Testing Endogeneity with Possibly Invalid Instruments and High Dimensional Covariates

Zijian Guo, Hyunseung Kang, T. Tony Cai and Dylan S. Small

Abstract

The Durbin-Wu-Hausman (DWH) test is a commonly used test for endogeneity in instrumental variables (IV) regression. Unfortunately, the DWH test depends, among other things, on assuming all the instruments are valid, a rarity in practice. In this paper, we show that the DWH test often has distorted size even if one IV is invalid. Also, the DWH test may have low power when many, possibly high dimensional, covariates are used to make the instruments more plausibly valid. To remedy these shortcomings, we propose a new endogeneity test which has proper size and better power when invalid instruments and high dimemsional covariates are present; in low dimensions, the new test is optimal in that its power is equivalent to the "oracle" DWH test's power that knows which instruments are valid. The paper concludes with a simulation study of the new test with invalid instruments and high dimensional covariates.

1 Introduction

Many empirical studies using instrumental variables (IV) regression are accompanied by the Durbin-Wu-Hausman test [Durbin, 1954, Wu, 1973, Hausman, 1978], hereafter called the DWH test. The primary purpose of the DWH test is to test the presence of endogeneity by comparing the ordinary least squares (OLS) estimate of the structural parameters in the IV regression to that of the two-stage least squares (TSLS). Consequently, the DWH test is often used to decide whether to use an IV analysis compared to a standard OLS analysis; the IV analysis has lower bias when the included possibly endogenous variable in the IV regression is truly endogenous whereas the standard OLS analysis has smaller variance [Davidson and MacKinnon, 1993].

Properly using the DWH test depends, among other things, on having instruments that are (i) strongly associated with the endogenous variable, often called strong instruments, and are (ii) known, with absolute certainty, to be exogenous[1], often referred to as valid instruments [Murray, 2006]. For example, when instruments are not strong, Staiger and Stock [1997] showed that the DWH test that used the TSLS estimator for variance, which is attributed to Durbin [1954] and Wu [1973], had distorted size under the null hypothesis while the DWH test that used the OLS estimator for variance, which is attributed to Hausman [1978], had proper size. Unfortunately, when instruments are not valid, which is perhaps a bigger concern in practice [Murray, 2006, Conley et al., 2012] and has arguably received far less attention than when the instruments are weak, there is a paucity of work on exactly characterizing the behavior of the DWH test. Also, with large datasets becoming more prevalent, there has been a trend toward conditioning on many, possibly high dimensional, exogenous covariates in structural models to make instruments more plausibly valid[2] [Gautier and Tsybakov, 2011, Belloni et al., 2012, Chernozhukov et al., 2015]. However, it's unclear how the DWH test is affected in the presence of many covariates. More importantly, even after conditioning, some IVs may still be invalid and subsequent analysis, including the DWH test, assuming that all the IVs are valid after conditioning can be misleading.

Prior work in analyzing the DWH test in instrumental variables is diverse. Estimation and inference under weak instruments are well-documented [Staiger and Stock, 1997, Nelson and Startz, 1990, Bekker, 1994, Bound et al., 1995, Dufour, 1997, Zivot et al., 1998, Wang and Zivot, 1998, Kleibergen, 2002, Moreira, 2003, Chao and Swanson, 2005, Andrews et al., 2007]. In particular, when the instruments are weak, the behavior of the DWH test under the null depends on the variance estimate [Staiger and Stock, 1997, Nakamura and Nakamura, 1981, Doko Tchatoka, 2015]. Some recent

[1] 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 [Holland, 1988, Imbens and Angrist, 1994, Angrist et al., 1996]. 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.

[2] For example, in causal inference and epidemiology, Hernán and Robins [2006] and Baiocchi et al. [2014] highlights the need to control for covariates for an instrument to remove violations of exogeneity.

work extends the specification test to handle growing number of instruments [Hahn and Hausman, 2002, Chao et al., 2014]. Unfortunately, all these works have not characterized the properties of the DWH test when instruments are invalid.

In addition, there is a growing literature on estimation and inference of structural effects in high dimensional instrumental variables models [Gautier and Tsybakov, 2011, Belloni et al., 2012, Chernozhukov et al., 2015, Belloni et al., 2011a, 2013, Fan and Liao, 2014, Chernozhukov et al., 2014]. Unfortunately, all of them assume that after controlling for high dimensional covariates, all the IVs are valid, which may not be true in practice.

Finally, for work related to invalid instruments, Fisher [1966, 1967], Newey [1985], Hahn and Hausman [2005], Guggenberger [2012], Berkowitz et al. [2012] and Caner [2014] considered properties of IV estimators or, more broadly, generalized method of moments estimators (GMM)s when there are local deviations from validity to invalidity. Andrews [1999] and Andrews and Lu [2001] considered selecting valid instruments within the context of GMMs. Small [2007] approached the invalid instrument problem via a sensitivity analysis. Conley et al. [2012] proposed various strategies, including union-bound correction, sensitivity analysis, and Bayesian analysis, to deal with invalid instruments. Liao [2013] and Cheng and Liao [2015] considered the setting where there is, a priori, a known set of valid instruments and another set of instruments that may not be valid. Recently, Kang et al. [2016] and Kolesár et al. [2015] worked under the case where this a priori information is absent. Kang et al. [2016] replaced the a priori knowledge with a sparsity-type assumption on the number of invalid instruments while Kolesár et al. [2015] replaced the a priori knowledge with an orthogonality assumption on the instruments' effects on the included endogenous variable and the outcome. In all cases, the main focus was on the selection of valid instruments or the estimation of structural parameters; none of the authors considered studying the properties of the endogeneity test.

Our main contribution is two-fold. First, we expand the theoretical analysis of the DWH test under weak instruments by showing negative results about the DWH test under invalid instruments and high dimensional covariates. We show that the DWH test fails to have the correct size in the presence of invalid instruments and we precisely characterize the deviation from the nominal level. We also show that the DWH test has low power when many covariates are present, especially when the number of covariates is similar to the sample size. Second, we remedy the failures of the DWH test by presenting an improved endogeneity test that is robust to both invalid instruments and high dimensional covariates and that works in settings where the number of structural parameters exceed the sample size. The key idea behind the new endogeneity test is based on a novel methodology that we call two-stage hard thresholding (TSHT) which allows selection of valid instruments and subsequent inference after selection. In the low dimensional setting, we show that our new test has optimal performance in that our test has the same asymptotic power as the "oracle" DWH test that knows which instruments are valid and invalid. In the high dimensional setting, we characterize the asymptotic power of the new proposed test and show that the power of the new test is better than the DWH test. We conclude the paper with simulation studies comparing the performance of our new test with the DWH test. We find that our test has the desired size and has better power than that of the DWH test when invalid instruments and high dimensional covariates are present. We also present technical proofs and extended simulation studies that further examine the power and sensitivity of our test to regularity assumptions in the supplement.

2 Instrumental Variables Regression and the DWH Test

2.1 Notation

For any p dimensional vector v, the jth element is denoted as v_j. Let $\|v\|_1$, $\|v\|_2$, and $\|v\|_\infty$ denote the $1, 2$ and ∞-norms, respectively. Let $\|v\|_0$ denote the number of non-zero elements in v and let $\text{supp}(v) = \{j : v_j \neq 0\} \subseteq \{1, \ldots, p\}$. For any n by p matrix M, denote the ith row and jth column entry as M_{ij}, the ith row vector as $M_{i.}$, the jth column vector as $M_{.j}$, and M' as the transpose of M. Also, given any n by 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. Let $\|M\|_\infty$ represent the element-wise matrix sup norm of matrix M. 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 $\Lambda_{J|J^C} = \Lambda_{JJ} - \Lambda_{JJ^C}\Lambda^{-1}_{J^C J^C}\Lambda_{J^C J}$ denote the submatrix Λ_{JJ} adjusted for the columns 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 $\mathbf{P}\left(|X_n/a_n| \geq C_0\right) \leq c_0$. 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 distribution of random variables, for any α, $0 < \alpha < 1$, and $B \in \mathbb{R}$, we define

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

where Φ and $z_{\alpha/2}$ are respectively the cumulative distribution function and $\alpha/2$ quantile of a standard normal distribution. We also denote $\chi^2_\alpha(d)$ to be 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 $X'_{i.}$ and the corresponding dimension of $W_{i.}$ to be $p = p_z + p_x$. The matrix W is indexed by the set $\mathcal{I} = \{1, \ldots, p_{\mathrm{z}}\}$ which consists of all the p_{z} candidate instruments and the set $\mathcal{I}^C = \{p_{\mathrm{z}} + 1, \ldots, p\}$ which consists of the p_{x} covariates. The variables (Y_i, D_i, Z_i, X_i) are governed by the following structural model

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

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

where β, π, ϕ, γ, and ψ are unknown parameters in the model and without loss of generality, we assume the variables are centered to mean zero[3]. The random disturbance terms δ_i and ϵ_i are independent of $(Z_{i.}, X_{i.})$ and, for simplicity, are assumed to be bivariate normal. Let the population covariance matrix of (δ_i, ϵ_i) be Σ, with $\Sigma_{11} = \mathrm{Var}(\delta_i | Z_{i.}, X_{i.})$, $\Sigma_{22} = \mathrm{Var}(\epsilon_i | Z_{i.}, X_{i.})$, and $\Sigma_{12} = \Sigma_{21} = \mathrm{Cov}(\delta_i, \epsilon_i | Z_{i.}, X_{i.})$. Let the second order moments of $W_{i.}$ be $\Lambda = \mathbf{E}\left(W_{i.}W'_{i.}\right)$ with $\Lambda_{\mathcal{I}|\mathcal{I}^c}$ as the adjusted covariance of $W_{i.}$. Let ω represent all the parameters $\omega = (\beta, \pi, \phi, \gamma, \psi, \Sigma)$ from the parameter space $\omega \in \Omega = \{\mathbb{R} \otimes \mathbb{R}^{p_z} \otimes \mathbb{R}^{p_x} \otimes \mathbb{R}^{p_z} \otimes \mathbb{R}^{p_z} \otimes \Sigma \succ 0\}$. Finally, we denote $s_{\mathrm{z2}} = \|\pi\|_0$, $s_{\mathrm{x2}} = \|\phi\|_0$, $s_{\mathrm{z1}} = \|\gamma\|_0$, $s_{\mathrm{x1}} = \|\psi\|_0$ and $s = \max\{s_{\mathrm{z2}}, s_{\mathrm{x2}}, s_{\mathrm{z1}}, s_{\mathrm{x1}}\}$.

If $\pi = 0$ in model (2), the models (2) and (3) represent the usual instrumental variables regression models with one endogenous variable, p_x exogenous covariates, and p_z instruments, all of which are assumed to be valid. On the other hand, if $\pi \neq 0$ and the support of π is unknown a priori,

[3] The mean-centering is equivalent to adding a constant 1 term (i.e. intercept term) in $X'_{i.}$

the instruments may have a direct effect on the outcome, thereby violating the exclusion restriction [Imbens and Angrist, 1994, Angrist et al., 1996] and making them potentially invalid, without knowing, a priori, which are invalid and valid [Murray, 2006, Conley et al., 2012, Kang et al., 2016]. In fact, among p_z candidate instruments, the support of π allows us to distinguish a valid instrument from an invalid one and provides us with a definition of a valid instrument.

Definition 1. *Suppose we have* p_{z} *candidate instruments along with the model* (2)*. We say that instrument* $j = 1, \ldots, p_z$ *is valid if* $\pi_j = 0$ *and invalid if* $\pi_j \neq 0$*.*

In addition, it is also useful to define relevant instruments from the 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 like Definition 1, we use the support of a vector to define the 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.

Definition 2. *Suppose we have* p_{z} *instruments along with the 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* $\mathcal{S}$ *be the set of relevant instruments.*

Definitions 1 and 2 combine to form valid and relevant instruments. We denote the set of instruments that satisfy both definitions as $\mathcal{V} = \{j \mid \pi_j = 0, \gamma_j \neq 0\}$. Note that Definitions 1 and 2 are related to the definition of instruments that is well known in the literature. In particular, if $p_z = 1$, an instrument that is relevant and valid is identical to the definition of an instrument in Holland [1988]. In particular, Definition 1 is the same as ignorability and exclusion restriction while Definition 2 is the same as the condition that the instrument is related to the exposure. Definitions 1 and 2 are also a special case of a definition of an instrument discussed in Angrist et al. [1996] where in our setup, we assume an additive, linear, and a constant treatment effect model. Hence, when multiple instruments, $p_{\mathrm{z}} > 1$, are present, Definitions 1 and 2, especially Definition 1, can be viewed as a generalization of the definition of an instrument in the $p_{\mathrm{z}} = 1$ case.

In the literature, Kang et al. [2016] considered the aforementioned framework and Definition 1 as a relaxation of instrumental variables assumptions where a priori information about which of the p_z instruments are valid or invalid is not available and provided sufficient and necessary conditions for identification of β. Guo et al. [2016] expanded the work in Kang et al. [2016] by allowing high dimensional covariates and instruments along with deriving honest confidence intervals for β. Kolesár et al. [2015] also analyzed the models (2) and (3) in the presence of invalid instruments, but they assumed orthogonality restrictions between π and γ.

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

$$C(\mathcal{V}) = \frac{\gamma_{\mathcal{V}}' \Lambda_{\mathcal{V}|\mathcal{V}^C} \gamma_{\mathcal{V}}}{|\mathcal{V}| \Sigma_{22}}. \tag{4}$$

If all instruments were relevant and valid, then $\mathcal{V} = \mathcal{I}$ and equation (4) is the usual definition of concentration parameter in Staiger and Stock [1997], Bound et al. [1995], Mariano [1973], Stock and Wright [2000] using population quantities, i.e. $\Lambda_{\mathcal{V}|\mathcal{V}^C}$. [4] However, if only a subset of all instruments are relevant and valid so that $\mathcal{V} \subset \{1, \ldots, p_{\mathrm{z}}\}$, then the concentration parameter represents the strength of the instruments for that subset $\mathcal{V}$, adjusted for the instruments in its complement $\mathcal{V}^C = \{1, \ldots, p_z, p_z + 1, \ldots, p\} \setminus \mathcal{V}$. Regardless, like the usual concentration parameter, a high value of $C(\mathcal{V})$ represents strong instruments in the set $\mathcal{V}$ while a low value of $C(\mathcal{V})$ represents weak instruments.

2.3 The DWH Test

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

$$H_0 : \Sigma_{12} = 0, \quad H_1 : \Sigma_{12} \neq 0, \tag{5}$$

The DWH test tests for endogeneity as specified by the hypothesis in equation (5) by comparing two consistent estimators of β under the null hypothesis H_0, i.e. no endogeneity, with different efficiencies. Specifically, the DWH test statistic, denoted as Q_{DWH}, is the quadratic difference of the the OLS estimate of β, denoted as $\widehat{\beta}_{\mathrm{OLS}} = (D' P_{X^\perp} D)^{-1} D' P_{X^\perp} Y$, and the TSLS estimate of β, denoted as $\widehat{\beta}_{\mathrm{TSLS}} = (D'(P_W - P_X)D)^{-1} D'(P_W - P_X) Y$,

$$Q_{\mathrm{DWH}} = \frac{(\widehat{\beta}_{\mathrm{TSLS}} - \widehat{\beta}_{\mathrm{OLS}})^2}{\widehat{\mathrm{Var}}(\widehat{\beta}_{\mathrm{TSLS}}) - \widehat{\mathrm{Var}}(\widehat{\beta}_{\mathrm{OLS}})} \tag{6}$$

[4] For example, if $\mathcal{V} = \mathcal{I}$ so that all IVs are valid, $C(\mathcal{V})$ corresponds exactly to the quantity $\lambda'\lambda/K_2$ on page 561 of Staiger and Stock [1997] for $n = 1$ and $K_1 = 0$. Without using population quantities, $nC(\mathcal{V})$ roughly corresponds to the usual concentration parameter using the estimated version of $\Lambda_{\mathcal{V}|\mathcal{V}^C}$

where $\widehat{\mathrm{Var}}(\widehat{\beta}_{\mathrm{OLS}}) = (D' P_{X^\perp} D)^{-1} \widehat{\Sigma}_{11}$, $\widehat{\mathrm{Var}}(\widehat{\beta}_{\mathrm{TSLS}}) = (D'(P_W - P_X)D)^{-1} \widehat{\Sigma}_{11}$, and $\widehat{\Sigma}_{11}$ can either be the OLS estimate of Σ, i.e. $\widehat{\Sigma}_{11} = \|Y - D\widehat{\beta}_{\mathrm{OLS}} - X\widehat{\phi}_{\mathrm{OLS}}\|_2^2/n$, or the TSLS estimate of Σ, i.e. $\widehat{\Sigma}_{11} = \|Y - D\widehat{\beta}_{\mathrm{TSLS}} - X\widehat{\phi}_{\mathrm{TSLS}}\|_2^2/n$[5]. Under H_0, both OLS and TSLS are consistent estimates of β, but the OLS estimate is more efficient than TSLS. Also, under H_0, both OLS and TSLS estimates of the variance Σ_{11} are consistent.

If $\pi = 0$ so that all the instruments are valid, the asymptotic null distribution of the DWH test in equation (6) is Chi-squared with one degree of freedom. With a known Σ_{11}, the DWH test has an exact Chi-squared null distribution with one degree of freedom. Regardless, both null distributions imply that for $\alpha, 0 < \alpha < 1$, we can reject the null hypothesis H_0 for the alternative H_1 by using the decision rule,

$$\text{Reject } H_0 \text{ if } \quad Q_{\mathrm{DWH}} \geq \chi^2_\alpha(1)$$

and the Type I error of the DWH test will be exactly or asymptotically controlled at level α. Also, under the local alternative hypotheses,

$$H_0 : \Sigma_{12} = 0, \quad H_2 : \Sigma_{12} = \frac{\Delta_1}{\sqrt{n}} \tag{7}$$

for some constant Δ_1, the asymptotic power of the DWH test is

$$\omega \in H_2 : \lim_{n \to \infty} \mathbf{P}(Q_{\mathrm{DWH}} \geq \chi^2_\alpha(1)) = G\left(\alpha, \frac{\Delta_1 \sqrt{C(\mathcal{I})}}{\sqrt{\left(C(\mathcal{I}) + \frac{1}{p_z}\right)\Sigma_{11}\Sigma_{22}}}\right). \tag{8}$$

For more information about the DWH test, see Davidson and MacKinnon [1993] and Wooldridge [2010] for textbook discussions.

3 Failure of the DWH Test

3.1 Invalid Instruments

While the DWH test performs as expected when all the instruments are valid, in practice, some instruments may be invalid and consequently, the DWH test can be a highly misleading assessment of the hypotheses (5). In Theorem 1, we show that the Type I error of the DWH test can be greater

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

than the nominal level for a wide range of IV configurations in which some IVs are invalid; we assume a known Σ_{11} in Theorem 1 for a cleaner technical exposition and to highlight the impact that invalid IVs have on the size and power of the DWH test, but the known Σ_{11} can be replaced by a consistent estimate of Σ_{11}. We also show that the power of the DWH test under the local alternative H_2 in equation (7) can be shifted.

Theorem 1. *Suppose we have models* (2) *and* (3) *with a known* Σ_{11}*. If* $\pi = \Delta_2/n^k$ *where* Δ_2 *is a fixed constant and* $0 \leq k < \infty$*, then for any* α*,* $0 < \alpha < 1$*, we have the following asymptotic phase-transition behaviors of the DWH test for different values of* k*.*

a. $0 \leq k < 1/2$*: The asymptotic Type I error of the DWH test under* H_0 *is* 1*, i.e.*

$$\omega \in H_0 : \lim_{n \to \infty} \mathbf{P}\left(Q_{\text{DWH}} \geq \chi^2_\alpha(1)\right) = 1 \tag{9}$$

and subsequently, the asymptotic power of the DWH test under H_2 *is 1.*

b. $k = 1/2$*: The asymptotic Type I error of the DWH test under* H_0 *is*

$$\omega \in H_0 : \lim_{n \to \infty} \mathbf{P}\left(Q_{\text{DWH}} \geq \chi^2_\alpha(1)\right) = G\left(\alpha, \frac{\frac{1}{p_z}\gamma' \Lambda_{\mathcal{I}|\mathcal{I}^c}\Delta_2}{\sqrt{C(\mathcal{I})\left(C(\mathcal{I}) + \frac{1}{p_z}\right)\Sigma_{11}\Sigma_{22}}}\right) \geq \alpha \tag{10}$$

and the asymptotic power of the DWH test under H_2 *is*

$$\begin{aligned}\omega \in H_2 : &\lim_{n \to \infty} \mathbf{P}\left(Q_{\text{DWH}} \geq \chi^2_\alpha(1)\right) \\ &= G\left(\alpha, \frac{\frac{1}{p_z}\gamma' \Lambda_{\mathcal{I}|\mathcal{I}^c}\Delta_2}{\sqrt{C(\mathcal{I})\left(C(\mathcal{I}) + \frac{1}{p_z}\right)\Sigma_{11}\Sigma_{22}}} + \frac{\Delta_1\sqrt{C(\mathcal{I})}}{\sqrt{\left(C(\mathcal{I}) + \frac{1}{p_z}\right)\Sigma_{11}\Sigma_{22}}}\right)\end{aligned} \tag{11}$$

c. $1/2 < k < \infty$*: The asymptotic Type I error of the DWH test is* α*, i.e.*

$$\omega \in H_0 : \lim_{n \to \infty} \mathbf{P}\left(Q_{\text{DWH}} \geq \chi^2_\alpha(1)\right) = \alpha \tag{12}$$

and the asymptotic power of the DWH test under H_2 *is equivalent to equation* (8)*.*

Theorem 1 presents the asymptotic behavior of the DWH test under a wide range of behaviors for the invalid IVs as represented by π. For example, when the instruments are invalid in the sense that their deviation from valid IVs (i.e. $\pi = 0$) to invalid IVs (i.e. $\pi \neq 0$) is at rates slower than $n^{-1/2}$, say $\pi = \Delta_2 n^{-1/4}$ or $\pi = \Delta_2$, equation (9) states that the DWH will always have Type I error that reaches 1. In other words, if some IVs, or even a single IV, are moderately (or strongly) invalid in the sense that they have moderate (or strong) direct effects on the outcome above the usual noise level of the model error terms at $n^{-1/2}$, then the DWH test will always reject the null hypothesis of no endogeneity even if there is truly no endogeneity present.

Next, suppose the instruments are invalid in the sense that their deviation from valid IVs to invalid IVs are exactly at $n^{-1/2}$ rate, also referred to as the Pitman drift.[6] This is the phase-transition point of the DWH test's Type I error as the error moves from 1 in equation (9) to α in equation (12). Under this type of invalidity, equation (10) shows that the Type I error of the DWH test depends on some factors, most prominently the factor $\gamma' \Lambda_{\mathcal{I}|\mathcal{I}^c} \Delta_2$. The factor $\gamma' \Lambda_{\mathcal{I}|\mathcal{I}^c} \Delta_2$ has been discussed in the literature, most recently by Kolesár et al. [2015] within the context of invalid IVs. Specifically, Kolesár et al. [2015] studied the case where $\Delta_2 \neq 0$ so that there are invalid IVs, but $\gamma' \Lambda_{\mathcal{I}|\mathcal{I}^c} \Delta_2 = 0$, which essentially amounted to saying that the IVs' effect on the endogenous variable D via γ is orthogonal to their direct effects on the outcome via Δ_2; see Assumption 5 of Section 3 in Kolesár et al. [2015] for details. Under their scenario, if $\gamma' \Lambda_{\mathcal{I}|\mathcal{I}^c} \Delta_2 = 0$, then the DWH test will have the desired size α. However, if $\gamma' \Lambda_{\mathcal{I}|\mathcal{I}^c} \Delta_2$ is not exactly zero, which will most likely be the case in practice, then the Type I error of the DWH test will always be larger than α and we can compute the exact deviation from α by using equation (10). Also, equation 11 computes the power under H_2 in the $n^{-1/2}$ setting, which again depends on the magnitude and direction of $\gamma' \Lambda_{\mathcal{I}|\mathcal{I}^c} \Delta_2$. For example, if there is only one instrument and that instrument has average negative effects on both D and Y, the overall effect on the power curve will be a positive shift away from the null value $\Sigma_{12} = 0$. Regardless, under the $n^{-1/2}$ invalid IV regime, the DWH test will always have size that is at least as large as α if invalid IVs are present.

Theorem 1 also shows that instruments' strength, as measured by the population concentration parameter $C(\mathcal{I})$ in equation (4), impacts the Type

[6] Fisher [1967] and Newey [1985] have used this type of $n^{-1/2}$ asymptotic argument to study misspecified econometrics models, specifically Section 2, equation (2.3) of Fisher [1967] and Section 2, Assumption 2 of Newey [1985]. More recently, Hahn and Hausman [2005] and Berkowitz et al. [2012] used the $n^{-1/2}$ asymptotic framework in their respective works to study plausibly exogenous variables.

I error rate of the DWH test when the IVs are invalid at the $n^{-1/2}$ rate. Specifically, if $\pi = \Delta_2 n^{-1/2}$ and the instruments are strong so that the concentration parameter $C(\mathcal{I})$ is large, then the deviation from α will be relatively minor even if $\gamma' \Lambda_{\mathcal{I}|\mathcal{I}^c} \Delta_2 \neq 0$. This phenomena has been mentioned in previous work, most notably Bound et al. [1995] and Angrist et al. [1996] where strong instruments can lessen the undesirable effects caused by invalid IVs. However, if the invalidity is at a rate other than $n^{-1/2}$, the strength of the IV has no impact on the Type I error of the DWH test. For instance, if π is slower than $n^{-1/2}$ at $\pi = \Delta_2 n^{-1/4}$, the Type I error rate of the DWH test will always be 1 for any value of the concentration parameter.

Finally, if the instruments are invalid in the sense that their deviation from $\pi = 0$ is faster than $n^{-1/2}$, say $\pi = \Delta n^{-1}$, then equation (12) shows that the DWH test maintains its desired size. To put this invalid IV regime in context, if the instruments are invalid at n^{-k} where $k > 1/2$, the convergence toward $\pi = 0$ is faster than the usual convergence rate of a sample mean from an i.i.d. sample towards a population mean. Also, this type of deviation is equivalent to saying that the invalid IVs are very weakly invalid and essentially act as if they are valid because the IVs are below the noise level of the model error terms at $n^{-1/2}$. Consequently, the DWH test is not impacted by these type of IVs with respect to size and power.

In short, Theorem 1 shows that the DWH test can fail with regards to not being able to control its Type I error. Indeed, the only cases when the DWH test achieves its desired size α are (i) when the invalid IVs essentially behave as valid IVs asymptotically, i.e. the case when $1/2 < k < \infty$, and (ii) when the IVs' effects on the endogenous variables are orthogonal to each other. However, in all other cases, which are arguably more realistic, the DWH test will have Type I error that is strictly greater than α.

3.2 Large Number of Covariates

Next, we consider the behavior of the DWH test when the instruments are assumed to be valid after conditioning on many covariates. As noted in Section 1, many empirical studies often condition on covariates to make instruments more plausibly valid, with the likelihood of having valid IVs increasing as one conditions on more covariates. Theoretically, this setting was studied by Gautier and Tsybakov [2011], Belloni et al. [2012, 2011a, 2013], Fan and Liao [2014], Chernozhukov et al. [2014] and Chernozhukov et al. [2015], who provided honest confidence intervals for a treatment effect, even in the case when the number of covariates and instruments exceeded the sample size. The authors have not studied the behavior of the DWH test under this scenario, specifically the effect on the DWH test by having many covariates to make IVs valid. As we will see below, a tradeoff of having more covariates to make IVs valid is a reduction in the power of the DWH test.

Formally, suppose the number of covariates and instruments are growing with sample size n, $p_{\mathrm{x}} = p_{\mathrm{x}}(n)$ and $p_{\mathrm{z}} = p_{\mathrm{z}}(n)$, so that $p = p_{\mathrm{x}} + p_{\mathrm{z}}$ and $n - p$ are increasing with respect to n. We assume $p \leq n$ 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. As in Section 3.1, we assume a known Σ_{11} for simpler exposition. But, more importantly, we assume that after conditioning on many covariates, our IVs are valid and $\pi = 0$. Theorem 2 characterizes the asymptotic behavior of the decision rule for the DWH test under this regime.

Theorem 2. *Suppose we have models* (2) *and* (3) *with a known* Σ_{11}, $\pi = 0$, *and* $W_{i.}$ *is a zero-mean multivariate Gaussian. If* $\sqrt{C(\mathcal{I})} \gg \sqrt{\log(n - p_{\mathrm{x}})/(n - p_{\mathrm{x}})p_{\mathrm{z}}}$, *for any* α, $0 < \alpha < 1$, *the asymptotic Type I error of the DWH test under* H_0 *is controlled at* α

$$\omega \in H_0 : \quad \limsup_{n, p_{\mathrm{x}}, p_{\mathrm{z}} \to \infty} \mathbf{P}\left(|Q_{\mathrm{DWH}}| \geq z_{\alpha/2}\right) = \alpha.$$

and the asymptotic power of the DWH test under H_2 *is*

$$\omega \in H_2 : \quad \lim_{n, p_{\mathrm{x}}, p_{\mathrm{z}} \to \infty} \left| \mathbf{P}\left(Q_{\mathrm{DWH}} \geq \chi^2_{\alpha}(1)\right) - G\left(\alpha, \frac{C(\mathcal{I})\Delta_1 \sqrt{1 - \frac{p}{n}}}{\sqrt{\left(C(\mathcal{I}) + \frac{1}{n - p_{\mathrm{x}}}\right)\left(C(\mathcal{I}) + \frac{1}{p_{\mathrm{z}}}\right)\Sigma_{11}\Sigma_{22}}}\right)\right| = 0 \tag{13}$$

Theorem 2 characterizes the asymptotic behavior of the DWH test under the regime where the covariates (or instruments) may grow with the sample size and the instruments are valid after conditioning on said covariates. Unlike Theorem 1, the DWH test asymptotically controls the Type I error under this regime at level α and equation (13) characterizes the asymptotic power of the DWH under the local alternative H_2. For example, if covariates and/or instruments are growing at $p/n \to 0$, equation (13) reduces to the usual power of the DWH test with fixed p in equation (8). 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 equation (13) equals α for any value of Δ_1.

This phenomena suggests that in the "middle ground" where $p/n \to c$, $0 < c < 1$, the usual DWH test may suffer in terms of power. As a concrete example, if $p_{\mathrm{x}} = n/2$ and $p_{\mathrm{z}} = n/3$ so that $p/n = 5/6$, then $G(\alpha, \cdot)$ in equation (13) reduces to

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

where the approximation sign is for n sufficiently large enough so that $C(\mathcal{I}) + 2/n \approx C(\mathcal{I})$. Under this setting, the power of the DWH test is smaller than the power in equation (8) under the fixed p regime; see also Section 6 for a numerical demonstration of this phenomena. Hence, including many covariates to make an IV valid may lead to poor power of the DWH test.

Finally, we make two additional remarks about Theorem 2. First, Theorem 2 controls the growth of the concentration parameter $C(\mathcal{I})$ to be faster than $\log(n - p_{\mathrm{x}})/(n - p_{\mathrm{x}})p_{\mathrm{z}}$. 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 $p_{\mathrm{z}}/n \to c$ for some constant c. The weak instrument asymptotics of Staiger and Stock [1997] is not directly applicable to our growth condition on $C(\mathcal{I})$ because the asymptotics keeps p_{z} and p_{x} fixed. Second, we can replace the condition that $W_{i.}$ is a zero-mean multivariate Gaussian in Theorem 2 by a condition used in Chernozhukov et al. [2015], specifically page 486 where (i) the vector of instruments $Z_{i\cdot}$ is a linear model of $X_{i\cdot}$, i.e. $Z_{i.}' = X_{i.}'B + \bar{Z}_{i.}'$, (ii) $\bar{Z}_{i\cdot}$ is independent of $X_{i\cdot}$, and (iii)$\bar{Z}_{i\cdot}$ is a multivariate normal distribution and the result in Theorem 2 will hold.

Sections 3.1 and 3.2 show that the usual DWH test can (i) fail to have Type I error control when invalid IVs are present and (ii) suffer from low power when many covariates are added to make the invalid IVs more plausibly valid. Both Theorems 1 and 2 characterize exactly how the DWH test fails and reveal different types of asymptotic phenomena depending on the parameter regime. In subsequent sections, we will address these deficiencies of the DWH test by proposing an improved endogeneity test that overcomes both invalid instruments and high dimensional covariates.

4 An Improved Endogeneity Test

4.1 Overview

Given the failures of the DWH test for endogeneity when invalid instruments and high dimensional variables are present, we present an improved endogeneity test that addresses both of these concerns. Our test covers many settings encountered in practice where it is guaranteed to have correct size in the presence of invalid instruments and have better power than the DWH test with many covariates. Our test also covers the case when the instruments are invalid even after controlling for high dimensional covariates, a generalization of the setting in Section 3.2. Furthermore, our endogeneity test can test endogeneity even if the number of parameters exceeds the sample size.

The key parts of our endogeneity test are (i) well-behaved estimates of reduced-form parameters and (ii) a novel two-stage hard thresholding (TSHT) to deal with invalid instruments. For the first part, consider the models (2) and (3) as reduced-forms models

$$Y_i = Z_{i.}' \Gamma + X_{i.}' \Psi + \xi_i, \tag{14}$$

$$D_i = Z_{i.}' \gamma + X_{i.}' \psi + \epsilon_i. \tag{15}$$

Here, $\Gamma = \beta\gamma + \pi$ and $\Psi = \phi + \beta\psi$ are the parameters of the reduced-form model (14) and $\xi_i = \beta\epsilon_i + \delta_i$ is the reduced-form error term. The errors in the reduced-models have the property that $\mathbf{E}(\xi_i | Z_{i.}, X_{i.}) = 0$ and $\mathbf{E}(\epsilon_i | Z_{i.}, X_{i.}) = 0$ and the covariance matrix of the error terms, denoted as Θ, are such that $\Theta_{11} = \text{Var}(\xi_i | Z_{i.}, X_{i.}) = \Sigma_{11} + 2\beta\Sigma_{12} + \beta^2\Sigma_{22}$, $\Theta_{22} = \text{Var}(\epsilon_i | Z_{i.}, X_{i.})$, and $\Theta_{12} = \text{Cov}(\xi_i, \epsilon_i | Z_{i.}, X_{i.}) = \Sigma_{12} + \beta\Sigma_{22}$. Each equation in the reduced-form model is the classic regression model with covariates $Z_{i.}$ and $X_{i.}$ and outcomes Y_i and D_i, respectively. Our endogeneity test requires any estimator of the reduced-form parameters that are well-behaved, which we define precisely in Section 4.2.

The second part of the endogeneity test is dealing with invalid instruments. Here, we take a novel two-stage hard thresholding approach that was introduced in Guo et al. [2016] to correctly select the valid IVs. Specifically, in the first step, we estimate the set of IVs that are relevant and in the second step, we use the relevant IVs as pilot estimates to find IVs that are valid. Section 4.3 details this procedure.

4.2 Well-Behaved Estimators

The first step of the endogeneity test is the estimation of the reduced-form parameters in equations (14) and (15). As mentioned before, our endogeneity test doesn't require a specific estimator for the reduced-form parameters. Rather, any estimator that is well-behaved as defined in Definition 3 will be sufficient for our endogeneity test.

Definition 3. *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 equations* (14) *and* (15). *The estimates* $(\widehat{\gamma}, \widehat{\Gamma}, \widehat{\Theta}_{11}, \widehat{\Theta}_{22}, \widehat{\Theta}_{12})$ *are well-behaved estimators if they satisfy the following criteria*

(W1) There exists a matrix $\widehat{V} = (\widehat{v}^{[1]}, \cdots, \widehat{v}^{[p_z]})$ *which is a function of* W *such that the reduced-form estimators of the coefficients* $\widehat{\gamma}$ *and* $\widehat{\Gamma}$ *satisfy*

$$\sqrt{n} \| \left(\widehat{\gamma} - \gamma\right) - \widehat{V}'\epsilon\|_\infty = O_p\left(\frac{s \log p}{\sqrt{n}}\right), \quad \sqrt{n} \| \left(\widehat{\Gamma} - \Gamma\right) - \widehat{V}'\xi\|_\infty = O_p\left(\frac{s \log p}{\sqrt{n}}\right). \tag{16}$$

and the matrix $\widehat{V}$ *satisfies*

$$\liminf_{n \to \infty} \inf_{\omega \in \Omega} \mathbf{P}\left(c \leq \min_{1 \leq j \leq p_z} \frac{\|\widehat{v}^{[j]}\|_2}{\sqrt{n}} \leq \max_{1 \leq j \leq p_z} \frac{\|\widehat{v}^{[j]}\|_2}{\sqrt{n}} \leq C,\ c\|\gamma\|_2 \leq \frac{1}{\sqrt{n}} \| \sum_{j \in \mathcal{V}} \gamma_j \widehat{v}^{[j]}\|_2 \right) = 1 \tag{17}$$

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

(W2) The reduced-form estimates of the error variances, $\widehat{\Theta}_{11}$, $\widehat{\Theta}_{22}$, *and* $\widehat{\Theta}_{12}$, *have the following behavior,*

$$\sqrt{n} \max\left\{\left|\widehat{\Theta}_{11} - \frac{1}{n}\xi'\xi\right|, \left|\widehat{\Theta}_{12} - \frac{1}{n}\epsilon'\xi\right|, \left|\widehat{\Theta}_{22} - \frac{1}{n}\epsilon'\epsilon\right|\right\} = O_p\left(\frac{s \log p}{\sqrt{n}}\right). \tag{18}$$

In the literature, there are many estimators for the reduced-form parameters that are well-behaved as specified in Definition 3. Some examples of well-behaved estimators are listed below.

1. (OLS): In low dimensional settings where p is fixed, the OLS estimates of the reduced-form parameters, i.e.

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

   trivially satisfy conditions for well-defined estimators. Specifically, let $\widehat{V}' = (W'W)^{-1}_{\mathcal{I}\cdot}W$. Then equation (16) holds because $(\widehat{\gamma} - \gamma) -$

$\widehat{V}'\epsilon = 0$ and $\left(\widehat{\Gamma} - \Gamma\right) - \widehat{V}'\xi = 0$. Also, equation (17) holds because $n^{-1/2}\|\widehat{v}^{[j]}\|_2 \xrightarrow{p} \Lambda_{jj}^{-1}$ and $n^{-1}\widehat{V}'\widehat{V} \xrightarrow{p} \Lambda_{\mathcal{I}\mathcal{I}}^{-1}$, 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 equation (18) goes to zero at $n^{-1/2}$ rate.

2. (Debiased Lasso Estimates) In high dimensional settings where p is growing with n, one of the most popular estimators for regression model parameters is the Lasso [Tibshirani, 1996]. Unfortunately, the Lasso estimator, let alone many penalized estimators, do not satisfy the definition of a well-defined estimator, specifically (W1), because penalized estimators are typically biased. Recent works by Zhang and Zhang [2014], Javanmard and Montanari [2014], van de Geer et al. [2014] and Cai and Guo [2016] remedied this bias problem by doing a bias correction on the original penalized estimates.

   As an example, suppose we use the square root Lasso estimator by Belloni et al. [2011b],

$$\{\widetilde{\Gamma}, \widetilde{\Psi}\} = \operatorname*{argmin}_{\Gamma \in \mathbb{R}^{p_z}, \Psi \in \mathbb{R}^{p_x}} \frac{\|Y - Z\Gamma - X\Psi\|_2}{\sqrt{n}} + \frac{\lambda_0}{\sqrt{n}} \left(\sum_{j=1}^{p_z} \|Z_{.j}\|_2 |\Gamma_j| + \sum_{j=1}^{p_x} \|X_{.j}\|_2 |\Psi_j| \right) \tag{19}$$

   for the reduced-form model (14) and

$$\{\widetilde{\gamma}, \widetilde{\psi}\} = \operatorname*{argmin}_{\Gamma \in \mathbb{R}^{p_z}, \Psi \in \mathbb{R}^{p_x}} \frac{\|D - Z\gamma - X\psi\|_2}{\sqrt{n}} + \frac{\lambda_0}{\sqrt{n}} \left(\sum_{j=1}^{p_z} \|Z_{.j}\|_2 |\gamma_j| + \sum_{j=1}^{p_x} \|X_{.j}\|_2 |\psi_j| \right) \tag{20}$$

   for the reduced-form model (15). The term λ_0 in both estimation problems (19) and (20) represents the penalty term in the square root Lasso estimator and typically, in practice, the penalty is set at $\lambda_0 = \sqrt{a_0 \log p/n}$ for some small constant a_0 greater than 2. To transform the above penalized estimators in equations (19) and (20) into well-behaved estimators as defined in Definition 3, we can 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

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

   Typically, the tuning parameter λ_n is chosen to be $12M_1^2\sqrt{\log p/n}$ where M_1 defined as the largest eigenvalue of Λ. Define $\widehat{v}^{[j]} = W\widehat{u}^{[j]}$

and $\widehat{V} = (\widehat{v}^{[1]}, \cdots, \widehat{v}^{[p_z]})$. Then, we can transform the penalized estimators in (19) and (20) into the debiased, well-behaved estimators, $\widehat{\Gamma}$ and $\widehat{\gamma}$

$$\widehat{\Gamma} = \widetilde{\Gamma} + \frac{1}{n}\widehat{V}'\left(Y - Z\widetilde{\Gamma} - X\widetilde{\Psi}\right), \quad \widehat{\gamma} = \widetilde{\gamma} + \frac{1}{n}\widehat{V}'\left(D - Z\widetilde{\gamma} - X\widetilde{\psi}\right). \tag{21}$$

Lemma 3, 4 and 11 in Guo et al. [2016] show that the above estimators for the reduced-form coefficients, $\widehat{\gamma}$ and $\widehat{\Gamma}$ in equation (21), 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 using the estimates from equations (19) and (20)

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

Lemma 3 and equation (180) of Guo et al. [2016] show that the above estimators of $\widehat{\Theta}_{11}, \widehat{\Theta}_{22}$ and $\widehat{\Theta}_{12}$ in equation (22) satisfy (W2). In summary, the debiased Lasso estimators in (21) and the variance estimators in (22) 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.

$$\widehat{\Gamma} = \widetilde{\Gamma} + \widehat{\Lambda^{-1}}_{\mathcal{I},\cdot}\left(Y - Z\widetilde{\Gamma} - X\widetilde{\Psi}\right), \quad \widehat{\gamma} = \widetilde{\gamma} + \widehat{\Lambda^{-1}}_{\mathcal{I},\cdot}\left(D - Z\widetilde{\gamma} - X\widetilde{\psi}\right). \tag{23}$$

where $\widetilde{\Gamma}$ and $\widetilde{\gamma}$ and $\widehat{\Lambda^{-1}}$ are initial estimators of Γ, γ and Λ^{-1}. 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, using the arguments in Theorem 2.1 of van de Geer et al. [2014], 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. 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 equation (22), satisfying (W2).

In short, the first part of our endogeneity test requires any estimator that is well-behaved. As illustrated above, many estimators in the literature satisfy the criteria for a well-behaved estimator laid out in Definition 3. Most notably, the OLS estimator in low dimensions and various versions of the Lasso estimator in high dimensions are well-behaved.

4.3 Two-Stage Hard Thresholding for Invalid Instruments

Once we have well-behaved estimators $(\widehat{\gamma}, \widehat{\Gamma}, \widehat{\Theta}_{11}, \widehat{\Theta}_{22}, \widehat{\Theta}_{12})$ satisfying Definition 3, we can proceed with the second step of our endogeneity test, which is dealing with invalid instruments. This step essentially amounts to selecting relevant and valid IVs among p_z candidate IVs, i.e. the set $\mathcal{V}$, so that we can use this set $\mathcal{V}$ to properly calibrate our well-behaved reduced-form estimates we obtained in Section 4.2. The estimation of this set $\mathcal{V}$ occurs in two stages, which are elaborated below.

In the first stage, we find IVs that are relevant, that is the set $\mathcal{S}$ in Definition 2 comprised of $\gamma_j \neq 0$, by thresholding the estimate $\widehat{\gamma}$ of γ

$$\widehat{\mathcal{S}} = \left\{ j : |\widehat{\gamma}_j| \geq \frac{\sqrt{\widehat{\Theta}_{22}} \|\widehat{v}^{[j]}\|_2}{\sqrt{n}} \sqrt{\frac{a_0 \log \max(p_{\mathrm{z}}, n)}{n}} \right\}. \tag{24}$$

The set $\widehat{\mathcal{S}}$ is an estimate of $\mathcal{S}$ and a_0 is some small constant greater than 2; from our experience, $a_0 = 2$ or $a_0 = 2.05$ work well in practice. The threshold is based on the noise level of $\widehat{\gamma}_j$ in equation (16) (represented by the term $n^{-1}\sqrt{\widehat{\Theta}_{22}} \|\widehat{v}^{[j]}\|_2$), adjusted by the dimensionality of the instrument size (represented by the term $\sqrt{a_0 \log \max(p_{\mathrm{z}}, n)}$).

In the second stage, we use the estimated set of relevant instruments in the first stage and select IVs that are valid, i.e. IVs where $\pi_j = 0$. Specifically, we take each instrument j in $\widehat{\mathcal{S}}$ that is estimated to be relevant and we define $\widehat{\beta}^{[j]}$ to be a "pilot" estimate of β by using this IV and dividing the reduced-form parameter estimates, i.e. $\widehat{\beta}^{[j]} = \widehat{\Gamma}_j / \widehat{\gamma}_j$. We also define $\widehat{\pi}^{[j]}$ to be a pilot estimate of π using this jth instrument's estimate of β, i.e. $\widetilde{\pi}^{[j]} = \widehat{\Gamma} - \widehat{\beta}^{[j]}\widehat{\gamma}$, and $\widehat{\Sigma}_{11}^{[j]}$ to be the pilot estimate of Σ_{11}, i.e. $\widehat{\Sigma}_{11}^{[j]} = \widehat{\Theta}_{11} + (\widehat{\beta}^{[j]})^2\widehat{\Theta}_{22} - 2\widehat{\beta}^{[j]}\widehat{\Theta}_{12}$. Then, for each $\widetilde{\pi}^{[j]}$ in $j \in \widehat{\mathcal{S}}$, we threshold each element of $\widetilde{\pi}^{[j]}$ to create the thresholded estimate $\widehat{\pi}^{[j]}$,

$$\widehat{\pi}_k^{[j]} = \widetilde{\pi}_k^{[j]} \mathbf{1}\left(k \in \widehat{\mathcal{S}} \ \cap \ |\widetilde{\pi}_k^{[j]}| \geq a_0 \sqrt{\widehat{\Sigma}_{11}^{[j]}} \frac{\|\widehat{v}^{[k]} - \frac{\widehat{\gamma}_k}{\widehat{\gamma}_j}\widehat{v}^{[j]}\|_2}{\sqrt{n}} \sqrt{\frac{\log \max(p_{\mathrm{z}}, n)}{n}}\right) \tag{25}$$

for all $1 \leq k \leq p_{\mathrm{z}}$. Each thresholded estimate $\widehat{\pi}^{[j]}$ is obtained by looking at the elements of the un-thresholded estimate, $\widetilde{\pi}^{[j]}$, and examining whether each element exceeds the noise threshold (represented by the term $n^{-1}\sqrt{\widehat{\Sigma}_{11}^{[j]}}\|\widehat{v}^{[k]} - \frac{\widehat{\gamma}_k}{\widehat{\gamma}_j}\widehat{v}^{[j]}\|_2$), adjusted for the multiplicity of the selection procedure (represented by the term $a_0\sqrt{\log\max(p_{\mathrm{z}}, n)}$). Among the $|\widehat{\mathcal{S}}|$ candidate estimates of π based on each instrument in $\widehat{\mathcal{S}}$, i.e. $\widehat{\pi}^{[j]}$, we choose $\widehat{\pi}^{[j]}$ with the most valid instruments, i.e. we choose $j^* \in \widehat{\mathcal{S}}$ where $j^* = \operatorname{argmin} \|\widehat{\pi}^{[j]}\|_0$; if there is a non-unique solution, we choose $\widehat{\pi}^{[j]}$ with the smallest ℓ_1 norm, the closest convex norm of ℓ_0. Subsequently, we can estimate the set of valid instruments $\widehat{\mathcal{V}} \subseteq \{1, \ldots, p_z\}$ as those elements of $\widehat{\pi}^{[j^*]}$ that are zero,

$$\widehat{\mathcal{V}} = \widehat{\mathcal{S}} \setminus \operatorname{supp}\left(\widehat{\pi}^{[j^*]}\right). \tag{26}$$

Then, using the estimated $\widehat{\mathcal{V}}$, we obtain our estimates of parameters Σ_{12}, Σ_{11}, and β

$$\widehat{\Sigma}_{12} = \widehat{\Theta}_{12} - \widehat{\beta}\widehat{\Theta}_{22}, \quad \widehat{\Sigma}_{11} = \widehat{\Theta}_{11} + \widehat{\beta}^2\widehat{\Theta}_{22} - 2\widehat{\beta}\widehat{\Theta}_{12}, \quad \widehat{\beta} = \frac{\sum_{j\in\widehat{\mathcal{V}}} \widehat{\gamma}_j \widehat{\Gamma}_j}{\sum_{j\in\widehat{\mathcal{V}}} \widehat{\gamma}_j^2} \tag{27}$$

Equation (27) provides us with the ingredients to construct our new test for endogeneity, which we denote as Q

$$Q = \frac{\sqrt{n}\widehat{\Sigma}_{12}}{\sqrt{\widehat{\mathrm{Var}}(\widehat{\Sigma}_{12})}}, \quad \widehat{\mathrm{Var}}(\widehat{\Sigma}_{12}) = \widehat{\Theta}_{22}^2\widehat{\mathrm{V}}_1 + \widehat{\mathrm{V}}_2 \tag{28}$$

where $\widehat{\mathrm{V}}_1 = \widehat{\Sigma}_{11}\left\|\sum_{j\in\widehat{\mathrm{V}}} \widehat{\gamma}_j \widehat{v}^{[j]}\right\|_2^2 / \left(\sum_{j\in\widehat{\mathrm{V}}} \widehat{\gamma}_j^2\right)^2$ and $\widehat{\mathrm{V}}_2 = \widehat{\Theta}_{11}\widehat{\Theta}_{22} + \widehat{\Theta}_{12}^2 + 2\widehat{\beta}^2\widehat{\Theta}_{22}^2 - 4\widehat{\beta}\widehat{\Theta}_{12}\widehat{\Theta}_{22}$. Here, V_1 is the variance associated with estimating β and the V_2 is the variance associated with estimating Θ.

Key differences between the original DWH test in equation (6) and our endogeneity test in equation (28) 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 consistency between the OLS and TSLS estimators under the null hypothesis. Also, as we will show in Section 5, unlike the DWH test, our endogeneity test will have proper size and superior power in the presence of invalid instruments and high dimensional covariates.

4.4 Special Cases

While our new endogeneity test in (28) is general in that it handles both low dimensional and high dimensional cases with potentially invalid IVs, we can simplify the procedure above in certain settings and, in some cases, achieve slightly better performance. We discuss two such scenarios, the low-dimensional invalid IV scenario described in Section 3.1 and the high dimensional valid IV scenario described in Section 3.2.

First, in the low dimensional invalid IV scenario described in Section 3.1, we can simply use the OLS estimators for our well-behaved estimators of the reduced-form parameters and replace the estimate of β in (27) with a slightly better estimate of β, which we denote as $\widehat{\beta}_E$.

$$\widehat{\beta}_E = \frac{\widehat{\gamma}'_{\widehat{\mathcal{V}}} \widehat{\Lambda}_{\widehat{\mathcal{V}}|\widehat{\mathcal{V}}^C} \widehat{\Gamma}_{\widehat{\mathcal{V}}}}{\widehat{\gamma}'_{\widehat{\mathcal{V}}} \widehat{\Lambda}_{\widehat{\mathcal{V}}|\widehat{\mathcal{V}}^C} \widehat{\gamma}_{\widehat{\mathcal{V}}}}. \tag{29}$$

We define the test statistic Q_E to be the test statistic Q except we replace the estimate $\widehat{\beta}$ with $\widehat{\beta}_E$. Then, in Section 5.1, under fixed p invalid IV setting, we show that Q_E, unlike the DWH test in this setting, not only controls for Type I error, but also achieves oracle performance in that the power of Q_E under H_2 is asymptotically equivalent to the power of the "oracle" DWH test that has information about instrument validity a priori, or equivalently, has knowledge about $\mathcal{V}$.

Second, in the high dimensional valid IV scenario described in Section 3.2, we can use any of the well-behaved estimators in high dimensional settings and simplify our thresholding procedure by skipping the second thresholding step. Specifically, once we obtained a set of relevant instruments $\widehat{\mathcal{S}}$ in the first thresholding step, we can estimate the set of valid instruments to be $\widehat{\mathcal{V}} = \widehat{\mathcal{S}}$ instead of doing a second thresholding with respect to π. Then, an estimate of β from this modification, denoted as $\widehat{\beta}_H$, would be

$$\widehat{\beta}_H = \frac{\sum_{j \in \widehat{\mathcal{V}}} \widehat{\gamma}_j \widehat{\Gamma}_j}{\sum_{j \in \widehat{\mathcal{V}}} \widehat{\gamma}_j^2}. \tag{30}$$

We define the test statistic Q_H to be the test statistic Q except we replace the estimate $\widehat{\beta}$ with $\widehat{\beta}_H$. Then, in Section 5.2 where p grows with respect to n and instruments are valid after conditioning on growing number of covariates, we show that Q_H has better power than the usual DWH test.

5 Properties of the New Endogeneity Test

5.1 Invalid Instruments in Low Dimensions

We start off the discussion about the properties of our endogeneity test by first, showing that our new test addresses the deficiencies of the DWH test in settings described in Section 3.1 where we are in the low dimensional, fixed p setting with invalid instruments. In addition to the modeling assumptions in equations (2) and (3), we make the following assumption

(IN1) (50% Rule) The number of valid IVs is more than half of the number of non-redundant IVs, that is $|\mathcal{V}| > \frac{1}{2}|\mathcal{S}|$.

We denote the assumption as "IN" since the assumption is specific to the case of invalid IVs. In a nutshell, Assumption (IN1) states that if the number of invalid instruments is not too large, then we can use the observed data to separate the invalid IVs from valid IVs, without knowing a priori which IVs are valid or invalid. Assumption (IN1) is a relaxation of the assumption typical in IV settings where all the IVs are assumed to be valid a priori so that $|\mathcal{V}| = p_{\mathrm{z}}$ and (IN1) holds automatically. In particular, Assumption (IN1) entertains the possibility that some IVs may be invalid, so $|\mathcal{V}| < p_{\mathrm{z}}$, but without knowing a priori which IVs are invalid, i.e. the exact set $\mathcal{V}$. Assumption (IN1) is also the generalization of the 50% rule in Han [2008] and Kang et al. [2016] in the presence of redundant IVs. Also, Kang et al. [2016] showed that this type of proportion-based assumption is a necessary component for identification of model parameters when instrument validity is uncertain.

Theorem 3 states that in the low dimensional setting, under Assumption (IN1), our new test controls for Type I error in the presence of possibly invalid instruments and the power of our test under H_2 is identical to the power of the DWH test that knows exactly which instruments are valid a priori.

Theorem 3. *Suppose we have models* (2) *and* (3) *with fixed* p *and Assumption* (IN1) *holds. Then, for any* α, $0 < \alpha < 1$, *the asymptotic Type I error* *of Q_E under H_0 is controlled at α, i.e.*

$$\omega \in H_0 : \quad \lim_{n \to \infty} \mathbf{P}\left(|Q_E| \geq z_{\alpha/2}\right) = \alpha.$$

In addition, the asymptotic power of Q_E under $\omega \in H_2$ is

$$\omega \in H_2 : \lim_{n \to \infty} \mathbf{P}\left(|Q_E| \geq z_{\alpha/2}\right) = G\left(\alpha, \frac{\Delta_1 \sqrt{C(\mathcal{V})}}{\sqrt{\left(C(\mathcal{V}) + \frac{1}{p_{\mathrm{z}}}\right)\Sigma_{11}\Sigma_{22}}}\right). \quad (31)$$

Theorem 3 characterizes the behavior of our new endogeneity test Q_E under all the regimes of π that satisfy (IN1). Also, so long as Assumption (IN1) holds, in the fixed p regime, our test Q_E has the same performance as the DWH test that incorporates the information about which instruments are valid or invalid, a priori. In short, our test Q_E is adaptive to the knowledge of IV validity.

We also make a technical note that in the low dimensional case, the assumptions of normal error terms or independence between the error terms and $W_{i.}$, which we made in Section 2.2 when we discussed the models (2) and (3) are not necessary to obtain the same results; Section **??** of the supplementary materials details this general case. We only made these assumptions in the text out of simplicity, especially in the high dimensional regime discussed below where assuming said conditions simplify the technical arguments.

5.2 Valid Instruments in High Dimensions

Next, we consider the high dimensional instruments and covariates setting where p is allowed to be larger than n, but we assume all the instruments are valid, i.e. $\pi = 0$, after conditioning on many covariates; this is a generalization of the setting discussed in Section 3.2 that encompasses the $p > n$ regime. Theorem 2 showed that the DWH test, while it controls Type I error, may have low power, especially when the ratio of p/n is close to 1. Theorem 4 shows that our new test Q_H remedies this deficiency of the DWH test by having proper Type I error control and exhibiting better power.

Theorem 4. *Suppose we have models* (2) *and* (3) *where all the instruments are valid, i.e.* $\pi = 0$*. If* $\sqrt{C(\mathcal{V})} \gg s_{\mathrm{z}1} \log p / \sqrt{n|\mathcal{V}|}$*, and* $\sqrt{s_{\mathrm{z}1}} s \log p / \sqrt{n} \to 0$*, then for any* α*,* $0 < \alpha < 1$*, the asymptotic Type I error of* Q_H *under* H_0 *is controlled at* α

$$\omega \in H_0 : \quad \lim_{n,p \to \infty} \mathbf{P}\left(|Q_H| \geq z_{\alpha/2}\right) = \alpha,$$

and the asymptotic power of Q_H under H_2 is

$$\limsup_{n,p\to\infty} \left| \mathbf{P}\left(|Q_H| \geq z_{\alpha/2}\right) - \mathbf{E}\left(G\left(\alpha, \frac{\Delta_1}{\sqrt{\Theta_{22}^2 \mathrm{V}_1 + \mathrm{V}_2}}\right)\right)\right| = 0, \tag{32}$$

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

In contrast to equation (13) that described the local power of the DWH test in high dimension, the term $\sqrt{1-p/n}$ is absent in the local power of our new endogeneity test in equation (32). Specifically, under H_2, our power is only affected by Δ_1 while the power of the DWH test is affected by $\Delta_1\sqrt{1-p/n}$. This suggests that our power will be better than the power of the DWH test when p is close to the sample size n. In the extreme case when $p/n \to 1$ and constant $C(\mathcal{V})$, the power of the DWH test will be α while the power of our test Q_H will be strictly greater than α. The simulation in Section 6 also numerically illustrates the discrepancies between the power of the two tests. We also stress that in the case $p > n$, our test still has proper size and non-trivial power while the DWH test is not even feasible in this setting.

Finally, like Theorem 2, Theorem 4 controls the growth of the concentration parameter $C(\mathcal{V})$ to be faster than $s_{z1}\log p/\sqrt{n|\mathcal{V}|}$, with a minor discrepancy in the growth rate due to the differences between the set of valid IVs, $\mathcal{V}$, and the set of candidate IVs, $\mathcal{I}$. But, similar to Theorem 2, 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, the growth condition on s, s_{z1}, p, n is related to the growth condition on p and n in Theorem 2 where $p \leq n$. Specifically, the sparsity condition $\sqrt{s_{z1}}s\log p/\sqrt{n} \to 0$ in Theorem 4 holds if $p \leq n$ and $s = O(n^{\delta_0})$ for $0 \leq \delta_0 < 1/3$.

5.3 General Case

We now analyze the properties of our test Q, which can jointly handle invalid instruments as well as high dimensional instruments and covariates, even when $p > n$. We start off by making two additional assumptions that essentially control the asymptotic behavior of relevant and invalid IVs as the dimension of the parameters grows.

(IN2) (Individual IV Strength) Among IVs in $\mathcal{S}$, we have $\min_{j\in\mathcal{S}} |\gamma_j| \geq \delta_{\min} \gg \sqrt{\log p/n}$.

(IN3) (Strong violation) Among IVs in the set $\mathcal{S} \setminus \mathcal{V}$, we have

$$\min_{j \in \mathcal{S} \setminus \mathcal{V}} \left| \frac{\pi_j}{\gamma_j} \right| \geq \frac{12(1+|\beta|)}{\delta_{\min}} \sqrt{\frac{M_1 \log p_{\mathrm{z}}}{n \lambda_{\min}(\Theta)}}. \tag{33}$$

Assumption (IN2) requires individual IV strength to be bounded away from zero. This assumption is needed primarily for cleaner technical exposition and the simulation studies in Section 6 along with additional simulation studies in the supplementary materials demonstrate that (IN2) is largely unnecessary for our test to have proper size and have good power. In the literature, (IN2) is similar to the "beta-min" condition assumption in high dimensional linear regression without IVs [Fan and Li, 2001, Zhao and Yu, 2006, Wainwright, 2007, Bühlmann and Van De Geer, 2011], with the exception that this condition is not imposed on our inferential quantity of interest, Σ_{12}. Next, Assumption (IN3) requires the ratio π_j/γ_j for invalid IVs to be large. Unlike (IN2), this assumption is needed to correctly select valid IVs in the presence of possibly invalid IVs and this sentiment is echoed in the model selection literature by Leeb and Pötscher [2005] who pointed out that "in general no model selector can be uniformly consistent for the most parsimonious true model" and hence the post-model-selection inference is generally non-uniform (or uniform within a limited class of models). Specifically, for any IV with a small, but non-zero $|\pi_j/\gamma_j|$, such a weakly invalid IV is hard to distinguish from valid IVs where $\pi_j/\gamma_j = 0$. If a weakly invalid IV is mistakenly declared as valid, the bias from this mistake is of the order $\sqrt{\log p_{\mathrm{z}}/n}$, which has consequences, not for consistency of the point estimation of Σ_{12}, but for a $\sqrt{n}$ inference of Σ_{12}; see the detailed discussion in Proposition **??** about point estimation of Σ_{12} in Section **??** of the supplementary materials.

Overall, Assumptions (IN1)-(IN3) allow detection of each valid IV with the observed data in order to obtain a good estimate of $\mathcal{V}$. Consequently, if all the instruments are valid, like the setting described in Section 5.2, we do not need Assumptions (IN1)-(IN3) to make any claims about our endogeneity test. However, in the presence of potentially invalid IVs that grow in dimension, these type of assumptions are needed to control the behavior of the invalid IVs asymptotically. Finally, Theorem 5 characterizes the asymptotic behavior of Q under these conditions.

Theorem 5. *Suppose we have models* (2) *and* (3) *where some instruments may be invalid, i.e.* $\pi \neq 0$, *and Assumptions* (IN1)-(IN3) *hold. If* $\sqrt{C(\mathcal{V})} \gg s_{\mathrm{z1}} \log p / \sqrt{n|\mathcal{V}|}$, *and* $\sqrt{s_{\mathrm{z1}}} s \log p / \sqrt{n} \to 0$, *then for any* α, $0 < \alpha < 1$, *the*

Type I error of Q under H_0 is controlled at α

$$\omega \in H_0 : \quad \lim_{n,p \to \infty} \mathbf{P}\left(|Q| \geq z_{\alpha/2}\right) = \alpha. \tag{34}$$

and the asymptotic power of Q under H_2 is in equation (32)*.*

Theorem 5 shows that our new test Q controls Type I error and has power that's similar to the power in Theorem 4 where we know about all the instruments' validity once we conditioned on many covariates. Indeed, similar to the property of Q_E in Theorem 3 and how it was adaptive to the knowledge about instrument validity in low dimension, Q also has this property, but in high dimensions.

6 Simulation

6.1 Setup

We conduct a simulation study to investigate (i) the performance of our new endogeneity test and DWH test and (ii) the sensitivity of our new test to violations of the regularity assumptions required for theoretical analysis, most notably (IN2) and (IN3). Here, we only show the results regarding the performance of the test and summarize the results about the sensitivity to regularity assumptions; see Section **??** of the supplementary materials for all the details. For the low dimensional case, we generate data from models (2) and (3) in Section 2.2 with $p_{\mathrm{z}} = 9$ instruments and $p_{\mathrm{x}} = 5$ covariates. The vector $\mathbf{W}_{i.}$ is a multivariate normal with mean zero and covariance $\Lambda_{ij} = 0.5^{|i-j|}$ for $1 \leq i, j \leq p$. The parameters of the models are: $\beta = 1$, $\phi = (0.6, 0.7, 0.8, 0.9, 1.0) \in \mathbb{R}^5$ and $\psi = (1.1, 1.2, 1.3, 1.4, 1.5) \in \mathbb{R}^5$. For the high dimensional case, we use the same models as the low dimensional case except $p_{\mathrm{z}} = 100$, $p_{\mathrm{x}} = 150$, $\phi = (0.6, 0.7, 0.8, \cdots, 1.5, 0, 0, \cdots, 0) \in \mathbb{R}^{p_{\mathrm{x}}}$ so that $s_{\mathrm{x}1} = 10$, and $\psi = (1.1, 1.2, 1.3, \cdots, 2.0, 0, 0, \cdots, 0) \in \mathbb{R}^{p_{\mathrm{x}}}$ so that $s_{\mathrm{x}2} = 10$. For both high dimensional and low dimensional cases, the relevant instruments are $\mathcal{S} = \{1, \ldots, 7\}$ and the instruments that are valid and relevant are $\mathcal{V} = \{1, 2, 3, 4, 5\}$; thus instruments 6 and 7 are relevant, but not valid. 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 sample size n, the endogeneity level via $\mathrm{Cov}(\delta_i, \epsilon_i)$, IV strength via γ, and IV validity via π. For sample size, we let $n = (300, 1000, 10000, 50000)$. 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 $\gamma_{\mathcal{V}} = K\,(1,1,1,1,\rho_1)$ and $\gamma_{\mathcal{S}^*\setminus\mathcal{V}} = K\,(1,1)$ and $\gamma_{(\mathcal{S}^*)^C} = 0$, where K is varied as a function of the concentration parameter (see below) and ρ_1 is varied based on the vector $(0, 0.1, 0.2)$ across simulations. Note that the value K controls the global strength of instruments, with higher $|K|$ indicating strong instruments in a global sense. Also, the value ρ_1 controls the relative individual strength of instruments, specifically between the first four instruments in $\mathcal{V}$ and the fifth instrument. For example, $\rho_1 = 0.2$ implies that the fifth IV's individual strength is only 20% of the other four valid instruments, i.e IVs 1 to 4. Also, varying ρ_1 would simulate the adherence to regularity assumption (IN2).

We specify K as follows. Suppose we set n at a baseline of 100 given the simulation parameters $\mathcal{V}, \rho_1$, Λ and Σ_{22}. For each value of $nC(\mathcal{V})$, say $nC(\mathcal{V}) = 25$, we find K that satisfies it. Here, the $nC(\mathcal{V})$ mimics the expected partial F-statistic one would obtain from doing the F-test for the null $\gamma_{\mathcal{V}} = 0$ from an OLS regression between D and W for a given sample size n. We vary $nC(\mathcal{V})$ from 25 to 150, specifying K for each value of $nC(\mathcal{V})$.

Finally, we vary π, which controls the validity of the IVs, by defining $\pi_j = \rho_2\gamma_j$ for $j = 6, 7$ and $\pi_j = 0$ for all other j so that ρ_2 controls the magnitude of IV invalidity from the 6th and 7th instruments. In the ideal case, we would have $\rho_2 = 0$ so that $\mathcal{S} = \mathcal{V} = \{1, 2, 3, 4, 5, 6, 7\}$. But, $\rho_2 \neq 0$ implies that the last two instruments are not valid and we vary ρ_2 based on the vector $(0, 1, 2)$. Note that varying ρ_2 would simulate the adherence to regularity assumption (IN3).

In summary, we vary the endogeneity level, the sample size n, IV strength, and IV validity in our simulation study, with ρ_1 and ρ_2 simulating the adherence to the new regularity assumptions in the paper, (IN2) and (IN3), respectively. Since the DWH test cannot be used when $n \leq p$, we restrict our simulation study to the case where $p < n$ for comparison purposes. In particular, we compare the power of our testing procedure to the DWH test and the oracle DWH test where an oracle provides us with knowledge about valid IVs, i.e. $\mathcal{V}$, which will not occur in practice.

6.2 Results

We present the result for $nC(\mathcal{V}) = 25$, which is representative of the results from the simulation study; for reference, all the results of the simulation study are in Section **??** of the supplementary materials. First, Figure 1 considers the power of three comparators, the proposed testing procedure Q_E, the regular DWH test and the oracle DWH test, under the low dimensional setting with $n = 1000, p_{\mathrm{x}} = 9$, and $p_{\mathrm{z}} = 5$. Columns "Strong","WeakIV1", and “WeakIV2” represent different values of ρ_1, specifically $\rho_1 = 0$, $\rho_1 = 0.1$, and $\rho_1 = 0.2$ respectively. Rows “Valid” and “Invalid” represent different values of ρ_2, specifically $\rho_2 = 0$ where all the instruments are valid and $\rho_2 = 2$ where the 6th and 7th instruments are invalid, respectively. The x-axis represents different values of endogeneity scaled by the error variances, i.e. $\Sigma_{12}/\sqrt{\Sigma_{11}\Sigma_{22}}$, and the y-axis is the empirical proportion of rejecting the null hypothesis H_0 over 500 simulations and is an approximation to the test's power.

When all the instruments are valid (i.e. the first row of Figure 1), the empirical power curves of the proposed test, the regular DWH test and the oracle DWH are identical. However, when some of the instruments are invalid (i.e. the second row of Figure 1), the regular DWH test cannot control Type I error and the power curve is shifted, as expected from Theorem 1. In contrast, the power curve of our proposed test is nearly identical to that of the oracle DWH test, which is expected based on our theoretical analysis in Theorem 3. In all cases, our test maintains Type I error control.

Second, Figure 2 considers the power of our test Q, the regular DWH test, and the oracle DWH test in the high dimensional setting with $n = 300, p_{\rm x} = 150$, and $p_{\rm z} = 100$. Again, when all the instruments are valid (i.e. the first row of Figure 2), all the three tests have proper size. However, as suggested by Theorem 2, the regular DWH test suffers from low power compared to our test Q. When some instruments are invalid in the high dimensional setting (i.e. the second row of Figure 2), the DWH test cannot controls Type I error while our proposed test not only control Type I error, but also has power that is close to the oracle DWH test which knows a priori which instruments are valid.

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 the presence of invalid instruments and high dimensional covariates. In the low dimensional setting, the power of our endogeneity test is identical to the power of the oracle DWH test while in the high dimensional setting, our endogeneity test has better power than the regular DWH test and has near-optimal performance with respect to the oracle.

Finally, we provide a brief summary of the simulation results reported in the supplementary materials, especially those concerning the violation of regularity assumptions (IN2) and (IN3); see Section **??** of the supplementary materials for details. First, in Figures **??** - **??** of the supplementary materials, our proposed testing procedure performs similarly to Figures 1 and 2 for different individual IV strengths (represented by different values of ρ_1), reiterating our comment in Section 5.3 that the proposed estimator is not sensitive to the individual instrument strength assumption (IN2) required in the theoretical analysis. However, as discussed in Section 5.3, in high dimensions, if the instruments are weakly invalid and consequently, violate (IN3), our procedure tends to suffer. In particular, in Figures **??** and **??** of the supplementary materials, when the invalid IVs weakly violated exogeneity with $\rho_2 = 1$, our test's size exceeds α, although at most by $5\% \sim 10\%$. But the power curve of the proposed test is still much better than that of DWH test assuming valid IVs after conditioning, which tends to be shifted away from the null value $\Sigma_{12} = 0$.

7 Conclusion

In this paper, we showed that the DWH test can be highly misleading with respect to Type I error in the presence of invalid instruments and have relatively low power in high dimensional settings. We propose an improved endogeneity test to remedy these failures of the DWH test and we show that our test has proper Type I error control in the presence of invalid IVs and has much better power than the DWH test in high dimensional settings.

References

J. Durbin. Errors in variables. *Review of the International Statistical Institute*, 22:23–32, 1954.

D. M. Wu. Alternative tests of independence between stochastic regressors and disturbances. *Econometrica*, 41:733–750, 1973.

J. Hausman. Specification tests in econometrics. *Econometrica*, 41:1251–1271, 1978.

R. Davidson and J. G. MacKinnon. *Estimation and Inference in Econometrics.* Oxford University Press, New York, 1993.

Paul W. Holland. Causal inference, path analysis, and recursive structural equations models. *Sociological Methodology*, 18(1):449–484, 1988.

Guido W Imbens and Joshua D Angrist. Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475, 1994.

Joshua D. Angrist, Guido W. Imbens, and Donald B. Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455, 1996.

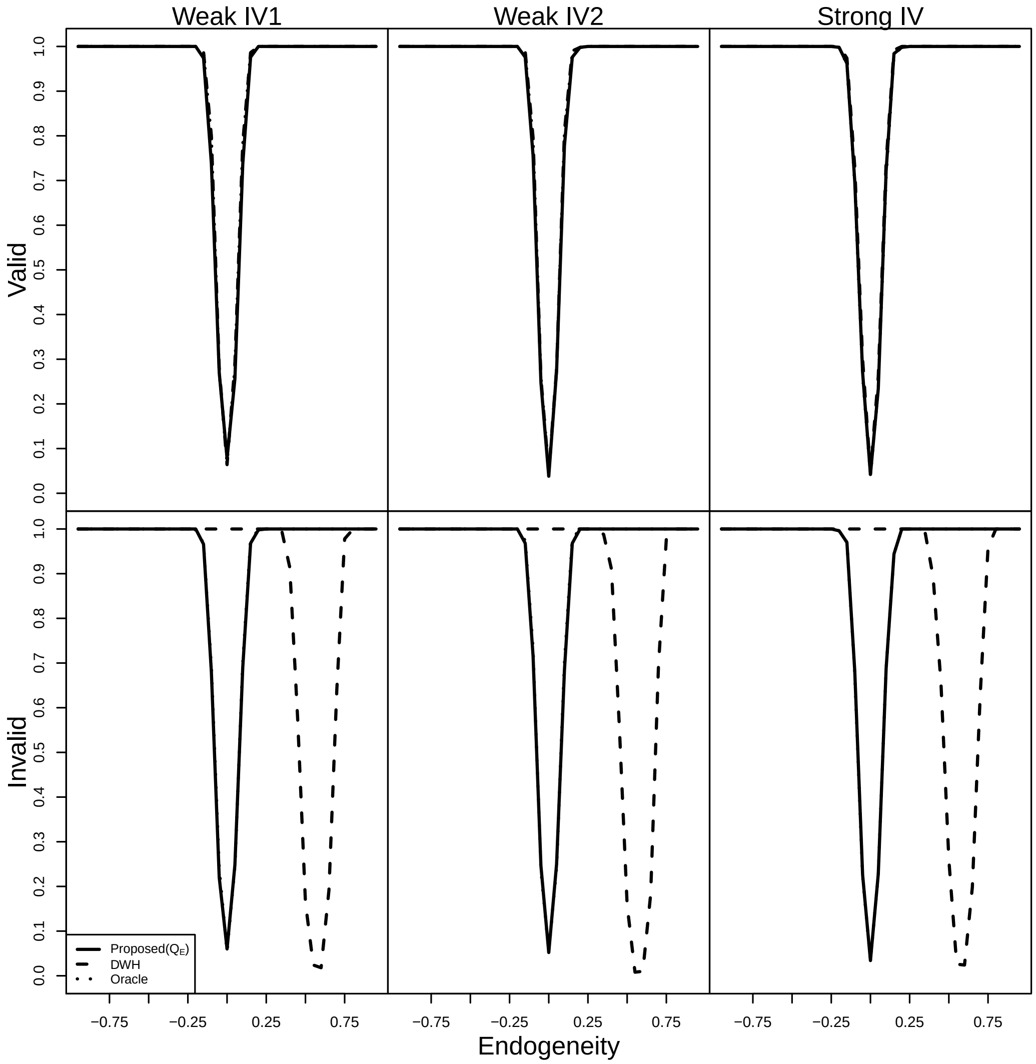

Figure 1: Power of endogeneity tests when $n = 1000$, $p_{\mathrm{x}} = 5$ and $p_{\mathrm{z}} = 9$. The x-axis represents the endogeneity $\frac{\Sigma_{12}}{\sqrt{\Sigma_{11}\Sigma_{22}}}$ and the y-axis represents the empirical power over 500 simulations. Each line represents a particular test's empirical power over various values of the endogeneity. The columns "Weak IV1", "Weak IV2", and "Strong IV" represent the cases when $\rho_1 = 0.1$, $\rho_1 = 0.2$, and $\rho_1 = 0$. The rows "Valid" and "Invalid" represent the cases when $\rho_2 = 0$ and $\rho_2 = 2$.

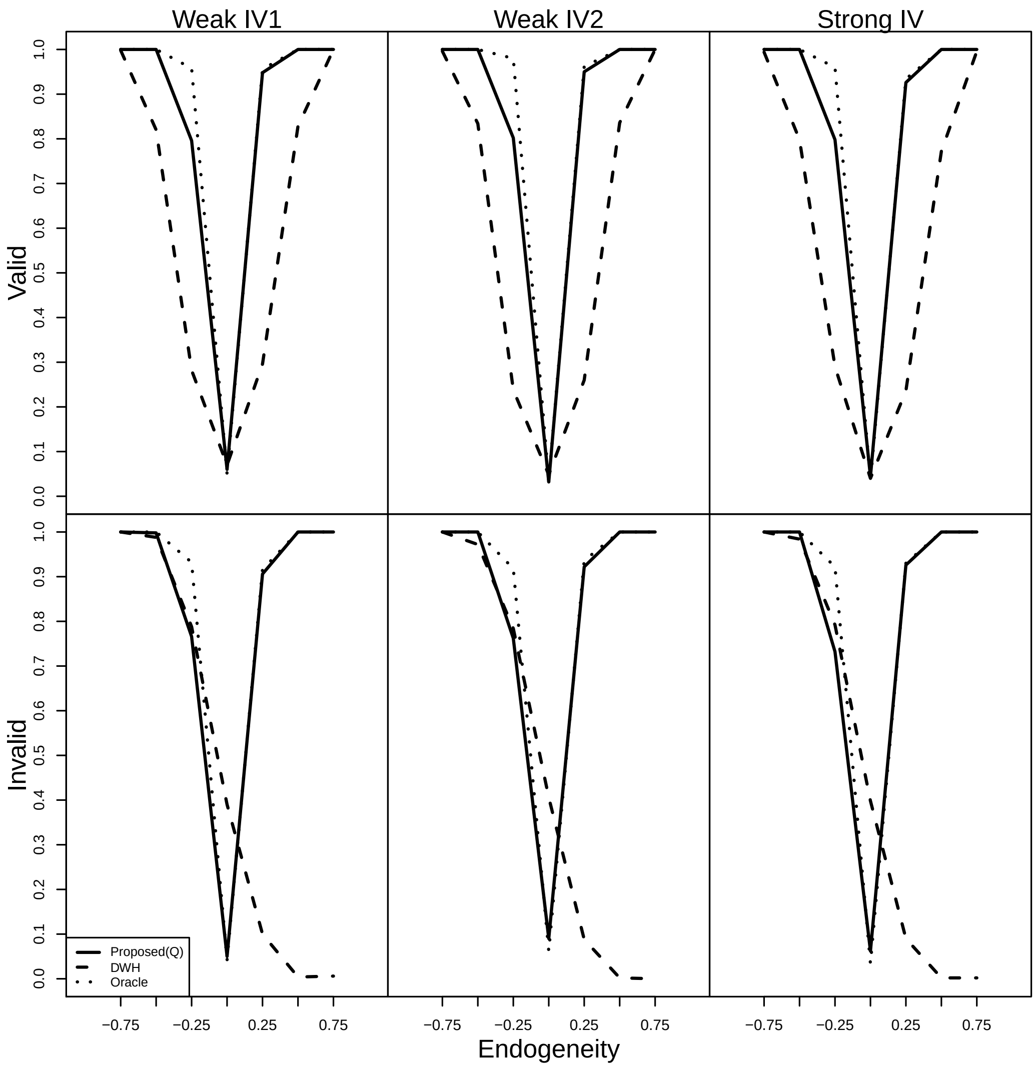

Figure 2: Power of endogeneity tests when $n = 300$, $p_{\mathrm{x}} = 150$ and $p_{\mathrm{z}} = 100$. The x-axis represents the endogeneity $\frac{\Sigma_{12}}{\sqrt{\Sigma_{11}\Sigma_{22}}}$ and the y-axis represents the empirical power over 500 simulations. Each line represents a particular test's empirical power over various values of the endogeneity. The columns "Weak IV1", "Weak IV2", and "Strong IV" represent the cases when $\rho_1 = 0.1$, $\rho_1 = 0.2$, and $\rho_1 = 0$. The rows "Valid" and "Invalid" represent the cases when $\rho_2 = 0$ and $\rho_2 = 2$.

Michael P. Murray. Avoiding invalid instruments and coping with weak instruments. *The Journal of Economic Perspectives*, 20(4):111–132, 2006.

Douglas Staiger and James H. Stock. Instrumental variables regression with weak instruments. *Econometrica*, 65(3):557–586, 1997.

Timothy G Conley, Christian B Hansen, and Peter E Rossi. Plausibly exogenous. *Review of Economics and Statistics*, 94(1):260–272, 2012.

Miguel A. Hernán and James M. Robins. Instruments for causal inference: An epidemiologist's dream? *Epidemiology*, 17(4):360–372, 2006.

Michael Baiocchi, Jing Cheng, and Dylan S. Small. Instrumental variable methods for causal inference. *Statistics in Medicine*, 33(13):2297–2340, 2014.

Eric Gautier and Alexandre B. Tsybakov. High-dimensional instrumental variables regression and confidence sets. *arXiv preprint arXiv:1105.2454*, 2011.

A. Belloni, D. Chen, V. Chernozhukov, and C. Hansen. Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2429, 2012.

Victor Chernozhukov, Christian Hansen, and Martin Spindler. Post-selection and post-regularization inference in linear models with many controls and instruments. *The American Economic Review*, 105(5):486–490, 2015.

C. R. Nelson and R. Startz. Some further results on the exact sample properties of the instrumental variables estimator. *Econometrica*, 58:967–976, 1990.

Paul A Bekker. Alternative approximations to the distributions of instrumental variable estimators. *Econometrica: Journal of the Econometric Society*, pages 657–681, 1994.

John Bound, David A. Jaeger, and Regina M. Baker. Problems with instrumental variables estimation when the correlation between the instruments and the endogeneous explanatory variable is weak. *Journal of the American Statistical Association*, 90(430):443–450, 1995.

Jean-Marie Dufour. Some impossibility theorems in econometrics with applications to structural and dynamic models. *Econometrica*, pages 1365–1387, 1997.

Eric Zivot, Richard Startz, and Charles R Nelson. Valid confidence intervals and inference in the presence of weak instruments. *International Economic Review*, pages 1119–1144, 1998.

J. Wang and E. Zivot. Inference on structural parameters in instrumental variables regression with weak instruments. *Econometrica*, 66(6):1389–1404, 1998.

Frank Kleibergen. Pivotal statistics for testing structural parameters in instrumental variables regression. *Econometrica*, 70(5):1781–1803, 2002.

Marcelo J. Moreira. A conditional likelihood ratio test for structural models. *Econometrica*, 71(4):1027–1048, 2003.

John C. Chao and Norman R. Swanson. Consistent estimation with a large number of weak instruments. *Econometrica*, 73(5):1673–1692, 2005.

Donald W. K. Andrews, Marcelo J. Moreira, and James H. Stock. Performance of conditional wald tests in {IV} regression with weak instruments. *Journal of Econometrics*, 139(1):116–132, 2007.

Alice Nakamura and Masao Nakamura. On the relationships among several specification error tests presented by durbin, wu, and hausman. *Econometrica: journal of the Econometric Society*, pages 1583–1588, 1981.

Firmin Doko Tchatoka. On bootstrap validity for specification tests with weak instruments. *The Econometrics Journal*, 18(1):137–146, 2015.

Jinyong Hahn and Jerry Hausman. A new specification test for the validity of instrumental variables. *Econometrica*, 70(1):163–189, 2002.

John C Chao, Jerry A Hausman, Whitney K Newey, Norman R Swanson, and Tiemen Woutersen. Testing overidentifying restrictions with many instruments and heteroskedasticity. *Journal of Econometrics*, 178:15–21, 2014.

Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference for high-dimensional sparse econometric models. *arXiv preprint arXiv:1201.0220*, 2011a.

Alexandre Belloni, Victor Chernozhukov, Ivan Fernández-Val, and Chris Hansen. Program evaluation with high-dimensional data. *arXiv preprint arXiv:1311.2645*, 2013.

Jianqing Fan and Yuan Liao. Endogeneity in high dimensions. *Annals of statistics*, 42(3):872, 2014.

Victor Chernozhukov, Christian Hansen, and Martin Spindler. Valid post-selection and post-regularization inference: An elementary, general approach. 2014.

Franklin M. Fisher. The relative sensitivity to specification error of different k-class estimators. *Journal of the American Statistical Association*, 61: 345–356, 1966.

Franklin M. Fisher. Approximate specification and the choice of a k-class estimator. *Journal of the American Statistical Association*, 62:1265–1276, 1967.

Whitney K. Newey. Generalized method of moments specification testing. *Journal of Econometrics*, 29(3):229 – 256, 1985.

Jinyong Hahn and Jerry Hausman. Estimation with valid and invalid instruments. *Annales d'conomie et de Statistique*, 79/80:25–57, 2005.

Patrik Guggenberger. On the asymptotic size distortion of tests when instruments locally violate the exogeneity assumption. *Econometric Theory*, 28(2):387–421, 2012.

Daniel Berkowitz, Mehmet Caner, and Ying Fang. The validity of instruments revisited. *Journal of Econometrics*, 166(2):255–266, 2012.

Mehmet Caner. Near exogeneity and weak identification in generalized empirical likelihood estimators: Many moment asymptotics. *Journal of Econometrics*, 182(2):247 – 268, 2014.

Donald W. K. Andrews. Consistent moment selection procedures for generalized method of moments estimation. *Econometrica*, 67(3):543–563, 1999.

Donald WK Andrews and Biao Lu. Consistent model and moment selection procedures for gmm estimation with application to dynamic panel data models. *Journal of Econometrics*, 101(1):123–164, 2001.

Dylan S. Small. Sensitivity analysis for instrumental variables regression with overidentifying restrictions. *Journal of the American Statistical Association*, 102(479):1049–1058, 2007.

Zhipeng Liao. Adaptive gmm shrinkage estimation with consistent moment selection. *Econometric Theory*, 29(05):857–904, 2013.

Xu Cheng and Zhipeng Liao. Select the valid and relevant moments: An information-based lasso for gmm with many moments. *Journal of Econometrics*, 186(2):443–464, 2015.

Hyunseung Kang, Anru Zhang, T Tony Cai, and Dylan S Small. Instrumental variables estimation with some invalid instruments and its application to mendelian randomization. *Journal of the American Statistical Association*, 111:132–144, 2016.

Michal Kolesár, Raj Chetty, John N. Friedman, Edward L. Glaeser, and Guido W. Imbens. Identification and inference with many invalid instruments. *Journal of Business & Economic Statistics*, 33(4):474–484, 2015.

Trevor Breusch, Hailong Qian, Peter Schmidt, and Donald Wyhowski. Redundancy of moment conditions. *Journal of Econometrics*, 91(1):89 – 111, 1999.

Alastair Hall and Fernanda P. M. Peixe. A consistent method for the selection of relevant instruments. *Econometric Reviews*, 22(3):269–287, 2003.

Zijian Guo, Hyunseung Kang, T Tony Cai, and S Dylan Small. Confidence interval for causal effects with possibly invalid instruments even after controlling for many confounders. 2016.

Roberto S Mariano. Approximations to the distribution functions of theil's k-class estimators. *Econometrica: Journal of the Econometric Society*, pages 715–721, 1973.

James H Stock and Jonathan H Wright. Gmm with weak identification. *Econometrica*, pages 1055–1096, 2000.

Jeffrey M. Wooldridge. *Econometric Analysis of Cross Section and Panel Data*. MIT press, 2nd ed. edition, 2010.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, 1996.

Cun-Hui Zhang and Stephanie S Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014.

Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909, 2014.

Sara van de Geer, Peter Bühlmann, Yaacov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.

T Tony Cai and Zijian Guo. Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *The Annals of statistics*, To appear, 2016.

Alexandre Belloni, Victor Chernozhukov, and Lie Wang. Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98 (4):791–806, 2011b.

Tingni Sun and Cun-Hui Zhang. Scaled sparse linear regression. *Biometrika*, 101(2):269–284, 2012.

Zhao Ren, Tingni Sun, Cun-Hui Zhang, and Harrison H Zhou. Asymptotic normality and optimalities in estimation of large gaussian graphical model. *arXiv preprint arXiv:1309.6024*, 2013.

Chirok Han. Detecting invalid instruments using l 1-gmm. *Economics Letters*, 101(3):285–287, 2008.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

Peng Zhao and Bin Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7(Nov):2541–2563, 2006.

Martin Wainwright. Information-theoretic bounds on sparsity recovery in the high-dimensional and noisy setting. In *2007 IEEE International Symposium on Information Theory*, pages 961–965. IEEE, 2007.

Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications.* Springer Science & Business Media, 2011.

Hannes Leeb and Benedikt M Pötscher. Model selection and inference: Facts and fiction. *Econometric Theory*, 21(01):21–59, 2005.