),\quad \mathrm{E}_P[(1-Z)(Y-g_0(0,X))|X,Z]=0,\quad \mathrm{E}_P[(1-Z)(D-m_0(0,X))|X,Z]=0.
\end{aligned}$$

Referring to the definitions of the score for the LATE in (2.4) and linear orthogonal score in (3.9), we

have

$$\psi^b(W;\eta) = g(1,X) - g(0,X) + \frac{Z(Y-g(1,X))}{p(X)} - \frac{(1-Z)(Y-g(0,X))}{1-p(X)},$$
$$\psi^a(W;\eta) = -m(1,X) + m(0,X) - \frac{Z(D-m(1,X))}{p(X)} + \frac{(1-Z)(D-m(0,X))}{1-p(X)}.$$

Then we have $\psi(W;\theta,\eta) = \psi^b(W;\eta) + \theta \times \psi^a(W;\eta)$. Therefore, all the conditions in Assumption 1 hold.

Step 2. Next, let us verify Assumption 2 (iii). Note that

$$\begin{aligned}
&\mathrm{E}_P\left[\psi(W;\theta,\eta_0)^2\right] = \mathrm{E}_P\left[\,(g_0(1,X) - g_0(0,X) - \theta(m_0(1,X) - m_0(0,X)))^2\,\right] \\
&+ \mathrm{E}_P\Big[\Big(\frac{Z(Y-g_0(1,X))}{p_0(X)} - \frac{(1-Z)(Y-g_0(0,X))}{1-p_0(X)} - \theta\big(\frac{Z(D-m_0(1,X))}{p_0(X)} - \frac{(1-Z)(D-m_0(0,X))}{1-p_0(X)}\big)\Big)^2\Big] \\
&\geq \mathrm{E}_P\Big[\Big(\frac{Z(Y-g_0(1,X))}{p_0(X)} - \frac{(1-Z)(Y-g_0(0,X))}{1-p_0(X)}\Big)^2\Big] - \theta^2\mathrm{E}_P\Big[\Big(\frac{Z(D-m_0(1,X))}{p_0(X)} - \frac{(1-Z)(D-m_0(0,X))}{1-p_0(X)}\Big)^2\Big] \\
&\geq \mathrm{E}_P\left[\frac{Z^2(Y-g_0(1,X))^2}{p_0(X)^2}\right] + \mathrm{E}_P\left[\frac{(1-Z)^2(Y-g_0(0,X))^2}{(1-p_0(X))^2}\right] \\
&\geq \frac{\mathrm{E}_P\left[Z(Y-g_0(1,X))^2 + (1-Z)(Y-g_0(0,X))^2\right]}{(1-\varepsilon)^2} \\
&= \frac{\mathrm{E}_P[u^2]}{(1-\varepsilon)^2} \geq \frac{c_0^2}{(1-\varepsilon)^2},
\end{aligned}$$

where the first equality holds since the interaction term equals to zero by the equations in (B.6), the third inequality follows from the facts that $\varepsilon \leq p_0(X) \leq 1-\varepsilon$, and the last equality follows from Assumption 3 (iv). Thus the Assumption 2 (iii) is satisfied.

Step 3. Next, we verify Assumption 2 (i). By Lemmas 2 and 3 invoked by Assumption 3-7, with probability $1 - o(1)$,

$$\|(\widehat{\beta}_{11},\widehat{\beta}_{12}) - (\beta_{11}^0,\beta_{12}^0)\| \vee \|(\widehat{\beta}_{21},\widehat{\beta}_{22}) - (\beta_{21}^0,\beta_{22}^0)\| \vee \|\widehat{\gamma} - \gamma^0\| \lesssim \sqrt{\frac{s_N \log(pN)}{N}}.$$

The proof of Lemmas 2 and 3 are given in Section B.4 and B.5. Thus Assumption 2 (i) is satisfied.

Step 4. Next, let us verify the condition in Assumption 2 (ii). Note that

$$\begin{aligned}\|g_0(D,X)\|_{P,q} &= (\mathrm{E}_P[|g_0(D,X)|^q])^{1/q} \\ &\geq (\mathrm{E}_P[|g_0(1,X)|^q P(D=1|X) + |g_0(0,X)|^q P(D=0|X)])^{1/q} \\ &\geq \varepsilon^{1/q}(\mathrm{E}_P[|g_0(1,X)|^q] + \mathrm{E}_P[|g_0(0,X)|^q])^{1/q} \\ &\geq \varepsilon^{1/q}(\mathrm{E}_P[|g_0(1,X)|^q] \vee \mathrm{E}_P[|g_0(0,X)|^q])^{1/q} \\ &\geq \varepsilon^{1/q}(\|g_0(1,X)\|_{P,q} \vee \|g_0(0,X)\|_{P,q}).\end{aligned}$$

Since $\|g_0(D,X)\|_{P,q} \leq \|Y\|_{P,q} \leq c_1$ by Assumption 3, we have

$$\|g_0(1,X)\|_{P,q} \leq c_1/\varepsilon^{1/q}, \text{and} \|g_0(0,X)\|_{P,q} \leq c_1/\varepsilon^{1/q}.$$

By using similar arguments, we obtain

$$\begin{aligned}&\|g(1,X) - g_0(1,X)\|_{P,q} \leq c_1/\varepsilon^{1/q}, \quad \|g(0,X) - g_0(0,X)\|_{P,q} \leq c_1/\varepsilon^{1/q}, \\ &\|m_0(1,X)\|_{P,q} \leq 1/\varepsilon^{1/q}, \quad \|m_0(0,X)\|_{P,q} \leq 1/\varepsilon^{1/q}, \\ &\|m(1,X) - m_0(1,X)\|_{P,q} \leq c_1/\varepsilon^{1/q}, \quad \|m(0,X) - m_0(0,X)\|_{P,q} \leq c_1/\varepsilon^{1/q},\end{aligned} \tag{B.7}$$

since $\|m_0(D,X)\|_{P,q} \leq 1$, $\|g(D,X) - g_0(D,X)\|_{P,q} \leq c_1$, and $\|m(Z,X) - m_0(Z,X)\|_{P,q} \leq c_1$. By calculation, we obtain

$$\begin{aligned}&\|\psi^a(W;\eta)\|_{P,q} \leq (1+\varepsilon^{-1})(\|m(1,X)\|_{P,q} + \|m(0,X)\|_{P,q}) + 2/\varepsilon \\ &\leq (1+\varepsilon^{-1})(\|m(1,X) - m_0(1,X)\|_{P,q} + \|m_0(1,X)\|_{P,q} + \|m(0,X) - m_0(0,X)\|_{P,q} + \|m_0(0,X)\|_{P,q}) + 2/\varepsilon \\ &\leq (1+\varepsilon^{-1})(2c_1\varepsilon^{-1/q} + 2\varepsilon^{-1/q}) + 2\varepsilon^{-1} := c_{\varepsilon 1}, \\ &\|\psi^b(W;\eta)\|_{P,q} \leq (1+\varepsilon^{-1})(\|g(1,X)\|_{P,q} + \|g(0,X)\|_{P,q}) + 2\|Y\|_{P,q}/\varepsilon \\ &\leq (1+\varepsilon^{-1})(2c_1\varepsilon^{-1/q} + 2\varepsilon^{-1/q}) + 2c_1\varepsilon^{-1} := c_{\varepsilon 2},\end{aligned}$$

where $c_{\varepsilon 1}$ and $c_{\varepsilon 2}$ are constants related with ε instead of N. Note that this completes the verification of Assumption 2 (b) Therefore, under the null, we have

$$\begin{aligned}&(\mathrm{E}_P[\|\psi(W;\theta,\eta)\|^q])^{1/q} = \|\psi(W;\theta,\eta)\|_{P,q} \leq \|\psi(W;\theta,\eta) - \psi(W;\theta_0,\eta)\|_{P,q} + \|\psi(W;\theta_0,\eta)\|_{P,q} \\ &\leq |\theta - \theta_0| \times \|\psi^a(W;\eta)\|_{P,q} + \|\psi^b(W;\eta)\|_{P,q} + |\theta_0| \times \|\psi^a(W;\eta)\|_{P,q} \\ &\leq |\theta - \theta_0| c_{\varepsilon 1} + c_{\varepsilon 2} + |\theta_0| c_{\varepsilon 1} \lesssim 1,\end{aligned}$$

where the last inequality need the assumption that Θ is a compact set by Assumption 3 (vi). This completes the verification of Assumption 2 (ii) (a).

Next, let us verify the condition in Assumption 2 (ii) (c). For any $\eta = (g, m, p)$, by the triangle inequality,

$$(\mathrm{E}_P[\|\psi(W;\theta,\eta) - \psi(W;\theta,\eta_0)\|^2])^{1/2} = \|\psi(W;\theta,\eta) - \psi(W;\theta,\eta_0)\|_{P,2} \leq \mathcal{I}_1 + \mathcal{I}_2 + \mathcal{I}_3 + \mathcal{I}_4,$$

where

$$\begin{aligned}
\mathcal{I}_1 &:= \|g(1,X) - g_0(1,X)\|_{P,2} + \|g(0,X) - g_0(0,X)\|_{P,2}, \\
\mathcal{I}_2 &:= |\theta| \times (\|m(1,X) - m_0(1,X)\|_{P,2} + \|m(0,X) - m_0(0,X)\|_{P,2}), \\
\mathcal{I}_3 &:= \left\| \frac{Z(Y - g(1,X))}{p(X)} - \frac{Z(Y - g_0(1,X))}{p_0(X)} \right\|_{P,2} + \left\| \frac{(1-Z)(Y - g(0,X))}{1-p(X)} - \frac{(1-Z)(Y - g_0(0,X))}{1-p_0(X)} \right\|_{P,2}, \\
\mathcal{I}_4 &:= |\theta| \times \left(\left\| \frac{Z(D - m(1,X))}{p(X)} - \frac{Z(D - m_0(1,X))}{p_0(X)} \right\|_{P,2} + \left\| \frac{(1-Z)(D - m(0,X))}{1-p(X)} - \frac{(1-Z)(D - m_0(0,X))}{1-p_0(X)} \right\|_{P,2} \right).
\end{aligned}$$

By using the similar argument as the one in obtaining equation (B.7), we have

$$\begin{aligned}
&\|g(1,X) - g_0(1,X)\|_{P,2} \leq \delta_N/\varepsilon^{1/q}, \quad \|g(0,X) - g_0(0,X)\|_{P,2} \leq \delta_N/\varepsilon^{1/q}, \\
&\|m(1,X) - m_0(1,X)\|_{P,2} \leq \delta_N/\varepsilon^{1/q}, \quad \|m(0,X) - m_0(0,X)\|_{P,2} \leq \delta_N/\varepsilon^{1/q}.
\end{aligned}$$

so $\mathcal{I}_1 \leq 2\delta_N/\varepsilon^{1/q}$ and $\mathcal{I}_2 \lesssim 2\delta_N/\varepsilon^{1/q}$. To bound $\mathcal{I}_3$, we have

$$\begin{aligned}
\mathcal{I}_3 &\leq \varepsilon^{-2} \times \Big(\|Zp_0(X)(Y - g(1,X)) - Zp(X)(Y - g_0(1,X))\|_{P,2} \\
&\quad + \|(1-Z)(1-p_0(X))(Y - g(0,X)) - (1-Z)(1-p(X))(Y - g_0(0,X))\|_{P,2} \Big) \\
&\leq \varepsilon^{-2} \times \Big(\|p_0(X)(u + g_0(1,X) - g(1,X)) - p(X)u\|_{P,2} \\
&\quad + \|(1-p_0(X))(u + g_0(0,X) - g(0,X)) - (1-p(X))u\|_{P,2} \Big) \\
&\leq \varepsilon^{-2} \times \Big(\|p_0(X)(g_0(1,X) - g(1,X))\|_{P,2} + \|(p(X) - p_0(X))u\|_{P,2} \\
&\quad + \|(1-p_0(X))(g_0(0,X) - g(0,X))\| + \|(p(X) - p_0(X))u\|_{P,2} \Big) \\
&\leq \varepsilon^{-2} \times \Big(\|(g_0(1,X) - g(1,X))\|_{P,2} + \sqrt{c_1}\|p(X) - p_0(X)\|_{P,2} \\
&\quad + \|(g_0(0,X) - g(0,X))\| + \sqrt{c_1}\|p(X) - p_0(X)\|_{P,2} \Big) \\
&\leq \varepsilon^{-2} \times (2/\varepsilon^{1/q} + 2\sqrt{c_1})\delta_N \leq c_{\varepsilon_3}\delta_N,
\end{aligned}$$

where $c_{\varepsilon_3} \geq \varepsilon^{-2} \times (2/\varepsilon^{1/q} + 2\sqrt{c_1})$, the first inequality follows from $\varepsilon \leq p(X) \leq 1-\varepsilon$ and $\varepsilon \leq 1 - p(X) \leq 1 - \varepsilon$, and the fourth one follows from Assumption 3 (vi). We use the similar argument to bound $\mathcal{I}_4$ and obtain that $\mathcal{I}_4 \lesssim \delta_N$. Therefore, we have $\|\psi(W;\theta,\eta) - \psi(W;\theta,\eta_0)\|_{P,2} \lesssim \delta_N$, which completes the verification of Assumption 2 (ii). □

B.3 Proof of Theorem 3

Proof. Theorem 2 shows that under Assumption 3 and $H_0 : \theta \in \Theta$, $\widehat{\mathbb{G}}_N(\cdot)$ weakly converges to a centered Gaussian process, and the variance estimator $\widehat{\Omega}(\theta_1, \theta_2)$ is a uniformly consistent estimator for $\Omega(\theta_1, \theta_2)$. The proof of Theorem 3 follows trivially from equation (3.5) and the fact that the distribution of $q_N(\theta_0) \sim N(0, \Omega(\theta_0, \theta_0))$ does not depend on S_N, the function $\widehat{\Omega}(\theta, \theta_0)\widehat{\Omega}(\theta_0, \theta_0)^{-1}$ is deterministic and known, and the definition of $c_\alpha(h_N)$ by equation (2.8). □

B.4 Proof of Lemma 2

Proof. We apply Lemma 1 in Belloni et al. (2016). In step 1, we will verify the condition for Lemma 1 in Belloni et al. (2016) holds. In step 2, we obtain a high-probability bound for λ_1 and λ_2. First, note that Assumption 4 implies that the restricted eigenvalue condition holds with probability $1-o(1)$ by Lemma 2.7 in Lecué and Mendelson (2017): for $T = \text{supp}(\beta_{11}^0, \beta_{12}^0)$, $|T| \geq 1$, and $c \geq 1$, we have $\kappa_{c_0} = \inf_{\delta \in \mathcal{D}_{c_0}} \frac{\|(Z_i, X_i')\delta\|_{2,N}}{\|\delta_T\|_1} > 0$, where $\mathcal{D}_{c_0} = \{\delta : \|\delta_{T^c}\| \leq c_0 \|\delta_T\|_1\}$ with $c_0 = (c+1)/(c-1)$.

Step 1. For a subset $A \subset \mathbb{R}^{p+1}$, define the nonlinear impact coefficient by

$$\bar{q}_A = \inf_{\delta \in A} \frac{\mathbb{E}_N\left[|(Z_i, X_i')\delta|^2\right]^{3/2}}{\mathbb{E}_N\left[|(Z_i, X_i')\delta|^3\right]}.$$

To apply Lemma 1 in Belloni et al. (2016), we verify the condition $\bar{q}_{\mathcal{D}_{c_0}} > 3\left(1 + \frac{1}{c}\right)\lambda_1\sqrt{s_N}/\left(N\kappa_{c_0}\right)$ with probability $1 = o(1)$. Observe that

$$\begin{aligned}
\bar{q}_{\mathcal{D}_{c_0}} &= \inf_{\delta \in \mathcal{D}_{c_0}} \frac{\mathbb{E}_N\left[|(Z_i, X_i')\delta|^2\right]^{3/2}}{\mathbb{E}_N\left[|(Z_i, X_i')\delta|^3\right]} \geq \inf_{\delta \in \mathcal{D}_{c_0}} \frac{\mathbb{E}_N\left[|(Z_i, X_i')\delta|^2\right]^{1/2}}{\max_{i \in [N]} \|(Z_i, X_i')\|_\infty \|\delta\|_1} \gtrsim_P \inf_{\delta \in \mathcal{D}_{c_0}} \frac{\mathbb{E}_N\left[|(Z_i, X_i')\delta|^2\right]^{1/2}}{N^{1/q} M_N \|\delta\|_1} \\
&\geq \inf_{\delta \in \mathcal{D}_{c_0}} \frac{\mathbb{E}_N\left[|(Z_i, X_i')\delta|^2\right]^{1/2}}{N^{1/q} M_N (1+c_0)\sqrt{s_N}\|\delta_T\|} \geq \frac{\kappa_{c_0}}{N^{1/q} M_N (1+c_0)\sqrt{s_N}} \geq \frac{1}{\Delta_N^{1/2} N^{1/4}} \gtrsim \sqrt{\frac{s_N \log a_N}{\Delta_N N}},
\end{aligned}$$

where the fourth inequality follows from the definition of κ_{c_0}, and the fifth comes from $\Delta_N \geq M_N s_N/N^{1/2-2/q}$ and the last one from $s_N \log a_N/N^{1/2} \leq \Delta_N$ by Assumption 7(iii), $\Delta_N = o(1)$, and $\lambda_1 = \sqrt{N \log a_N}$. Therefore, we obtain,

$$\|Z_i(\widetilde{\beta}_{11} - \beta_{11}^0) + X_i'(\widetilde{\beta}_{12} - \beta_{12}^0)\|_{2,N} = O(\frac{\lambda_1\sqrt{s_N}}{N}), \quad \|(\widetilde{\beta}_{11}, \widetilde{\beta}_{12}) - (\beta_{11}^0, \beta_{12}^0)\|_1 = O(\frac{\lambda_1 s_N}{N}).$$

Step 2. In this step, we show for some large $K > 0$, let $\zeta \in (0, 1)$ and

$$\lambda_1 = K\sqrt{N \log(p/\zeta)},$$

then with probability $1 - \zeta - o(1)$, for a fixed $c > 1$, it holds that

$$P(\lambda_1/N \geq c\|\nabla L_1(\beta_{11}^0, \beta_{12}^0)\|_\infty) \geq 1 - \zeta - o(1).$$

The proof relies on Theorem 2.1 and 2.2 in Chernozhukov et al. (2013). We need to verify the conditions in Chernozhukov et al. (2013). Conditions are directly implied by Assumption 7 (i)(ii). Now, by Theorem 2.1 and 2.2 in Chernozhukov et al. (2013), we have

$$\sup_{t\in\mathbb{R}} |P(\|\sqrt{N}\nabla L_1(\beta_{11}^0, \beta_{12}^0)\|_\infty \leq t) - P(\|\mathbf{G}\|_\infty \leq t)| = o(1),$$

where $\mathbf{G} \sim N(0, \Sigma)$, Σ is the asymptotic variance of $\sqrt{N}\nabla L_1(\beta_{11}^0, \beta_{12}^0)$. Then the Gaussian concentration inequality implies that with probability $1 - \zeta - o(1)$,

$$P(\lambda_1/N \geq c\|\nabla L_1(\beta_{11}^0, \beta_{12}^0)\|_\infty) \geq 1 - \zeta - o(1).$$

Now, combining the result with the bound from Step 1 concludes the convergence rate for $(\widehat{\beta}_{11}, \widehat{\beta}_{12})$. Replacing λ_1, (Z_i, X_i'), (β_{11}, β_{12}), $L_1(\beta_{11}, \beta_{12})$ by λ_2, X_i', γ, $L_2(\gamma)$ respectively, we could obtain the convergence rate for $\widehat{\gamma}$. □

B.5 Proof of Lemma 3

Proof. The proof replies on Theorem 1 in Belloni and Chernozhukov (2013). We need to verify the conditions in Theorem 1 in Belloni and Chernozhukov (2013) hold. Note that Assumption 4 directly implies the restricted eigenvalue condition in Belloni and Chernozhukov (2013). Condition V in Belloni and Chernozhukov (2013) follows from Assumption 3 (iv) and (v). Therefore, with the choice of $\lambda_3 = K_3\sqrt{N \log(pN)}$, we have with probability $1 - o(1)$, $\|(\widehat{\beta}_{21}, \widehat{\beta}_{22}) - (\beta_{21}^0, \beta_{22}^0)\| \lesssim \sqrt{\frac{s_N \log(pN)}{N}}$. □

References

Abadie, A. (2003): "Semiparametric instrumental variable estimation of treatment response models," *Journal of econometrics*, 113, 231–263.

Acemoglu, D. and J. Angrist (2000): "How large are human-capital externalities? Evidence from compulsory schooling laws," *NBER macroeconomics annual*, 15, 9–59.

Aizer, A. and J. J. Doyle Jr (2015): "Juvenile incarceration, human capital, and future crime: Evidence from randomly assigned judges," *The Quarterly Journal of Economics*, 130, 759–803.

Ambrus, A., E. Field, and R. Gonzalez (2020): "Loss in the time of cholera: Long-run impact of a disease epidemic on the urban landscape," *American Economic Review*, 110, 475–525.

Andrews, D. and J. H. Stock (2005): "Inference with weak instruments," .

Andrews, D. W. (1994): "Asymptotics for semiparametric econometric models via stochastic equicontinuity," *Econometrica: Journal of the Econometric Society*, 43–72.

Andrews, D. W. and P. Guggenberger (2019): "Identification-and singularity-robust inference for moment condition models," *Quantitative Economics*, 10, 1703–1746.

Andrews, D. W., M. J. Moreira, and J. H. Stock (2006): "Optimal two-sided invariant similar tests for instrumental variables regression," *Econometrica*, 74, 715–752.

Andrews, I. and A. Mikusheva (2016): "Conditional inference with a functional nuisance parameter," *Econometrica*, 84, 1571–1612.

Andrews, I., J. H. Stock, and L. Sun (2019): "Weak instruments in instrumental variables regression: Theory and practice," *Annual Review of Economics*, 11, 727–753.

Angrist, J. and G. Imbens (1995a): "Identification and estimation of local average treatment effects," .

Angrist, J., V. Lavy, and A. Schlosser (2010): "Multiple experiments for the causal link between the quantity and quality of children," *Journal of Labor Economics*, 28, 773–824.

Angrist, J. D. (2022): "Empirical strategies in economics: Illuminating the path from cause to effect," *Econometrica*, 90, 2509–2539.

Angrist, J. D., K. Graddy, and G. W. Imbens (2000): "The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish," *The Review of Economic Studies*, 67, 499–527.

Angrist, J. D. and G. W. Imbens (1995b): "Two-stage least squares estimation of average causal effects in models with variable treatment intensity," *Journal of the American statistical Association*, 90, 431–442.

Angrist, J. D. and A. B. Krueger (1991): "Does compulsory school attendance affect schooling and earnings?" *The Quarterly Journal of Economics*, 106, 979–1014.

Belloni, A. and V. Chernozhukov (2013): "Least squares after model selection in high-dimensional sparse models," *Bernoulli*, 19, 521–547.

Belloni, A., V. Chernozhukov, D. Chetverikov, and Y. Wei (2018): "Uniformly valid post-regularization confidence regions for many functional parameters in z-estimation framework," *Annals of statistics*, 46, 3643.

Belloni, A., V. Chernozhukov, I. Fernandez-Val, and C. Hansen (2017): "Program evaluation and causal inference with high-dimensional data," *Econometrica*, 85, 233–298.

Belloni, A., V. Chernozhukov, and C. Hansen (2014): "High-dimensional methods and inference on structural and treatment effects," *Journal of Economic Perspectives*, 28, 29–50.

Belloni, A., V. Chernozhukov, and K. Kato (2015): "Uniform post-selection inference for least absolute deviation regression and other Z-estimation problems," *Biometrika*, 102, 77–94.

Belloni, A., V. Chernozhukov, and Y. Wei (2016): "Post-selection inference for generalized linear models with many controls," *Journal of Business & Economic Statistics*, 34, 606–619.

Benzell, S. G. and K. Cooke (2021): "A network of thrones: Kinship and conflict in europe, 1495–1918," *American Economic Journal: Applied Economics*, 13, 102–133.

Bickel, P. J., C. A. Klaassen, P. J. Bickel, Y. Ritov, J. Klaassen, J. A. Wellner, and Y. Ritov (1993): *Efficient and adaptive estimation for semiparametric models*, vol. 4, Springer.

Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009): "Simultaneous analysis of Lasso and Dantzig selector," .

Bisbee, J., R. Dehejia, C. Pop-Eleches, and C. Samii (2017): "Local instruments, global extrapolation: External validity of the labor supply–fertility local average treatment effect," *Journal of Labor Economics*, 35, S99–S147.

Bound, J., D. A. Jaeger, and R. M. Baker (1995): "Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak," *Journal of the American statistical association*, 90, 443–450.

Card, D. (1993): "Using geographic variation in college proximity to estimate the return to schooling," .

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018): "Double/debiased machine learning for treatment and structural parameters," *Econometrics Journal*, 21, C1 – C68.

Chernozhukov, V., D. Chetverikov, and K. Kato (2013): "Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors," *The Annals of Statistics*, 41, 2786–2819.

——— (2016): "Empirical and multiplier bootstraps for suprema of empirical processes of increasing complexity, and related Gaussian couplings," *Stochastic Processes and their Applications*, 126, 3632–3651.

——— (2017): "Central limit theorems and bootstrap in high dimensions," *The Annals of Probability*, 45, 2309–2352.

Chernozhukov, V., J. C. Escanciano, H. Ichimura, W. K. Newey, and J. M. Robins (2022): "Locally robust semiparametric estimation," *Econometrica*, 90, 1501–1535.

Dahl, G. B., A. R. Kostøl, and M. Mogstad (2014): "Family welfare cultures," *The Quarterly Journal of Economics*, 129, 1711–1752.

French, E. and J. Song (2014): "The effect of disability insurance receipt on labor supply," *American economic Journal: economic policy*, 6, 291–337.

Frölich, M. (2007): "Nonparametric IV estimation of local average treatment effects with covariates," *Journal of Econometrics*, 139, 35–75.

Galiani, S., M. A. Rossi, and E. Schargrodsky (2011): "Conscription and crime: Evidence from the Argentine draft lottery," *American Economic Journal: Applied Economics*, 3, 119–136.

Hahn, J., P. Todd, and W. Van der Klaauw (2001): "Identification and estimation of treatment effects with a regression-discontinuity design," *Econometrica*, 69, 201–209.

Hirano, K., G. W. Imbens, D. B. Rubin, and X.-H. Zhou (2000): "Assessing the effect of an influenza vaccine in an encouragement design," *Biostatistics*, 1, 69–88.

Hornung, E. (2015): "Railroads and growth in Prussia," *Journal of the European Economic Association*, 13, 699–736.

Kane, T. J. and C. E. Rouse (1993): "Labor market returns to two-and four-year colleges: is a credit a credit and do degrees matter?" .

Kleibergen, F. (2002): "Pivotal statistics for testing structural parameters in instrumental variables regression," *Econometrica*, 70, 1781–1803.

——— (2005): "Testing parameters in GMM without assuming that they are identified," *Econometrica*, 73, 1103–1123.

Kling, J. R. (2006): "Incarceration length, employment, and earnings," *American Economic Review*, 96, 863–876.

Lecué, G. and S. Mendelson (2017): "Sparse recovery under weak moment assumptions," *Journal of the European Mathematical Society*, 19, 881–904.

Lee, D. S., J. McCrary, M. J. Moreira, and J. Porter (2022): "Valid t-ratio Inference for IV," *American Economic Review*, 112, 3260–90.

Linton, O. (1996): "Edgeworth approximation for MINPIN estimators in semiparametric regression models," *Econometric Theory*, 12, 30–60.

Maestas, N., K. J. Mullen, and A. Strand (2013): "Does disability insurance receipt discourage work? Using examiner assignment to estimate causal effects of SSDI receipt," *American economic review*, 103, 1797–1829.

Mikusheva, A. and L. Sun (2022): "Inference with many weak instruments," *The Review of Economic Studies*, 89, 2663–2686.

Moreira, H. and M. J. Moreira (2019): "Optimal two-sided tests for instrumental variables regression with heteroskedastic and autocorrelated errors," *Journal of Econometrics*, 213, 398–433.

Moreira, M. J. (2003): "A conditional likelihood ratio test for structural models," *Econometrica*, 71, 1027–1048.

——— (2009): "Tests with correct size when instruments can be arbitrarily weak," *Journal of Econometrics*, 152, 131–140.

Moser, P., A. Voena, and F. Waldinger (2014): "German Jewish émigrés and US invention," *American Economic Review*, 104, 3222–3255.

Newey, W. K. (1990): "Semiparametric efficiency bounds," *Journal of applied econometrics*, 5, 99–135.

——— (1994): "The asymptotic variance of semiparametric estimators," *Econometrica: Journal of the Econometric Society*, 1349–1382.

Oreopoulos, P. (2006): "Estimating average and local average treatment effects of education when compulsory schooling laws really matter," *American Economic Review*, 96, 152–175.

Pfanzagl, J. and W. Wefelmeyer (1985): "Contributions to a general asymptotic statistical theory," *Statistics & Risk Modeling*, 3, 379–388.

Robins, J. M. and A. Rotnitzky (1995): "Semiparametric efficiency in multivariate regression models with missing data," *Journal of the American Statistical Association*, 90, 122–129.

Staiger, D. O. and J. H. Stock (1994): "Instrumental variables regression with weak instruments," .

Stock, J. H. and J. H. Wright (2000): "GMM with weak identification," *Econometrica*, 68, 1055–1096.

Stock, J. H., J. H. Wright, and M. Yogo (2002): "A survey of weak instruments and weak identification in generalized method of moments," *Journal of Business & Economic Statistics*, 20, 518–529.

Stock, J. H. and M. Yogo (2002): "Testing for weak instruments in linear IV regression," .

Tan, Z. (2006): "Regression and weighting methods for causal inference using instrumental variables," *Journal of the American Statistical Association*, 101, 1607–1618.

Tsiatis, A. A. (2006): "Semiparametric theory and missing data," .

Yau, L. H. and R. J. Little (2001): "Inference for the complier-average causal effect from longitudinal data subject to noncompliance and missing data, with application to a job training assessment for the unemployed," *Journal of the American Statistical Association*, 96, 1232–1244.