Journal of Econometrics 207 (2018) 175–187

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Testing endogeneity with high dimensional covariates

Zijian Guo [a,*], Hyunseung Kang [b], T. Tony Cai [c], Dylan S. Small [c]

[a] Department of Statistics and Biostatistics, Rutgers University, United States
[b] Department of Statistics, University of Wisconsin-Madison, United States
[c] Department of Statistics, The Wharton School, University of Pennsylvania, United States

ARTICLE INFO

Article history:
Received 22 March 2017
Received in revised form 9 March 2018
Accepted 14 July 2018
Available online 10 August 2018

JEL classification:
C12
C36

Keywords:
Durbin–Wu–Hausman test
Endogeneity test
High dimensions
Instrumental variable
Invalid instruments
Power function

ABSTRACT

Modern, high dimensional data has renewed investigation on instrumental variables (IV) analysis, primarily focusing on estimation of effects of endogenous variables and putting little attention towards specification tests. This paper studies in high dimensions the Durbin–Wu–Hausman (DWH) test, a popular specification test for endogeneity in IV regression. We show, surprisingly, that the DWH test maintains its size in high dimensions, but at an expense of power. We propose a new test that remedies this issue and has better power than the DWH test. Simulation studies reveal that our test achieves near-oracle performance to detect endogeneity.

1. Introduction

1.1. Endogeneity testing with high dimensional data

Recent growth in both the size and dimension of the data has led to a resurgence in analyzing instrumental variables (IV) regression in high dimensional settings (Belloni et al., 2011a, 2012, 2013; Chernozhukov et al., 2014, 2015; Fan and Liao, 2014; Gautier and Tsybakov, 2011) where the number of regression parameters, especially those associated with exogenous covariates, is growing with, and may exceed, the sample size.[1] The primary focus in these works has been providing tools for estimation and inference of a single endogenous variable's effect on the outcome under some low-dimensional structural assumptions on the structural parameters associated with the instruments and the covariates, such as sparsity. (Belloni et al., 2011a, 2012, 2013; Chernozhukov et al., 2014, 2015; Gautier and Tsybakov, 2011). This line of work has generally not focused on specification tests in the high dimensional IV setting.

The main goal of this paper is to study the high dimensional behavior of one of the most common specification tests in IV regression, the test for endogeneity, which assumes the validity of the IV and tests whether the included endogenous variable (e.g., a treatment variable) is actually exogenous. Historically, the most widely used test for endogeneity is the Durbin–Wu–Hausman test (Durbin, 1954; Hausman, 1978; Wu, 1973), hereafter called the DWH test, and is widely implemented in

* Corresponding author.
E-mail address: zijguo@stat.rutgers.edu (Z. Guo).

[1] In the paper, we use the term "high dimensional setting" more broadly where the number of parameters is growing with the sample size; see Sections 3 and 4.3 for details and examples. Note that the modern usage of the term "high dimensional setting" where the sample size exceeds the parameter is one case of this broader setting.

https://doi.org/10.1016/j.jeconom.2018.07.002
0304-4076/

software, such as ivreg2 in Stata (Baum et al., 2007). The DWH test detects the presence of endogeneity in the structural model by studying the difference between the ordinary least squares (OLS) estimate of the structural parameters in the IV regression to that of the two-stage least squares (TSLS) under the null hypothesis of no endogeneity; see Section 2.3 for the exact characterization of the DWH test. In low dimensional settings, the primary requirements for the DWH test to correctly control Type I error are having instruments that are (i) strongly associated with the included endogenous variable, often called strong instruments, and (ii) exogenous to the structural errors,[2] often referred to as valid instruments (Murray, 2006). When instruments are not strong, Staiger and Stock (1997) showed that the DWH test that used the TSLS estimator for variance, developed by Durbin (1954) and Wu (1973), had distorted size under the null hypothesis while the DWH test that used the OLS estimator for variance, developed by Hausman (1978), had proper size. When instruments are invalid, which is perhaps a bigger concern in practice (Conley et al., 2012; Murray, 2006), the DWH test will usually fail because the TSLS estimator is inconsistent under the null hypothesis; see the Supplementary materials for a simple theoretical justification of this phenomenon. Indeed, some recent work with high dimensional data (Belloni et al., 2012; Chernozhukov et al., 2015) advocated conditioning on many, possibly high dimensional, exogenous covariates to make instruments more plausibly valid.[3] However, while adding additional covariates can potentially make instruments more plausibly valid, it is unclear what price one has to pay with respect to the performance of specification tests like the DWH test.

1.2. Prior work and contribution

Prior work in analyzing the DWH test in instrumental variables is diverse. Estimation and inference under weak and/or many instruments are well documented (Andrews et al., 2007; Bekker, 1994; Bound et al., 1995; Chao and Swanson, 2005; Dufour, 1997; Han and Phillips, 2006; Hansen et al., 2008; Kleibergen, 2002; Moreira, 2003; Morimune, 1983; Nelson and Startz, 1990; Newey and Windmeijer, 2005; Staiger and Stock, 1997; Stock and Yogo, 2005; Wang and Zivot, 1998; Zivot et al., 1998). In particular, when the instruments are weak, the behavior of the DWH test under the null depends on the variance estimate (Doko Tchatoka, 2015; Nakamura and Nakamura, 1981; Staiger and Stock, 1997). Other works study the behavior of the DWH test under different strengths of instruments and/or weak instrument asymptotics (Hahn et al., 2011; Staiger and Stock, 1997) and under a two-stage testing scheme (Guggenberger, 2010). Some recent work extended the specification test to handle growing number of instruments (Chao et al., 2014; Hahn and Hausman, 2002; Lee and Okui, 2012). Other recent works extended specification tests based on overidentification (Hahn and Hausman, 2005; Hausman et al., 2005) and to heteroskedastic data (Chmelarova et al., 2007). Fan et al. (2015) considered testing endogeneity in the high dimensional non-IV setting and approximated the null distribution of their test statistic by the bootstrap; the distribution under the alternative was not identifiable. None of these works have characterized the properties of the DWH test used in IV regression under the high dimensional setting.

Our main contributions are two-fold. First, we characterize the behavior of the popular DWH test in high dimensions. The theoretical analysis reveals that the DWH test actually controls Type I error at the correct level in high dimensions, but pays a significant price with respect to power, especially for small to moderate degrees of endogeneity; we also confirm our finding numerically with a simulation study of the empirical power of the DWH test. Our finding also suggests that, although conditioning on a large number of covariates makes instruments more plausibly valid, the power of the DWH test is reduced because of the large number of covariates. Second, we remedy the low power of the DWH test by presenting a simple and improved endogeneity test that is robust to high dimensional covariates and/or instruments and that works in settings where the number of structural parameters is allowed to exceed the sample size. In particular, our new endogeneity test applies a hard thresholding procedure to popular estimators for reduced-form models, such as OLS in low dimensions or bias-corrected Lasso estimators in high dimensions (see Section 4.1 for details). This hard thresholding procedure is an essential step of the new endogeneity test, where relevant instruments are selected for testing endogeneity. We also highlight that the success of the proposed endogeneity test does not require the correct selection of all relevant instruments. That is, even if the relevant instruments are not correctly selected, the proposed testing procedure still controls Type I error and achieves non-trivial power under regularity conditions. Additionally, we briefly discuss an extension of our endogeneity test to incorporate invalid instruments, especially when many covariates are conditioned upon to avoid invalid IVs.

This paper is closely connected to the paper Guo et al. (2016) by the same group of authors, where Guo et al. (2016) proposed confidence intervals for the treatment effect in the presence of both high-dimensional instruments, covariates and invalid instruments. The current paper considers a related but different problem about endogeneity testing and extends the idea proposed in Guo et al. (2016) to testing endogeneity in high dimensional settings. In particular, we are the first to

[2] The term exogeneity is sometimes used in the IV literature to encompass two assumptions, (a) independence of the IVs to the disturbances in the structural model and (b) IVs having no direct effect on the outcome, sometimes referred to as the exclusion restriction (Angrist et al., 1996; Holland, 1988; Imbens and Angrist, 1994). As such, an instrument that is perfectly randomized from a randomized experiment may not be exogenous in the sense that while the instrument is independent to any structural error terms, the instrument may still have a direct effect on the outcome.

[3] For example, in Section 7 of the empirical example of Belloni et al. (2012), the authors studied the effect of federal appellate court decisions on economic outcomes by using the random assignment of judges to decide appellate cases. They state that once the distribution of characteristics of federal circuit court judges in a given circuit-year is controlled for, "the realized characteristics of the randomly assigned three-judge panel should be unrelated to other factors besides judicial decisions that may be related to economic outcomes" (page 2405). More broadly, in empirical practice, adding covariates to make IVs more plausibly valid is commonplace; see Card (1999), Cawley et al. (2013), and Kosec (2014) for examples as well as review papers in epidemiology and causal inference by Hernán and Robins (2006) and Baiocchi et al. (2014).

provide a test for endogeneity when $n < p$. In addition, the characterization of the power of DWH test in high dimensions and the technical tools used in the paper are new. In particular, the technical tools can be used to study other specification tests like the Sargan or the J test (e.g. Sargan test (Hansen, 1982; Sargan, 1958) in high dimensions.

We conduct simulation studies comparing the performance of our new test with the usual DWH test and apply the proposed endogeneity test to an empirical data analysis following Belloni et al. (2012, 2014). We find that our test has the desired size and has better power than the DWH test for all degrees of endogeneity and performs similarly to the oracle DWH test which knows the support of relevant instruments and covariates a priori. In the supplementary materials, we also present technical proofs and extended simulation studies that further examine the power of our test.

2. Instrumental variables regression and the DWH Test

2.1. Notation

For any vector $v \in \mathbb{R}^p$, v_j denotes its jth element, and $\|v\|_1$, $\|v\|_2$, and $\|v\|_\infty$ denote the 1, 2 and ∞-norms, respectively. Let $\|v\|_0$ denote the number of non-zero elements in v and define $\text{supp}(v) = \{j : v_j \neq 0\} \subseteq \{1, \ldots, p\}$. For any $n \times p$ matrix M, denote the (i, j) entry by M_{ij}, the ith row by $M_{i.}$, the jth column by $M_{.j}$, and the transpose of M by M'. Also, given any $n \times p$ matrix M with sets $I \subseteq \{1, \ldots, n\}$ and $J \subseteq \{1, \ldots, p\}$ denote M_{IJ} as the submatrix of M consisting of rows specified by the set I and columns specified by the set J, $M_{I.}$ as the submatrix of M consisting of rows indexed by the set I and all columns, and $M_{.J}$ as the submatrix of M consisting of columns specified by the set J and all rows. Also, for any $n \times p$ full-rank matrix M, define the orthogonal projection matrices $P_M = M(M'M)^{-1}M'$ and $P_{M^\perp} = \mathrm{I} - M(M'M)^{-1}M'$ where $P_M + P_{M^\perp} = \mathrm{I}$ and I is an identity matrix. For a $p \times p$ matrix Λ, $\Lambda \succ 0$ denotes that Λ is a positive definite matrix. For any $p \times p$ positive definite Λ and set $J \subseteq \{1, \ldots, p\}$, let $\Lambda_{J|J^C} = \Lambda_{JJ} - \Lambda_{JJ^C}\Lambda^{-1}_{J^CJ^C}\Lambda_{J^CJ}$ denote the submatrix Λ_{JJ} adjusted for the columns in the complement of the set J, J^C.

For a sequence of random variables X_n indexed by n, we use $X_n \xrightarrow{p} X$ to represent that X_n converges to X in probability. For a sequence of random variables X_n and numbers a_n, we define $X_n = o_p(a_n)$ if X_n/a_n converges to zero in probability and $X_n = O_p(a_n)$ if for every $c_0 > 0$, there exists a finite constant C_0 such that $\mathbf{P}(|X_n/a_n| \geq C_0) \leq c_0$. For any two sequences of numbers a_n and b_n, we will write $b_n \ll a_n$ if $\limsup b_n/a_n = 0$.

For notational convenience, for any α, $0 < \alpha < 1$, let Φ and $z_{\alpha/2}$ denote, respectively, the cumulative distribution function and $\alpha/2$ quantile of a standard normal distribution. Also, for any $B \in \mathbb{R}$, we define the function $G(\alpha, B)$ to be the tail probabilities of a normal distribution shifted by B, i.e.

$$G(\alpha, B) = 1 - \Phi(z_{\alpha/2} - B) + \Phi(-z_{\alpha/2} - B). \tag{1}$$

We use $\chi^2_\alpha(d)$ to denote the $1 - \alpha$ quantile of the Chi-squared distribution with d degrees of freedom.

2.2. Model and definitions

Suppose we have n individuals where for each individual $i = 1, \ldots, n$, we measure the outcome Y_i, the included endogenous variable D_i, p_z candidate instruments $Z'_{i.}$, and p_x exogenous covariates $X'_{i.}$ in an i.i.d. fashion. We denote $W'_{i.}$ to be concatenated vector of $Z'_{i.}$ and $X'_{i.}$ with dimension $p = p_z + p_x$. The columns of the matrix W are indexed by two sets, the set $\mathcal{I} = \{1, \ldots, p_z\}$, which consists of all the p_z candidate instruments, and the set $\mathcal{I}^C = \{p_z + 1, \ldots, p\}$, which consists of the p_x covariates. The variables (Y_i, D_i, Z_i, X_i) are governed by the following structural model.

$$Y_i = D_i\beta + X'_{i.}\phi + \delta_i, \qquad E(\delta_i \mid Z_{i.}, X_{i.}) = 0 \tag{2}$$

$$D_i = Z'_{i.}\gamma + X'_{i.}\psi + \epsilon_i, \qquad E(\epsilon_i \mid Z_{i.}, X_{i.}) = 0 \tag{3}$$

where β, ϕ, γ, and ψ are unknown parameters in the model and without loss of generality, we assume the variables are centered to mean zero.[4] Let the population covariance matrix of (δ_i, ϵ_i) be Σ, with $\Sigma_{11} = \text{Var}(\delta_i \mid Z_{i.}, X_{i.})$, $\Sigma_{22} = \text{Var}(\epsilon_i \mid Z_{i.}, X_{i.})$, and $\Sigma_{12} = \Sigma_{21} = \text{Cov}(\delta_i, \epsilon_i \mid Z_{i.}, X_{i.})$. Let the second order moments of $W_{i.}$ be $\Lambda = E\left(W_{i.}W'_{i.}\right)$ and let $\Lambda_{\mathcal{I}|\mathcal{I}^c}$ denote the adjusted covariance of variables belonging to the index set $\mathcal{I}$. Let ω represent all parameters $\omega = (\beta, \pi, \phi, \gamma, \psi, \Sigma)$ and define the parameter space

$$\Omega = \{\omega = (\beta, \pi, \phi, \gamma, \psi, \Sigma) : \beta \in \mathbb{R}, \pi, \gamma \in \mathbb{R}^{p_z}, \phi, \psi \in \mathbb{R}^{p_x}, \Sigma \in \mathbb{R}^{2\times 2}, \Sigma \succ 0\}. \tag{4}$$

Finally, we denote $s_{x2} = \|\phi\|_0, s_{z1} = \|\gamma\|_0, s_{x1} = \|\psi\|_0$ and $s = \max\{s_{x2}, s_{z1}, s_{x1}\}$.

We also define relevant and irrelevant instruments. This is, in many ways, equivalent to the notion that the instruments $Z_{i.}$ are associated with the endogenous variable D_i, except we use the support of a vector to define instruments' association to the endogenous variable; see Breusch et al. (1999); Hall and Peixe (2003), and Cheng and Liao (2015) for some examples in the literature of defining relevant and irrelevant instruments based on the support of a parameter.

[4] Mean-centering is equivalent to adding a constant 1 term (i.e. intercept term) in $X'_{i.}$; see Section 1.4 of Davidson and MacKinnon (1993) for details.

Definition 1. Suppose we have p_z instruments along with model (3). We say that instrument $j = 1, \ldots, p_z$ is relevant if $\gamma_j \neq 0$ and irrelevant if $\gamma_j = 0$. Let $\mathcal{S} \subseteq \mathcal{I} = \{1, 2, \ldots, p_z\}$ denote the set of relevant IVs.

Finally, for $\mathcal{S}$, the set of relevant IVs, we define the concentration parameter, a common measure of instrument strength,

$$C(\mathcal{S}) = \frac{\gamma_{\mathcal{S}}' \Lambda_{\mathcal{S}|\mathcal{S}^C} \gamma_{\mathcal{S}}}{|\mathcal{S}|\, \Sigma_{22}}. \tag{5}$$

If all the instruments were relevant, then $\mathcal{S} = \mathcal{I}$ and Eq. (5) is the usual definition of concentration parameter in Bound et al. (1995); Mariano (1973); Staiger and Stock (1997) and Stock and Wright (2000) using population quantities, i.e. $\Lambda_{\mathcal{S}|\mathcal{S}^C}$. In particular, $C(\mathcal{S})$ corresponds exactly to the quantity $\lambda'\lambda/K_2$ on page 561 of Staiger and Stock (1997) when $n = 1$ and $K_1 = 0$. Without using population quantities, the function $nC(\mathcal{S})$ roughly corresponds to the usual concentration parameter using the sample estimator of $\Lambda_{\mathcal{S}|\mathcal{S}^C}$. However, if only a subset of all instruments are relevant so that $\mathcal{S} \subset \mathcal{I}$, then the concentration parameter in Eq. (5) represents the strength of instruments for that subset $\mathcal{S}$, adjusted for the exogenous variables in its complement $\mathcal{S}^C$. Regardless, like the usual concentration parameter, a high value of $C(\mathcal{S})$ represents strong instruments in the set $\mathcal{S}$ while a low value of $C(\mathcal{S})$ represents weak instruments.

2.3. The DWH test

Consider the following hypotheses for detecting endogeneity in models (2) and (3),

$$H_0 : \Sigma_{12} = 0 \quad \text{versus} \quad H_1 : \Sigma_{12} \neq 0. \tag{6}$$

The DWH test tests the hypothesis of endogeneity in Eq. (6) by comparing two consistent estimators of β under the null hypothesis H_0 (i.e. no endogeneity) with different efficiencies. Formally, the DWH test statistic, denoted as Q_{DWH}, is the quadratic difference between the OLS estimator of β, $\widehat{\beta}_{\text{OLS}} = (D'P_{X^\perp}D)^{-1}D'P_{X^\perp}Y$, and the TSLS estimator of β, $\widehat{\beta}_{\text{TSLS}} = (D'(P_W - P_X)D)^{-1}D'(P_W - P_X)Y$.

$$Q_{\text{DWH}} = \frac{(\widehat{\beta}_{\text{TSLS}} - \widehat{\beta}_{\text{OLS}})^2}{\widehat{\text{Var}}(\widehat{\beta}_{\text{TSLS}}) - \widehat{\text{Var}}(\widehat{\beta}_{\text{OLS}})}. \tag{7}$$

The terms $\widehat{\text{Var}}(\widehat{\beta}_{\text{OLS}})$ and $\widehat{\text{Var}}(\widehat{\beta}_{\text{TSLS}})$ are standard error estimates of the OLS and TSLS estimators, respectively, and have the following forms

$$\widehat{\text{Var}}(\widehat{\beta}_{\text{OLS}}) = \left(D'P_{X^\perp}D\right)^{-1}\widehat{\Sigma}_{11}, \quad \widehat{\text{Var}}(\widehat{\beta}_{\text{TSLS}}) = \left(D'(P_W - P_X)D\right)^{-1}\widehat{\Sigma}_{11}. \tag{8}$$

The $\widehat{\Sigma}_{11}$ in Eq. (8) is the estimate of Σ_{11} and can either be based on the OLS estimate, i.e. $\widehat{\Sigma}_{11} = \|Y - D\widehat{\beta}_{\text{OLS}} - X\widehat{\phi}_{\text{OLS}}\|_2^2/n$, or the TSLS estimate, i.e. $\widehat{\Sigma}_{11} = \|Y - D\widehat{\beta}_{\text{TSLS}} - X\widehat{\phi}_{\text{TSLS}}\|_2^2/n$.[5] Under H_0, both OLS and TSLS estimators of the variance Σ_{11} are consistent. Also, under H_0, both OLS and TSLS estimators are consistent estimators of β, but the OLS estimator is more efficient than the TSLS estimator.

The asymptotic null distribution of the DWH test in Eq. (7) is a Chi-squared distribution with one degree of freedom (the DWH test has an exact Chi-squared null distribution with one degree of freedom if Σ_{11} is known). Hence for any $0 < \alpha < 1$, an asymptotically (or exactly if Σ_{11} is known) level α test is given by

$$\text{Reject } H_0 \text{ if } \quad Q_{\text{DWH}} \geq \chi_\alpha^2(1).$$

Also, under the local alternative hypothesis,

$$H_0 : \Sigma_{12} = 0 \quad \text{versus} \quad H_2 : \Sigma_{12} = \frac{\Delta_1}{\sqrt{n}} \tag{9}$$

for some constant $\Delta_1 \neq 0$, the asymptotic power of the DWH test is

$$\lim_{n\to\infty} \mathbf{P}(Q_{\text{DWH}} \geq \chi_\alpha^2(1)) = G\left(\alpha, \frac{\Delta_1\sqrt{C(\mathcal{I})}}{\sqrt{\left(C(\mathcal{I}) + \frac{1}{p_z}\right)\Sigma_{11}\Sigma_{22}}}\right), \tag{10}$$

where $G(\alpha, \cdot)$ is defined in Eq. (1); see Theorem 3 of the Supplementary Materials for a proof of Eq. (10). For textbook discussions on the DWH test, see Section 7.9 of Davidson and MacKinnon (1993) and Section 6.3.1 of Wooldridge (2010).

[5] The OLS and TSLS estimates of ϕ can be obtained as follows: $\widehat{\phi}_{\text{OLS}} = (X'P_{D^\perp}X)^{-1}X'P_{D^\perp}Y$ and $\widehat{\phi}_{\text{TSLS}} = (X'P_{\hat{D}^\perp}X)^{-1}X'P_{\hat{D}^\perp}Y$ where $\hat{D} = P_W D$.

3. The DWH test with many covariates

We now consider the behavior of the DWH test in the presence of many covariates and/or instruments. Formally, suppose the number of covariates and instruments are growing with sample size n, that is, $p_x = p_x(n)$, $p_z = p_z(n)$ and $\lim_{n\to\infty} \min\{p_z, p_x\} = \infty$, so that $p = p_x + p_z$ and $n - p$ are increasing with respect to n. For this section only, we focus on the case where $p < n$ since the DWH test with OLS and TSLS estimators cannot be implemented when the sample size is smaller than the dimension of the model parameters; later sections, specifically Section 4, will consider endogeneity testing including both $p < n$ and $p \geq n$ settings. We assume a known Σ_{11} for a cleaner technical exposition and to highlight the deficiencies of the DWH test that are not specific to estimating Σ_{11}, but specific to the form of the DWH test, the quadratic differencing of estimators in Eq. (7). However, the known Σ_{11} assumption can be replaced by a consistent estimate of Σ_{11}. Theorem 1 characterizes the asymptotic behavior of the DWH test under this setting.

Theorem 1. *Suppose we have models (2) and (3) where Σ_{11} is known, $W_{i\cdot}$ is a zero-mean multivariate Gaussian, the errors δ_i and ϵ_i are independent of $W_{i\cdot}$ and they are assumed to be bivariate normal. If $\sqrt{C(\mathcal{I})} \gg \sqrt{\log(n-p_x)/(n-p_x)p_z}$, for each α, $0 < \alpha < 1$, the asymptotic Type I error of the DWH test under H_0 is controlled at α, that is,*

$$\limsup_{n\to\infty} \mathbf{P}_\omega\left(|Q_{\mathrm{DWH}}| \geq z_{\alpha/2}\right) = \alpha \quad \text{for any } \omega \text{ with corresponding } \Sigma_{12} = 0.$$

Furthermore, for any ω with $\Sigma_{12} = \Delta_1/\sqrt{n}$, the asymptotic power of Q_{DWH} satisfies

$$\lim_{n\to\infty} \left| \mathbf{P}_\omega\left(Q_{\mathrm{DWH}} \geq \chi^2_\alpha(1)\right) - G\left(\alpha, \frac{C(\mathcal{I})\Delta_1\sqrt{1-\frac{p}{n}}}{\sqrt{\left(C(\mathcal{I}) + \frac{1}{n-p_x}\right)\left(C(\mathcal{I}) + \frac{1}{p_z}\right)\Sigma_{11}\Sigma_{22}}}\right)\right| = 0. \tag{11}$$

Note that the convergence in Theorem 1 is pointwise convergence instead of uniform convergence. Theorem 1 states that the Type I error of the DWH test is actually controlled at the desired level α if one were to naively use it in the presence of many covariates and/or instruments and Σ_{11} is known a priori. However, the power of the DWH test under the local alternative H_2 in Eq. (11) behaves differently in high dimensions than in low dimensions, as specified in Eq. (10). For example, if covariates and/or instruments are growing at $p/n \to 0$, Eq. (11) reduces to the usual power of the DWH test under low dimensional settings in Eq. (10). On the other hand, if covariates and/or instruments are growing at $p/n \to 1$, then the usual DWH test essentially has no power against any local alternative in H_2 since $G(\alpha, \cdot)$ in Eq. (11) equals α for any value of Δ_1.

This phenomenon suggests that in the "middle ground" where $p/n \to c$, $0 < c < 1$, the usual DWH test will likely suffer in terms of power. As a concrete example, if $p_x = n/2$ and $p_z = n/3$ so that $p/n = 5/6$, then $G(\alpha, \cdot)$ in Eq. (11) reduces to

$$G\left(\alpha, \frac{C(\mathcal{I})\Delta_1}{\sqrt{2\left(C(\mathcal{I}) + \frac{2}{n}\right)\left(C(\mathcal{I}) + \frac{1}{p_z}\right)\Sigma_{11}\Sigma_{22}}}\right) \approx G\left(\alpha, \frac{1}{\sqrt{6}} \cdot \frac{\sqrt{C(\mathcal{I})}\Delta_1}{\sqrt{\left(C(\mathcal{I}) + \frac{1}{p_z}\right)\Sigma_{11}\Sigma_{22}}}\right)$$

where the approximation sign is for n sufficiently large so that $C(\mathcal{I}) + 2/n \approx C(\mathcal{I})$. In this setting, the power of the DWH test is smaller than the power of the DWH test in Eq. (10) for the low dimensional setting. Section 5 also shows this phenomenon numerically.

Also, Theorem 1 provides some important guidelines for empiricists using the DWH test. First, Theorem 1 suggests that with modern cross-sectional data where the number of covariates may be very large, the DWH test should not be used to test endogeneity. Not only is the DWH test potentially incapable of detecting the presence of endogeneity under this scenario, but also an empiricist may be misled into a non-IV type of analysis, say the OLS or the Lasso, based on the result of the DWH test (Wooldridge, 2010). If the empiricist used a more powerful endogeneity test under this setting, he or she would have correctly concluded that there is endogeneity and used an IV analysis. Second, as discussed in Section 1, if empirical works add many covariates to make an IV more plausibly valid, one pays a price in terms of the power of the specification test; consequently, additional samples may be needed to get the desired level of power for detecting endogeneity.

Finally, we make two remarks about the regularity conditions in Theorem 1. First, Theorem 1 controls the growth of the concentration parameter $C(\mathcal{I})$ to be faster than $\log(n-p_x)/(n-p_x)p_z$. This growth condition is satisfied under the many instrument asymptotics of Bekker (1994) and the many weak instrument asymptotics of Chao and Swanson (2005) where $C(\mathcal{I})$ converges to a constant as $p_z/n \to c$ for some constant c. The weak instrument asymptotics of Staiger and Stock (1997) are not directly applicable to our growth condition on $C(\mathcal{I})$ because its asymptotics keeps p_z and p_x fixed. Second, we can replace the condition that $W_{i\cdot}$ is a zero-mean multivariate Gaussian in Theorem 1 by another condition used in high dimensional IV regression, for instance page 486 of Chernozhukov et al. (2015) where (i) the vector of instruments $Z_{i\cdot}$ is a linear model of $X_{i\cdot}$, i.e. $Z_{i\cdot}' = X_{i\cdot}'B + \bar{Z}_{i\cdot}'$, (ii) $\bar{Z}_{i\cdot}$ is independent of $X_{i\cdot}$, and (iii) $\bar{Z}_{i\cdot}$ is a multivariate normal distribution and the results in Theorem 1 will hold.

4. An improved endogeneity test

Given that the DWH test for endogeneity may have low power in high dimensional settings, we present a simple and improved endogeneity test that has better power to detect endogeneity. In particular, our endogeneity test takes any popular estimator that is "well-behaved" for estimating reduced-form parameters (see Definition 2 for details) and applies a simple hard thresholding procedure to choose the most relevant instruments. We also stress that our endogeneity test is the first test capable of testing endogeneity if the number of parameters exceeds the sample size.

4.1. Well-behaved estimators

Consider the following reduced-form models

$$Y_i = Z_{i.}'\Gamma + X_{i.}'\Psi + \xi_i, \tag{12}$$
$$D_i = Z_{i.}'\gamma + X_{i.}'\psi + \epsilon_i. \tag{13}$$

The terms $\Gamma = \beta\gamma$ and $\Psi = \phi + \beta\psi$ are the parameters for the reduced-form model (12) and $\xi_i = \beta\epsilon_i + \delta_i$ is the reduced-form error term. The errors in the reduced-form models have the property that $\mathbf{E}(\xi_i|Z_{i.}, X_{i.}) = 0$ and $\mathbf{E}(\epsilon_i|Z_{i.}, X_{i.}) = 0$. Also, the covariance matrix of these error terms, denoted as Θ, has the following forms: $\Theta_{11} = \mathrm{Var}(\xi_i|Z_{i.}, X_{i.}) = \Sigma_{11} + 2\beta\Sigma_{12} + \beta^2\Sigma_{22}$, $\Theta_{22} = \mathrm{Var}(\epsilon_i|Z_{i.}, X_{i.})$, and $\Theta_{12} = \mathrm{Cov}(\xi_i, \epsilon_i|Z_{i.}, X_{i.}) = \Sigma_{12} + \beta\Sigma_{22}$.

As mentioned before, our improved endogeneity test does not require a specific estimator for the reduced-form parameters. Rather, any estimator that is well-behaved, as defined below, will be sufficient.

Definition 2. Consider estimators $(\widehat{\gamma}, \widehat{\Gamma}, \widehat{\Theta}_{11}, \widehat{\Theta}_{22}, \widehat{\Theta}_{12})$ of the reduced-form parameters, $(\gamma, \Gamma, \Theta_{11}, \Theta_{22}, \Theta_{12})$ respectively, in Eqs. (12) and (13). The estimators $(\widehat{\gamma}, \widehat{\Gamma}, \widehat{\Theta}_{11}, \widehat{\Theta}_{22}, \widehat{\Theta}_{12})$ are well-behaved estimators if they satisfy the two criteria below.

(W1) The reduced-form estimators of the coefficients $\widehat{\gamma}$ and $\widehat{\Gamma}$ satisfy

$$\sqrt{n}\|(\widehat{\gamma} - \gamma) - \frac{1}{n}\widehat{V}'\epsilon\|_\infty = O_p\left(\frac{s\log p}{\sqrt{n}}\right),\ \sqrt{n}\|\left(\widehat{\Gamma} - \Gamma\right) - \frac{1}{n}\widehat{V}'\xi\|_\infty = O_p\left(\frac{s\log p}{\sqrt{n}}\right). \tag{14}$$

for some matrix $\widehat{V} = (\widehat{V}_{\cdot 1}, \ldots, \widehat{V}_{\cdot p_z})$ which is only a function of W and satisfies

$$\liminf_{n\to\infty}\mathbf{P}\left(c \le \min_{1\le j\le p_z}\frac{\|\widehat{V}_{\cdot j}\|_2}{\sqrt{n}} \le \max_{1\le j\le p_z}\frac{\|\widehat{V}_{\cdot j}\|_2}{\sqrt{n}} \le C,\ c\|\gamma\|_2 \le \frac{1}{\sqrt{n}}\|\sum_{j\in\mathcal{S}}\gamma_j\widehat{V}_{\cdot j}\|_2\right) = 1 \tag{15}$$

for some constants $c > 0$ and $C > 0$.

(W2) The reduced-form estimators of the error variances, $\widehat{\Theta}_{11}$, $\widehat{\Theta}_{22}$, and $\widehat{\Theta}_{12}$ satisfy

$$\sqrt{n}\max\left\{\left|\widehat{\Theta}_{11} - \frac{1}{n}\xi'\xi\right|, \left|\widehat{\Theta}_{12} - \frac{1}{n}\epsilon'\xi\right|, \left|\widehat{\Theta}_{22} - \frac{1}{n}\epsilon'\epsilon\right|\right\} = O_p\left(\frac{s\log p}{\sqrt{n}}\right). \tag{16}$$

There are many estimators of the reduced-form parameters in the literature that are well-behaved. Some examples of well-behaved estimators are listed below.

1. (OLS): In settings where p is fixed or p is growing with n at a rate $p/n \to 0$, the OLS estimators of the reduced-form parameters, i.e.

$$(\widehat{\Gamma}, \widehat{\Psi})' = (W'W)^{-1}W'Y \quad , (\widehat{\gamma}, \widehat{\psi})' = (W'W)^{-1}W'D,$$
$$\widehat{\Theta}_{11} = \frac{\left\|Y - Z\widehat{\Gamma} - X\widehat{\Psi}\right\|_2^2}{n-1} \quad , \widehat{\Theta}_{22} = \frac{\left\|D - Z\widehat{\gamma} - X\widehat{\psi}\right\|_2^2}{n-1}$$
$$\widehat{\Theta}_{12} = \frac{\left(Y - Z\widehat{\Gamma} - X\widehat{\Psi}\right)'\left(D - Z\widehat{\gamma} - X\widehat{\psi}\right)}{n-1}$$

trivially satisfy conditions for well-behaved estimators. Specifically, let $\widehat{V}' = (\frac{1}{n}W'W)^{-1}_{\mathcal{I}\cdot}W$. Then Eq. (14) holds because $(\widehat{\gamma} - \gamma) - \widehat{V}'\epsilon = 0$ and $\left(\widehat{\Gamma} - \Gamma\right) - \widehat{V}'\xi = 0$. Also, Eq. (15) holds because, in probability, $n^{-1/2}\|\widehat{V}_{\cdot j}\|_2 \to \Lambda_{jj}^{-1}$ and $n^{-1}\widehat{V}'\widehat{V} \to \Lambda_{\mathcal{I}\mathcal{I}}^{-1}$, thus satisfying (W1). Also, (W2) holds because $\|\widehat{\Gamma} - \Gamma\|_2^2 + \|\widehat{\Psi} - \Psi\|_2^2 = O_p\left(n^{-1}\right)$ and $\|\widehat{\gamma} - \gamma\|_2^2 + \|\widehat{\psi} - \psi\|_2^2 = O_p\left(n^{-1}\right)$, which implies Eq. (16) is going to zero at $n^{-1/2}$ rate.
2. (Debiased Lasso Estimators) In high dimensional settings where p is growing with n and often exceeds n, one of the most popular estimators for regression model parameters is the Lasso (Tibshirani, 1996). Unfortunately, the Lasso estimator and many penalized estimators do not satisfy the definition of a well-behaved estimator, specifically (W1), because penalized estimators are typically biased. Fortunately, recent works by Javanmard and Montanari (2014); van de Geer et al. (2014); Zhang and Zhang (2014) and Cai and Guo (2016) remedied this bias problem by doing a bias correction on the original penalized estimates.

More concretely, suppose we use the square root Lasso estimator by Belloni et al. (2011b),

$$\{\widetilde{\Gamma}, \widetilde{\Psi}\} = \underset{\Gamma\in\mathbb{R}^{p_Z},\Psi\in\mathbb{R}^{p_X}}{\operatorname{argmin}} \frac{\|Y - Z\Gamma - X\Psi\|_2}{\sqrt{n}} + \frac{\lambda_0}{\sqrt{n}}\left(\sum_{j=1}^{p_z}\|Z_j\|_2|\Gamma_j| + \sum_{j=1}^{p_x}\|X_j\|_2|\Psi_j|\right) \tag{17}$$

for the reduced-form model in Eq. (12) and

$$\{\widetilde{\gamma}, \widetilde{\psi}\} = \underset{\Gamma\in\mathbb{R}^{p_Z},\Psi\in\mathbb{R}^{p_X}}{\operatorname{argmin}} \frac{\|D - Z\gamma - X\psi\|_2}{\sqrt{n}} + \frac{\lambda_0}{\sqrt{n}}\left(\sum_{j=1}^{p_z}\|Z_j\|_2|\gamma_j| + \sum_{j=1}^{p_x}\|X_j\|_2|\psi_j|\right) \tag{18}$$

for the reduced-form model in Eq. (13). The term λ_0 in both estimation problems (17) and (18) represents the penalty term in the square root Lasso estimator and typically, the penalty is set at $\lambda_0 = \sqrt{a_0 \log p/n}$ for some constant a_0 slightly greater than 2, say 2.01 or 2.05. To transform the above penalized estimators in Eqs. (17) and (18) into well-behaved estimators, we follow Javanmard and Montanari (2014) to debias the penalized estimators. Specifically, we solve p_z optimization problems where the solution to each p_z optimization problem, denoted as $\widehat{u}^{[j]} \in \mathbb{R}^p$, $j = 1, \ldots, p_z$, is

$$\widehat{u}^{[j]} = \underset{u\in\mathbb{R}^p}{\operatorname{argmin}} \frac{1}{n}\|Wu\|_2^2 \quad \text{s.t.} \quad \|\frac{1}{n}W'Wu - I_j\|_\infty \le \lambda_n.$$

Typically, the tuning parameter λ_n is chosen to be $12M_1^2\sqrt{\log p/n}$ where M_1 is defined as the largest eigenvalue of Λ. Define $\widehat{V}_{\cdot j} = W\widehat{u}^{[j]}$ and $\widehat{V} = (\widehat{V}_{\cdot 1}, \ldots, \widehat{V}_{\cdot p_z})$. Then, we can transform the penalized estimators in (17) and (18) into debiased, well-behaved estimators, $\widehat{\Gamma}$ and $\widehat{\gamma}$,

$$\widehat{\Gamma} = \widetilde{\Gamma} + \frac{1}{n}\widehat{V}'\left(Y - Z\widetilde{\Gamma} - X\widetilde{\Psi}\right), \quad \widehat{\gamma} = \widetilde{\gamma} + \frac{1}{n}\widehat{V}'\left(D - Z\widetilde{\gamma} - X\widetilde{\psi}\right). \tag{19}$$

Guo et al. (2016) showed that $\widehat{\Gamma}$ and $\widehat{\gamma}$ satisfy (W1). As for the error variances, following Belloni et al. (2011b), Sun and Zhang (2012) and Ren et al. (2013), we estimate the covariance terms $\Theta_{11}, \Theta_{22}, \Theta_{12}$ by

$$\begin{aligned}\widehat{\Theta}_{11} &= \frac{\left\|Y - Z\widetilde{\Gamma} - X\widetilde{\Psi}\right\|_2^2}{n}, \quad \widehat{\Theta}_{22} = \frac{\left\|D - Z\widetilde{\gamma} - X\widetilde{\psi}\right\|_2^2}{n} \\ \widehat{\Theta}_{12} &= \frac{\left(Y - Z\widetilde{\Gamma} - X\widetilde{\Psi}\right)'\left(D - Z\widetilde{\gamma} - X\widetilde{\psi}\right)}{n}.\end{aligned} \tag{20}$$

Lemma 3 of Guo et al. (2016) showed that the above estimators of $\widehat{\Theta}_{11}, \widehat{\Theta}_{22}$ and $\widehat{\Theta}_{12}$ in Eq. (20) satisfy (W2). In summary, the debiased Lasso estimators in Eq. (19) and the variance estimators in Eq. (20) are well-behaved estimators.

3. (One-Step and Orthogonal Estimating Equations Estimators) Recently, Chernozhukov et al. (2015) proposed the one-step estimator of the reduced-form coefficients, i.e.

$$\widehat{\Gamma} = \widetilde{\Gamma} + \frac{1}{n}\widetilde{\Lambda^{-1}}_{\mathcal{I},\cdot}W^\intercal\left(Y - Z\widetilde{\Gamma} - X\widetilde{\Psi}\right), \ \widehat{\gamma} = \widetilde{\gamma} + \frac{1}{n}\widetilde{\Lambda^{-1}}_{\mathcal{I},\cdot}W^\intercal\left(D - Z\widetilde{\gamma} - X\widetilde{\psi}\right).$$

where $\widetilde{\Gamma}, \widetilde{\gamma}$, and $\widetilde{\Lambda^{-1}}$ are initial estimators of Γ, γ and Λ^{-1}, respectively. The initial estimators must satisfy conditions (18) and (20) of Chernozhukov et al. (2015) and many popular estimators like the Lasso or the square root Lasso satisfy these two conditions. Then, the arguments in Theorem 2.1 of van de Geer et al. (2014) showed that the one-step estimator of Chernozhukov et al. (2015) satisfies (W1). Relatedly, Chernozhukov et al. (2015) proposed estimators for the reduced-form coefficients based on orthogonal estimating equations and in Proposition 4 of Chernozhukov et al. (2015), the authors showed that the orthogonal estimating equations estimator is asymptotically equivalent to their one-step estimator.

For variance estimation, one can use the variance estimator in Belloni et al. (2011b), which reduces to the estimators in Eq. (20) and thus, satisfies (W2).

In short, the first part of our endogeneity test requires any estimator that is well-behaved and, as illustrated above, many estimators, such as the OLS in low dimensions and bias corrected penalized estimators in high dimensions, satisfy the criteria for a well-behaved estimator.

4.2. Estimating relevant instruments via hard thresholding

Once we have well-behaved estimators $(\widehat{\gamma}, \widehat{\Gamma}, \widehat{\Theta}_{11}, \widehat{\Theta}_{22}, \widehat{\Theta}_{12})$ satisfying Definition 2, the next step in our endogeneity test is finding IVs that are relevant, that is the set $\mathcal{S}$ in Definition 1 comprised of $\gamma_j \neq 0$. We do this by hard thresholding the estimate $\widehat{\gamma}$ by the dimension and the noise of $\widehat{\gamma}$.

$$\widehat{\mathcal{S}} = \left\{j : |\widehat{\gamma}_j| \ge \frac{\sqrt{\widehat{\Theta}_{22}}\|\widehat{V}_{\cdot j}\|_2}{\sqrt{n}}\sqrt{\frac{a_0 \log\max\{p_z, n\}}{n}}\right\}. \tag{21}$$

The set $\widehat{\mathcal{S}}$ is an estimate of $\mathcal{S}$ and a_0 is some constant greater than 2; from our experience and like many Lasso problems, $a_0 = 2.01$ or $a_0 = 2.05$ works well in practice. The threshold in (21) is based on the noise level of $\widehat{\gamma}_j$ in Eq. (14) (represented by the term $n^{-1}\sqrt{\widehat{\Theta}_{22}}\|\widehat{V}_{\cdot j}\|_2$), adjusted by the dimensionality of the instrument size (represented by the term $\sqrt{a_0 \log \max\{p_z, n\}}$).

Using the estimated set $\widehat{\mathcal{S}}$ of relevant IVs leads to the estimates of Σ_{12}, Σ_{11}, and β,

$$\widehat{\Sigma}_{12} = \widehat{\Theta}_{12} - \widehat{\beta}\widehat{\Theta}_{22}, \quad \widehat{\Sigma}_{11} = \widehat{\Theta}_{11} + \widehat{\beta}^2\widehat{\Theta}_{22} - 2\widehat{\beta}\widehat{\Theta}_{12}, \quad \widehat{\beta} = \frac{\sum_{j\in\widehat{\mathcal{S}}}\widehat{\gamma}_j\widehat{\Gamma}_j}{\sum_{j\in\widehat{\mathcal{S}}}\widehat{\gamma}_j^2}. \tag{22}$$

Eq. (22) provides us with the ingredients to construct our new test for endogeneity, which we denote as Q

$$Q = \frac{\sqrt{n}\widehat{\Sigma}_{12}}{\sqrt{\widehat{\mathrm{Var}}(\widehat{\Sigma}_{12})}} \quad \text{and} \quad \widehat{\mathrm{Var}}(\widehat{\Sigma}_{12}) = \widehat{\Theta}_{22}^2\widehat{\mathrm{Var}}_1 + \widehat{\mathrm{Var}}_2 \tag{23}$$

where $\widehat{\mathrm{Var}}_1 = \widehat{\Sigma}_{11}\left\|\sum_{j\in\widehat{\mathcal{S}}}\widehat{\gamma}_j\widehat{V}_{\cdot j}/\sqrt{n}\right\|_2^2/\left(\sum_{j\in\widehat{\mathcal{S}}}\widehat{\gamma}_j^2\right)^2$ and $\widehat{\mathrm{Var}}_2 = \widehat{\Theta}_{11}\widehat{\Theta}_{22}+\widehat{\Theta}_{12}^2+2\widehat{\beta}^2\widehat{\Theta}_{22}^2-4\widehat{\beta}\widehat{\Theta}_{12}\widehat{\Theta}_{22}$. Here, Var_1 is the variance associated with estimating β and Var_2 is the variance associated with estimating Θ.

A major difference between the original DWH test in Eq. (7) and our endogeneity test in Eq. (23) is that our endogeneity test directly estimates and tests the endogeneity parameter Σ_{12} while the original DWH test implicitly tests for the endogeneity parameter by checking the quadratic distance between the OLS and TSLS estimators under the null hypothesis. More importantly, our endogeneity test efficiently uses the sparsity of the regression vectors while the DWH test does not incorporate such information. As shown in Section 4.3, our endogeneity test in this form where we make use of the sparsity information to estimate Σ_{12} will have superior power in high dimension compared to the DWH test.

4.3. Properties of the new endogeneity test

We study the properties of our new test in high dimensional settings where p is a function of n and is allowed to be larger than n; note that this is a generalization of the setting discussed in Section 3 where $p < n$ because the DWH test is not feasible when $p \geq n$. Theorem 1 showed that the DWH test, while it controls Type I error at the desired level, may have low power, especially when the ratio of p/n is close to 1. Theorem 2 shows that our new test Q remedies this deficiency of the DWH test by having proper Type I error control and exhibiting better power than the DWH test.

Theorem 2. *Suppose we have models (2) and (3) where the errors δ_i and ϵ_i are independent of $W_{i\cdot}$ and are assumed to be bivariate normal and we use a well-behaved estimator in our test statistic Q. If $\sqrt{C(\mathcal{S})} \gg s_{z1}\log p/\sqrt{n|\mathcal{V}|}$, and $\sqrt{s_{z1}}s\log p/\sqrt{n} \to 0$, then for any α, $0 < \alpha < 1$, the asymptotic Type I error of Q under H_0 is controlled at α, that is,*

$$\lim_{n\to\infty}\mathbf{P}_w\left(|Q| \geq z_{\alpha/2}\right) = \alpha, \quad \text{for any } \omega \text{ with corresponding } \Sigma_{12} = 0. \tag{24}$$

For any ω with $\Sigma_{12} = \Delta_1/\sqrt{n}$, the asymptotic power of Q is

$$\lim_{n\to\infty}\left|\mathbf{P}_\omega\left(|Q| \geq z_{\alpha/2}\right) - \mathbf{E}\left(G\left(\alpha, \frac{\Delta_1}{\sqrt{\Theta_{22}^2\mathrm{Var}_1 + \mathrm{Var}_2}}\right)\right)\right| = 0, \tag{25}$$

where $\mathrm{Var}_1 = \Sigma_{11}\left\|\sum_{j\in\mathcal{S}}\gamma_j\widehat{V}_{\cdot j}/\sqrt{n}\right\|_2^2/\left(\sum_{j\in\mathcal{S}}\gamma_j^2\right)^2$ *and* $\mathrm{Var}_2 = \Theta_{11}\Theta_{22} + \Theta_{12}^2 + 2\beta^2\Theta_{22}^2 - 4\beta\Theta_{12}\Theta_{22}$.

In contrast to Eq. (11) that described the power of the usual DWH test in high dimensions, the term $\sqrt{1-p/n}$ is absent in the power of our new endogeneity test Q in Eq. (25). Specifically, under the local alternative H_2, our power is only affected by Δ_1 while the power of the DWH test is affected by $\Delta_1\sqrt{1-p/n}$. Consequently, the power of our test Q does not suffer from the growing dimensionality of p. For example, in the extreme case when $p/n \to 1$ and $C(\mathcal{S})$ is a constant, the power of the usual DWH test will be α while the power of our test Q will always be greater than α. For further validation, Section 5 numerically illustrates the discrepancies between the power of the two tests. Finally, we stress that in the case $p > n$, our test still has proper size and non-trivial power while the DWH test is not feasible in this setting.

With respect to the regularity conditions in Theorem 2, like Theorem 1, Theorem 2 controls the growth of the concentration parameter $C(\mathcal{S})$ to be faster than $s_{z1}\log p/\sqrt{n|\mathcal{S}|}$, with a minor discrepancy in the growth rate due to the differences between the set of relevant IVs, $\mathcal{S}$, and the set of candidate IVs, $\mathcal{I}$. But, similar to Theorem 1, this growth condition is satisfied under the many instrument asymptotics of Bekker (1994) and the many weak instrument asymptotics of Chao and Swanson (2005). Also, note that unlike the negative result in Theorem 1, the "positive" result in Theorem 2 is more general in that we do not require W to be Gaussian and require Σ_{11} to be known a priori. Instead, we only need the conditions of well-behaved estimators to hold. Also, we follow other high-dimensional inference works Javanmard and Montanari (2014); van de Geer et al. (2014); Zhang and Zhang (2014) in assuming independence and normality assumptions on the error terms δ_i and ϵ_i, where such assumptions are made out of technicalities in establishing the distribution of test statistics in high dimensions. Finally, we remark that the expectation inside Eq. (25) is with respect to W and $\widehat{V}$ is a function of W.

4.4. An extension: endogeneity test in high dimensions with possibly invalid IVs

As discussed in Section 1, one of the motivations for having high dimensional covariates in empirical IV work is to avoid invalid instruments. While adding more covariates can potentially make instruments more plausibly valid, as demonstrated in Section 3, there is a price to pay with respect to the power of the DWH test. More importantly, even after conditioning on many covariates, some IVs may still be invalid and subsequent analysis, including the DWH test, assuming that all the IVs are valid after conditioning, can be seriously misleading. Inspired by these concerns, there has been a recent literature in estimation and inference of structural parameters in IV regression when invalid instruments are present (Guo et al., 2016; Kang et al., 2016; Kolesár et al., 2015). Our new endogeneity test Q can be extended to handle the case of invalid instruments through the voting method proposed in Guo et al. (2016). The methodological and theoretical details are presented in Section 3.3 of the Supplementary Materials. To summarize the results, the extension of Q to handle invalid instruments still controls the Type I error rate and has non-negligible power under high dimension with possibly invalid instruments.

5. Simulation and data example

5.1. Setup

We conduct a simulation study to investigate the performance of our new endogeneity test and the DWH test in high dimensional settings. Specifically, we generate data from models (2) and (3) in Section 2.2 with $n = 200$ or 300, $p_z = 100$ and $p_x = 150$. The vector $W_{i\cdot}$ is a multivariate normal with mean zero and covariance $\Lambda_{ij} = 0.5^{|i-j|}$ for $1 \le i, j \le p$. We set the parameters as follows $\beta = 1$, $\phi = (0.6, 0.7, 0.8, \ldots, 1.5, 0, 0, \ldots, 0) \in \mathbb{R}^{p_x}$ so that $s_{x1} = 10$, and $\psi = (1.1, 1.2, 1.3, \ldots, 2.0, 0, 0, \ldots, 0) \in \mathbb{R}^{p_x}$ so that $s_{x2} = 10$. The relevant instruments are $\mathcal{S} = \{1, \ldots, 7\}$. Variance of the error terms are set to $\mathrm{Var}(\delta_i) = \mathrm{Var}(\epsilon_i) = 1.5$.

The parameters we vary in the simulation study are: the endogeneity level via $\mathrm{Cov}(\delta_i, \epsilon_i)$, and IV strength via γ. For the endogeneity level, we set $\mathrm{Cov}(\delta_i, \epsilon_i) = 1.5\rho$, where ρ is varied and captures the level of endogeneity; a larger value of $|\rho|$ indicates a stronger correlation between the endogenous variable D_i and the error term δ_i. For IV strength, we set $\gamma_{\mathcal{S}} = K\,(1, 1, 1, 1, 1, 1, \rho_1)$ and $\gamma_{\mathcal{S}^c} = 0$, where K is varied as a function of the concentration parameter (see below) and ρ_1 is either 0 or 0.2. Specifically, the value K controls the global strength of instruments, with higher $|K|$ indicating strong instruments in a global sense. In contrast, the value ρ_1 controls the relative individual strength of instruments, specifically between the first six instruments in $\mathcal{S}$ and the seventh instrument. For example, $\rho_1 = 0.2$ implies that the seventh instrument's individual strength is only 20% of the first six instruments. Note that varying ρ_1 essentially stress-tests the thresholding step in our endogeneity test to numerically verify whether our testing procedure can handle relevant IVs with very small magnitudes of γ.

We specify K as follows. Suppose we have a set of simulation parameters $\mathcal{S}$, ρ_1, Λ and Σ_{22}. For each value of $100 \cdot C(\mathcal{S})$, we find the corresponding K that satisfies $100 \cdot C(\mathcal{S}) = 100 \cdot K^2 \|\Lambda^{1/2}_{\mathcal{S}|\mathcal{S}^c}(1, 1, 1, 1, 1, 1, \rho_1)\|_2^2/(7 \cdot 1.5)$. We vary $100 \cdot C(\mathcal{S})$ from 25 to 100, specifying K for each value of $100 \cdot C(\mathcal{S})$.

For each simulation setting, we repeat the data generation 1000 times. For each simulation setting, we compare the power of our testing procedure Q to the DWH test and the oracle DWH test where an oracle knows the support of the parameter vectors ϕ, ψ and γ. We set the desired α level for all three tests to be $\alpha = 0.05$.

5.2. Results

Table 1 and Fig. 1 consider the high dimensional setting with $n = 200, 300$, $p_x = 150$, and $p_z = 100$. Table 1 measures the Type I error rate across three methods; for $n = 200$, the regular DWH test was not used since both the OLS and TSLS estimators are infeasible in this regime. We see a few clear trends in Table 1. First, generally speaking, all three methods control their Type I error around the desired $\alpha = 0.05$. Our proposed test has a slight upward bias of Type I error in some high dimensional settings with weak IV, i.e. where the C value is around 25. But, the worst case upward bias is no more than 0.03 off from the target 0.05 and is within simulation error as C gets larger. Additionally, as Fig. 1 shows, the slight bias in Type I error in small C regimes is offset by substantial power gains compared to the regular DWH test. Second, as the instrument gets stronger, both individually via ρ_2 and overall via C, the Type I error control generally gets better across all three methods, which is not surprising given the literature on strong instruments.

Figs. 1 and 2 consider the power of our test Q, the regular DWH test, and the oracle DWH test in the high dimensional setting with $n = 200, 300$, $p_x = 150$, and $p_z = 100$. As predicted by Theorem 1, the regular DWH test suffers from low power, especially if the degree of endogeneity is around 0.25 where the gap between the regular DWH test and the oracle DWH test is the greatest across most simulation settings. In fact, even if the global strength of the IV increases, the DWH test still has low power. In contrast, as predicted from Theorem 2, our test Q can handle $n \approx p$ or $n < p$. It also has uniformly better power than the regular DWH test across all degrees of endogeneity and across all simulation settings in the plot. Our test also achieves near-oracle performance as the global instrument strength grows.

In summary, all the simulation results indicate that our endogeneity test controls Type I error and is a much better alternative to the regular DWH test in high dimensional settings, with near-optimal performance with respect to the oracle. Our test is also capable of handling the regime $n < p$. In the supplementary materials, we also conduct low dimensional simulations and show that all three tests, the oracle DWH test, the regular DWH test, and our proposed test behave identically with respect to power and Type I error control.

Table 1
Empirical Type I error when $p_x = 150$ and $p_z = 100$ after 1000 simulations. The value n represents the sample size and $\alpha = 0.05$. "Regular," "Ours," and "Oracle" represent the regular DWH test, the proposed test (Q), and the oracle DWH test, respectively. "Weak", and "Strong " represent the cases when $\rho_1 = 0.2$ and $\rho_1 = 0$, respectively. C represents the overall strength of the instruments, as measured by $100 \cdot C(\mathcal{S})$. NA indicates not applicable.

		Weak			Strong		
C	n	Regular	Ours	Oracle	Regular	Ours	Oracle
25	300	0.040	0.079	0.034	0.061	0.048	0.038
	200	NA	0.080	0.054	NA	0.075	0.054
50	300	0.049	0.046	0.032	0.043	0.065	0.048
	200	NA	0.072	0.055	NA	0.069	0.050
75	300	0.053	0.059	0.044	0.043	0.062	0.048
	200	NA	0.065	0.038	NA	0.063	0.048
100	300	0.067	0.055	0.048	0.050	0.064	0.044
	200	NA	0.057	0.045	NA	0.049	0.045

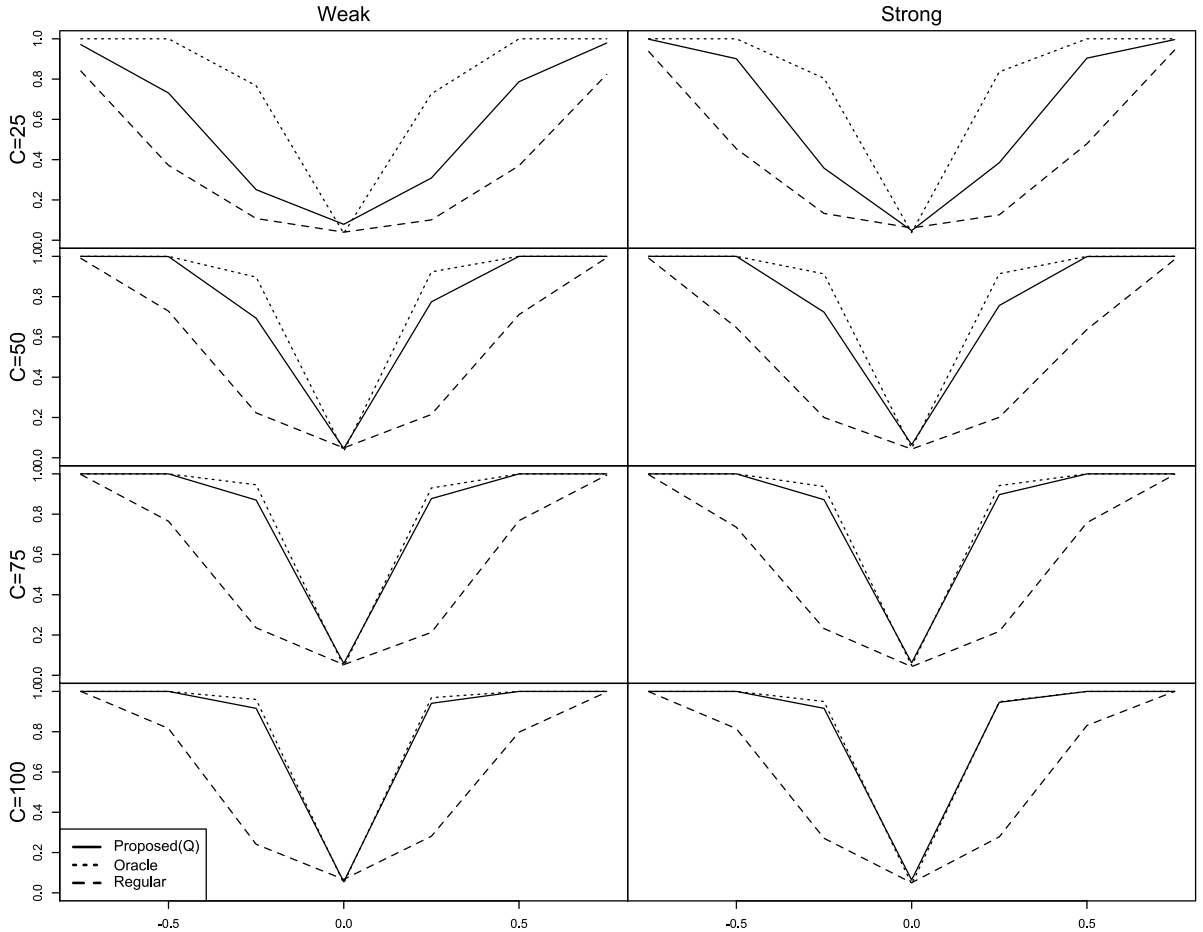

Fig. 1. Power of endogeneity tests when $n = 300$, $p_x = 150$ and $p_z = 100$. The x-axis represents the endogeneity ρ and the y-axis represents the empirical power over 1000 simulations. Each line represents a particular test's empirical power over various values of the endogeneity, where the solid line, the dashed line and the dotted line represent the proposed test (Q), the regular DWH test and the oracle DWH test, respectively. The columns represent the individual IV strengths, with column names "Weak" and "Strong" denoting the cases when $\rho_1 = 0.2$, and $\rho_1 = 0$, respectively. The rows represent the overalls strength of the instruments, as measured by $100 \cdot C(\mathcal{S})$.

5.3. Data example

To highlight the usefulness of the proposed test statistic Q, specifically its ability to run DWH test in dimensions where $n < p$, we re-analyze a high dimensional data analysis done in Belloni et al. (2012, 2014). Specifically, the outcome Y is the log of average Case–Shiller home price index and the endogenous variable D is the number of federal appellate court decisions that were against seizure of property via eminent domain. There are $n = 183$ individuals and $p_z = 147$ instruments which are derived from indicators that represent the random assignment of judges to different cases, characteristics of judges, and

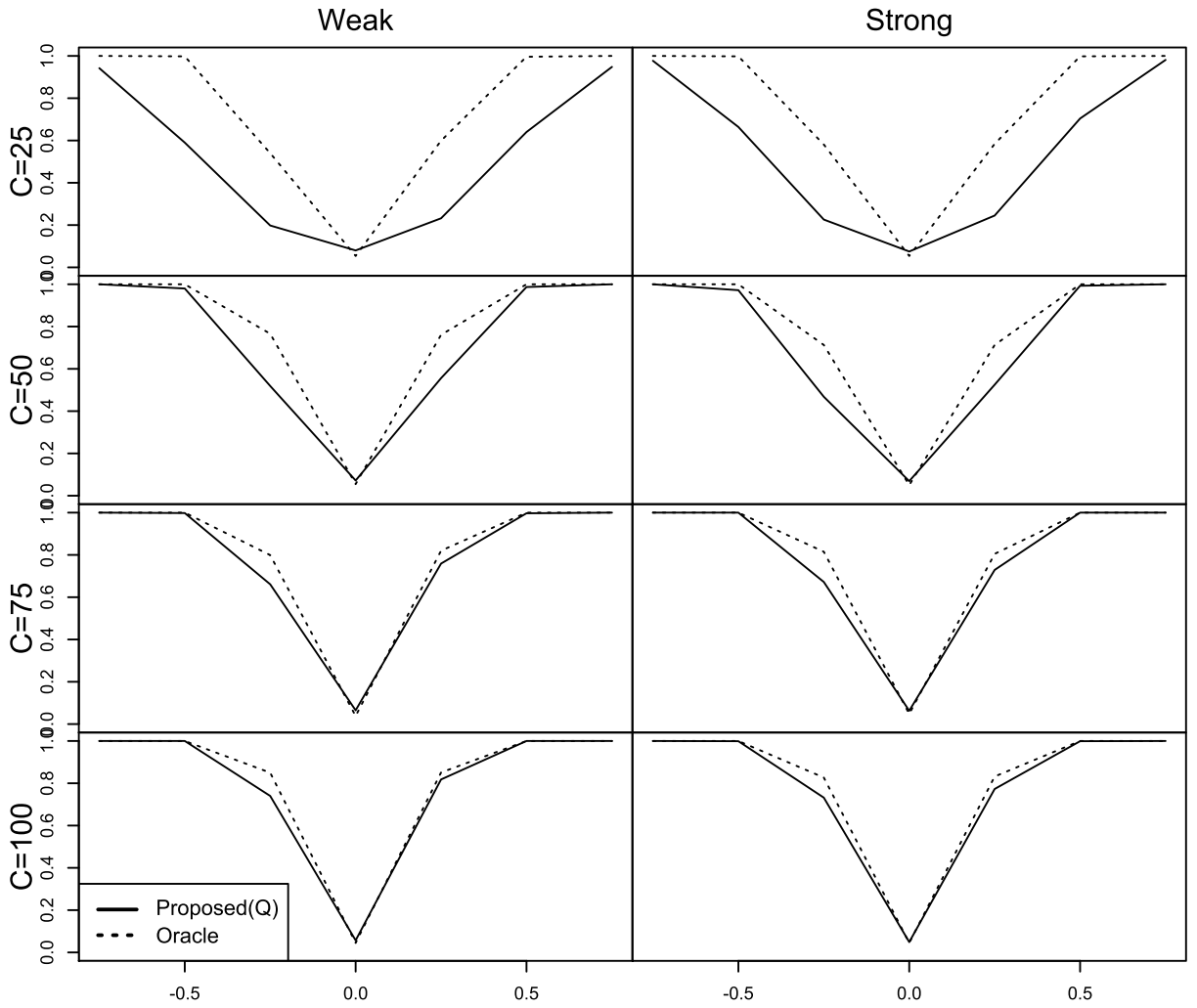

Fig. 2. Power of endogeneity tests when $n = 200$, $p_x = 150$ and $p_z = 100$. The x-axis represents the endogeneity ρ and the y-axis represents the empirical power over 1000 simulations. Each line represents a particular test's empirical power over various values of the endogeneity, where the solid line and the dotted line represent the proposed test (Q) and the oracle DWH test, respectively. The columns represent the individual IV strengths, with column names "Weak" and "Strong" denoting the cases when $\rho_1 = 0.2$, and $\rho_1 = 0$, respectively. The rows represent the overalls strength of the instruments, as measured by $100 \cdot C(\mathcal{S})$.

other interactions. Additionally, there are $p_x = 71$ exogenous variables that describe the type of cases, number of court decisions, circuit specific and time-specific effects; see Belloni et al. (2012) and Belloni et al. (2014) for more details about the instruments and the exogenous variables. We use the code provided in Belloni et al. (2012) to replicate the data set.

Because $n < p$, the DWH test or other tests for endogeneity cannot be used. Consequently, investigators are forced to remove covariates and/or instruments to run their usual specification test. For example, in our analysis, we drop the covariates and use the AER package (Kleiber and Zeileis, 2008), which is a popular R package to run IV analysis, to run the DWH test. The package reports back that the p-value for the DWH test is 0.683.

In contrast, our new test Q allows data where $n < p$. As such, we are not forced to remove covariates from the original analysis when we run our test Q on this data. Our test reports the p-value for the Q test is 0.21, meaning that there is not evidence for the number of federal appellate court decisions against seizure of property or eminent domain being endogenous. Unlike the DWH test, our test was able to accommodate these high dimensional covariates rather than dropping them from the analysis.

6. Conclusion and discussion

In this paper, we showed that the popular DWH test, while being able to control Type I error, can have low power in high dimensional settings. We propose a simple and improved endogeneity test to remedy the low power of the DWH test

by modifying popular reduced-form parameters with a thresholding step. We also show that this modification leads to drastically better power than the DWH test in high dimensional settings.

For empirical work, the results in the paper suggest that one should be cautious in interpreting high p-values produced by the DWH test in IV regression settings when many covariates and/or instruments are present. In particular, as shown in Section 3, in modern data settings with a potentially large number of covariates and/or instruments, the DWH test may declare that there is no endogeneity in the structural model, even if endogeneity is truly present. Our proposed test, which is a simple modification of the popular estimators for reduced-forms parameters, does not suffer from this problem, as it achieves near-oracle performance to detect endogeneity, and can even handle general settings when $n < p$ and invalid IVs are present.

Acknowledgments

The research of Hyunseung Kang was supported in part by NSF Grant DMS-1502437. The research of T. Tony Cai was supported in part by NSF Grants DMS-1403708 and DMS-1712735, and NIH Grant R01 GM-123056. The research of Dylan S. Small was supported in part by NSF Grant SES-1260782.

Appendix A. Supplementary data

Supplementary material including additional simulation results, endogeneity test for invalid instrumental variables and all the proofs can be found online at https://www.sciencedirect.com/science/article/pii/S0304407618301325?via%3Dihub.

References

Andrews, D.W.K., Moreira, M.J., Stock, J.H., 2007. Performance of conditional Wald tests in {IV} regression with weak instruments. J. Econometrics 139 (1), 116–132.
Angrist, J.D., Imbens, G.W., Rubin, D.B., 1996. Identification of causal effects using instrumental variables. J. Amer. Statist. Assoc. 91 (434), 444–455.
Baiocchi, M., Cheng, J., Small, D.S., 2014. Instrumental variable methods for causal inference. Stat. Med. 33 (13), 2297–2340.
Baum, C.F., Schaffer, M.E., Stillman, S., 2007. ivreg2: Stata module for extended instrumental variables/2SLS, GMM and AC/HAC, LIML and k-class regression, vol. 2007. Boston College Department of Economics, Statistical Software Components S, p. 425401.
Bekker, P.A., 1994. Alternative approximations to the distributions of instrumental variable estimators. Econometrica 657–681.
Belloni, A., Chen, D., Chernozhukov, V., Hansen, C., 2012. Sparse models and methods for optimal instruments with an application to eminent domain. Econometrica 80 (6), 2369–2429.
Belloni, A., Chernozhukov, V., Fernández-Val, I., Hansen, C., 2013. Program evaluation with high-dimensional data. arXiv preprint arXiv:1311.2645.
Belloni, A., Chernozhukov, V., Hansen, C., 2011a. Inference for high-dimensional sparse econometric models, arXiv preprint arXiv:1201.0220.
Belloni, A., Chernozhukov, V., Hansen, C., 2014. High-dimensional methods and inference on structural and treatment effects. J. Econ. Perspect. 28 (2), 29–50.
Belloni, A., Chernozhukov, V., Wang, L., 2011b. Square-root lasso: pivotal recovery of sparse signals via conic programming. Biometrika 98 (4), 791–806.
Bound, J., Jaeger, D.A., Baker, R.M., 1995. Problems with instrumental variables estimation when the correlation between the instruments and the endogeneous explanatory variable is weak. J. Amer. Statist. Assoc. 90 (430), 443–450.
Breusch, T., Qian, H., Schmidt, P., Wyhowski, D., 1999. Redundancy of moment conditions. J. Econometrics 91 (1), 89–111.
Cai, T.T., Guo, Z., 2017. Confidence intervals for high-dimensional linear regression: minimax rates and adaptivity. Ann. Statist. 45 (2), 615–646.
Card, D., 1999. Chapter 30 - the causal effect of education on earnings. In: Ashenfelter, O.C., Card, D. (Eds.), Handbook of Labor Economics, vol. 3, Part A. Elsevier, pp. 1801–1863.
Cawley, J., Frisvold, D., Meyerhoefer, C., 2013. The impact of physical education on obesity among elementary school children. J. Health Econ. 32 (4), 743–755.
Chao, J.C., Hausman, J.A., Newey, W.K., Swanson, N.R., Woutersen, T., 2014. Testing overidentifying restrictions with many instruments and heteroskedasticity. J. Econometrics 178, 15–21.
Chao, J.C., Swanson, N.R., 2005. Consistent estimation with a large number of weak instruments. Econometrica 73 (5), 1673–1692.
Cheng, X., Liao, Z., 2015. Select the valid and relevant moments: an information-based lasso for gmm with many moments. J. Econometrics 186 (2), 443–464.
Chernozhukov, V., Hansen, C., Spindler, M., 2014. Valid post-selection and post-regularization inference: an elementary, general approach.
Chernozhukov, V., Hansen, C., Spindler, M., 2015. Post-selection and post-regularization inference in linear models with many controls and instruments. Amer. Econ. Rev. 105 (5), 486–490.
Chmelarova, V., University, L.S., College, A..M., 2007. The Hausman Test, and Some Alternatives, with Heteroskedastic Data. Louisiana State University and Agricultural & Mechanical College.
Conley, T.G., Hansen, C.B., Rossi, P.E., 2012. Plausibly exogenous. Rev. Econ. Stat. 94 (1), 260–272.
Davidson, R., MacKinnon, J.G., 1993. Estimation and Inference in Econometrics. Oxford University Press, New York.
Doko Tchatoka, F., 2015. On bootstrap validity for specification tests with weak instruments. Econom. J. 18 (1), 137–146.
Dufour, J.M., 1997. Some impossibility theorems in econometrics with applications to structural and dynamic models. Econometrica 1365–1387.
Durbin, J., 1954. Errors in variables. Rev. Int. Stat. Inst. 22, 23–32.
Fan, J., Liao, Y., 2014. Endogeneity in high dimensions. Ann. Statist. 42 (3), 872.
Fan, J., Shao, Q.M., Zhou, W.X., 2015. Are discoveries spurious? distributions of maximum spurious correlations and their applications, arXiv preprint arXiv:1502.04237.
Gautier, E., Tsybakov, A.B., 2011. High-dimensional instrumental variables regression and confidence sets, arXiv preprint arXiv:1105.2454.
Guggenberger, P., 2010. The impact of a hausman pretest on the asymptotic size of a hypothesis test. Econ. Theory 26 (2), 369–382.
Guo, Z., Kang, H., Cai, T.T., Small, S.D., 2016. Confidence intervals for causal effects with invalid instruments using two-stage hard thresholding. arXiv preprint arXiv:1603.05224.
Hahn, J., Ham, J.C., Moon, H.R., 2011. The hausman test and weak instruments. J. Econometrics 160 (2), 289–299.
Hahn, J., Hausman, J., 2002. A new specification test for the validity of instrumental variables. Econometrica 70 (1), 163–189.
Hahn, J., Hausman, J., 2005. Estimation with valid and invalid instruments. Ann. Stat. 79/80, 25–57.
Hall, A., Peixe, F.P.M., 2003. A consistent method for the selection of relevant instruments. Econometric Rev. 22 (3), 269–287.

Han, C., Phillips, P.C., 2006. GMM with many moment conditions. Econometrica 74 (1), 147–192.
Hansen, C., Hausman, J., Newey, W., 2008. Estimation with many instrumental variables. J. Bus. Econom. Statist. 398–422.
Hansen, L.P., 1982. Large sample properties of generalized method of moments estimators. Econometrica 1029–1054.
Hausman, J., 1978. Specification tests in econometrics. Econometrica 41, 1251–1271.
Hausman, J., Stock, J.H., Yogo, M., 2005. Asymptotic properties of the Hahn–Hausman test for weak-instruments. Econom. Lett. 89 (3), 333–342.
Hernán, M.A., Robins, J.M., 2006. Instruments for causal inference: an epidemiologist's dream?. Epidemiology 17 (4), 360–372.
Holland, P.W., 1988. Causal inference, path analysis, and recursive structural equations models. Sociol. Methodol. 18 (1), 449–484.
Imbens, G.W., Angrist, J.D., 1994. Identification and estimation of local average treatment effects. Econometrica 62 (2), 467–475.
Javanmard, A., Montanari, A., 2014. Confidence intervals and hypothesis testing for high-dimensional regression. J. Mach. Learn. Res. 15 (1), 2869–2909.
Kang, H., Zhang, A., Cai, T.T., Small, D.S., 2016. Instrumental variables estimation with some invalid instruments and its application to mendelian randomization. J. Amer. Statist. Assoc. 111, 132–144.
Kleiber, C., Zeileis, A., 2008. Applied Econometrics with R. Springer-Verlag, New York.
Kleibergen, F., 2002. Pivotal statistics for testing structural parameters in instrumental variables regression. Econometrica 70 (5), 1781–1803.
Kolesár, M., Chetty, R., Friedman, J.N., Glaeser, E.L., Imbens, G.W., 2015. Identification and inference with many invalid instruments. J. Bus. Econom. Statist. 33 (4), 474–484.
Kosec, K., 2014. The child health implications of privatizing Africa's urban water supply. J. Health Econ. 35, 1–19.
Lee, Y., Okui, R., 2012. Hahn–Hausman test as a specification test. J. Econometrics 167 (1), 133–139.
Mariano, R.S., 1973. Approximations to the distribution functions of Theil's k-class estimators. Econometrica 715–721.
Moreira, M.J., 2003. A conditional likelihood ratio test for structural models. Econometrica 71 (4), 1027–1048.
Morimune, K., 1983. Approximate distributions of k-class estimators when the degree of overidentifiability is large compared with the sample size. Econometrica 821–841.
Murray, M.P., 2006. Avoiding invalid instruments and coping with weak instruments. J. Econ. Perspect. 20 (4), 111–132.
Nakamura, A., Nakamura, M., 1981. On the relationships among several specification error tests presented by Durbin, Wu, and Hausman. Econometrica 1583–1588.
Nelson, C.R., Startz, R., 1990. Some further results on the exact sample properties of the instrumental variables estimator. Econometrica 58, 967–976.
Newey, W.K., Windmeijer, F., 2005. GMM with many weak moment conditions.
Ren, Z., Sun, T., Zhang, C.H., Zhou, H.H., 2013. Asymptotic normality and optimalities in estimation of large gaussian graphical model arXiv preprint arXiv:1309.6024.
Sargan, J.D., 1958. The estimation of economic relationships using instrumental variables. Econometrica 393–415.
Staiger, D., Stock, J.H., 1997. Instrumental variables regression with weak instruments. Econometrica 65 (3), 557–586.
Stock, J., Yogo, M., 2005. Testing for weak instruments in linear iv regression. In: Andrews, D.W. (Ed.), Identification and Inference for Econometric Models. Cambridge University Press, New York, pp. 80–108.
Stock, J.H., Wright, J.H., 2000. GMM with weak identification. Econometrica 1055–1096.
Sun, T., Zhang, C.H., 2012. Scaled sparse linear regression. Biometrika 101 (2), 269–284.
Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Stat. Methodol. 58 (1), 267–288.
van de Geer, S., Bühlmann, P., Ritov, Y., Dezeure, R., 2014. On asymptotically optimal confidence regions and tests for high-dimensional models. Ann. Statist. 42 (3), 1166–1202.
Wang, J., Zivot, E., 1998. Inference on structural parameters in instrumental variables regression with weak instruments. Econometrica 66 (6), 1389–1404.
Wooldridge, J.M., 2010. Econometric Analysis of Cross Section and Panel Data, second ed.. MIT press.
Wu, D.M., 1973. Alternative tests of independence between stochastic regressors and disturbances. Econometrica 41, 733–750.
Zhang, C.H., Zhang, S.S., 2014. Confidence intervals for low dimensional parameters in high dimensional linear models. J. R. Stat. Soc. Ser. B Stat. Methodol. 76 (1), 217–242.
Zivot, E., Startz, R., Nelson, C.R., 1998. Valid confidence intervals and inference in the presence of weak instruments. Internat. Econom. Rev. 1119–1144.