An Identification-and Dimensionality-Robust Test for Instrumental Variables Models

Manu Navjeevan*
Texas A&M University

Abstract. Using novel modifications of Lindeberg's interpolation technique, I propose a new identification-robust test for the structural parameter in a heteroskedastic instrumental variables model. While I allow the number of instruments to be large, my analysis does not require this fact, making my test applicable in many settings that have not been well studied. Instead, the proposed test statistic has a limiting chi-squared distribution so long as an auxiliary parameter can be consistently estimated. This is possible using machine learning methods even when the number of instruments is much larger than the sample size. To improve power, a simple combination with the sup-score statistic of Belloni et al. (2012) is proposed. I point out that first-stage F-statistics calculated on LASSO selected variables may be misleading indicators of identification strength and investigate the performance of my proposed methods in both empirical data and simulation study.

Keywords: Instrumental Variables, Weak Identification, High-Dimensional
JEL Codes: C12, C36, C55

1. Introduction

Consider a linear instrumental variables (IV) model

$$y_i = x_i'\beta + z_{1i}'\Gamma + \epsilon_i, \quad \mathbb{E}[\epsilon_i|z_i] = 0 \tag{1.1}$$

where $y_i \in \mathbb{R}$ is an outcome of interest and $x_i \in \mathbb{R}^{d_x}$ is a vector of endogenous variables that may be correlated with the structural error $\epsilon_i \in \mathbb{R}$. The variable $z_i = (z_{1i}, z_{2i})' \in \mathbb{R}^{d_c} \times \mathbb{R}^{d_z}$ represents a vector of instrumental variables, of which a subvector of fixed dimension, $z_{1i}' \in \mathbb{R}^{d_c}$, is included in the structural equation (1.1) as exogenous control. I assume that the researcher has access to n independent observations of $(y_i, x_i', z_i')'$. In this setting, I propose a new test for a two-sided restriction on the structural parameter; $H_0 : \beta = \beta_0$ versus $H_1 : \beta \neq \beta_0$. The proposed test has exact asymptotic size even when instruments are potentially high-dimensional ($d_z \gg n$) and arbitrarily weak.

When instruments are suspected to be weak, researchers may want to test hypotheses about structural parameters using testing procedures that are robust to identification strength. These procedures all rely on some conditions on the rate of growth of the number of instruments, d_z, in relation to the sample size, n. The initial identification robust tests developed in Staiger and Stock (1997), Moreira (2003), and Kleibergen (2005) are shown by Andrews and Stock (2007) to control size under heteroskedasticity when the number of instruments cubed is small relative to the sample size, $d_z^3/n \to 0$. Meanwhile, recent and interesting "many-instrument" tests (Crudu et al. (2021), Mikusheva and Sun (2021), Matsushita and Otsu (2022), Lim et al. (2022)) allow the number of instruments to be proportional to the sample size, $d_z/n \to \varrho \in [0, 1)$, but require that the number of instruments itself be large, $d_z \to \infty$.

*Email: mnavjeevan@ucla.edu. Revised July 8, 2024. The latest version of this paper can be found here. I thank Denis Chetverikov for his guidance, numerous discussions and permanent support. I am also grateful to the other members of my dissertation committee, Andres Santos and Zhipeng Liao, and participants in UCLA's econometrics proseminar, Jinyong Hahn, Rosa Matzkin, Shuyang Sheng, and Daniel Ober-Reynolds for a half-decade of useful comments.

In practice, these conditions can be difficult to interpret and the variety of tests available under alternate regimes may make it difficult for the researcher to know which test, if any, should be applied in her exact setting. As examples, consider settings such as that of Derenoncourt (2022), where $d_z = 9$ and $n = 239$, Paravisini et al. (2014), where $d_z = 10$ and $n = 5{,}995$, and Gilchrist and Sands (2016) where $d_z = 52$ and $n = 1{,}671$. In all three cases, the number of instruments cubed, $d_z^3 = 729$, 1,000 and 140,608, respectively, cannot be treated as negligible relative to the sample size. Indeed, all of these papers use post-LASSO estimates of the first-stage, suggesting concern about the large number of instruments by the authors. However, as the number of instruments is only moderately large in all of these settings, asymptotic approximations that rely on $d_z \to \infty$ may not accurately resemble the finite sample distribution of the test statistic. This can make size control of many-instrument tests questionable.

By comparison, the test proposed in this paper has correct asymptotic size so long as an auxiliary "conditional slope" parameter can be consistently estimated. This is possible using standard methods when the number of instruments is small or with machine learning methods even when the number of instruments is much larger than the sample size, $d_z \gg n$, a setting that has received limited attention until this point. Existing identification-robust tests that allow $d_z \gg n$ (Belloni et al. (2012), Gautier and Rose (2021), Mikusheva (2023)) either fail to incorporate first-stage information or rely on sample splitting, both of which may reduce power in overidentified models. In practice, I suggest an ℓ_1-penalized estimator of the conditional slope parameter which can work in both low- and high-dimensional settings and is easily implementable using out-of-the-box methods. This estimator has the additional benefit of being trivially consistent when first- and second-stage errors are homoskedastic.

To construct the test statistic I borrow an idea from Kleibergen (2002, 2005) and leverage the conditional slope parameter to partial out the structural error from first-stage jackknife-ridge estimates.[1] Combining the jackknife and partialling out approaches leaves the researcher with first-stage estimates that are uncorrelated not only with an observations' own structural error but also from the structural error of other observations. If observations are jointly Gaussian this could be then used, along with the fact that uncorrelated jointly Gaussian are independent, to condition on these first-stage estimates and show the test statistic has a chi-squared distribution under the null hypothesis.

When the number of instruments is small, treating all observations as if they were jointly Gaussian can be justified in large samples through central limit and continuous mapping theorems. However, when the number of instruments is large these standard tools cannot be applied. Instead, my asymptotic analysis uses modifications of Lindeberg's interpolation argument (Lindeberg, 1922) to directly show that the distribution of the test statistic in the general model can be approximated by the distribution of the test statistic in a Gaussian model. The modifications of Lindeberg's original argument are required to deal with a "divide-by-zero" problem that arises under weak identification. When there is a single endogeneous variable analysis can be simplified by taking advantage of the particular form of the test statistic, while in the general case a more involved argument is needed which relies on stronger moment conditions. These modified approaches may be of independent interest to a growing literature on direct Gaussian approximation techniques (Chatterjee (2006), Chernozhukov et al. (2013), Pouzo (2015), Celentano et al. (2020)).

Through the Gaussian approximation result, I examine the power properties of my proposed testing procedure in local neighborhoods of the null. These local neighborhoods are characterized by a bounded local power index. In the case of a single endogenous variable I show that, under an additional regularity condition, the local power index diverging implies that the test

[1] Interestingly, the asymptotic analysis does not require the first-stage jackknife-ridge estimates to be consistent, allowing some flexibility if the researcher wanted to use an alternative method to construct first-stage estimates.

is consistent. Unfortunately, the process of partialling out the structural error can introduce bias into the first-stage estimate under the alternative hypothesis. Against certain alternatives, this bias can be particularly pronounced and erase the first-stage signal from the instruments, a problem pointed out by Moreira (2001), Andrews et al. (2006), and Andrews (2016) in the context of the original K-statistic. To address this, I propose a simple combination with the sup-score test of Belloni et al. (2012), which does not incorporate first-stage information but does not face a power decline against particular alternatives.

Identification-robust testing procedures may be of particular interest in high-dimensional settings due to a lack of clarity on how to pretest for weak identification. Current empirical practice when using post-LASSO estimates of the first-stage appears to be to use standard t-test inference if the first stage F-statistic on the LASSO selected variables is larger than 10 (Paravisini et al. (2014), Gilchrist and Sands (2016), Derenoncourt (2022)). Using a simple numerical demonstration, I argue that first stage F-statistics on LASSO selected variables may not be reliable indicators of identification strength. Given uncertainty about the strength of identification I apply the newly proposed testing procedures to the data of Gilchrist and Sands (2016) and generate weak instrument-robust confidence intervals for the effect of social spillovers on movie consumption. The new confidence intervals are larger than those implied by the author's initial analysis but are considerably smaller than those obtained by inverting the sup-score test.

Finally, I examine the applicability of the theoretical results in this paper through a simulation study. While existing tests seem to face size distortions in alternate regimes, the test based on my proposed test statistic has nearly exact size in a variety of settings. While this new test may have diminished power against certain alternatives, this deficiency is ameliorated through combination with the sup-score test. Compared to the many-instrument tests of Mikusheva and Sun (2021) and Matsushita and Otsu (2022) and the sup-score test of Belloni et al. (2012), the tests proposed in this paper also appear to have favorable power properties, particularly when the instruments are highly correlated.

The outline of this paper is as follows. Section 2 defines the model and the test statistic, Section 3 provides results for a single endogenous variable. Section 4 examines the power properties of the test and proposes a combination to improve power. Section 5 extends the analysis to the general case, Section 6 contains the empirical application, and Section 7 provides evidence from simulation study.

2. Model and Setup

Though the analysis below allows for exogenous regressors, to simplify the exposition I follow Mikusheva and Sun (2021) and assume that they have already been partialed out of both the outcome, y_i, and the endogenous regressors, x_i. As the controls are assumed to be of fixed dimension, this is without loss of generality. [1] The IV model considered can be written as a system of simultaneous equations:

$$\begin{aligned} y_i &= x_i'\beta + \varepsilon_i \\ x_i &= \Pi_i + v_i \end{aligned} \tag{2.1}$$

The researcher observes the outcome $y_i \in \mathbb{R}$, the endogenous variable $x_i \in \mathbb{R}^{d_x}$, and the instruments $z_i \in \mathbb{R}^{d_z}$ but neither the structural error $\varepsilon_i \in \mathbb{R}$ nor the first-stage errors $v_i \in \mathbb{R}^{d_x}$. The structural error is assumed to be conditional-mean independent of the instruments, $\mathbb{E}[\varepsilon_i|z_i] = 0$. I denote $\mathbb{E}[x_i|z_i]$ as $\Pi_i := \mathbb{E}[x_i|z_i]$ and make no assumptions about the functional form of the conditional expectation so the instruments are allowed to affect the endogenous

[1] For discussion refer to Appendix E in Navjeevan (2023).

variable in a nonlinear fashion.

The random variables $\{(z_i, \varepsilon_i, v_i)\}_{i=1}^n$ are assumed to be independent across observations. Observations need not be identically distributed but the errors are assumed to have a common covariance structure conditional on the instruments z_i:

$$\text{Var}((\varepsilon_i, v_i)'|z_i) := \Omega(z_i) = \begin{pmatrix} \sigma^2_{\epsilon\epsilon}(z_i) & \Sigma_{v\epsilon}(z_i) \\ \Sigma_{\epsilon v}(z_i) & \Sigma_{vv}(z_i) \end{pmatrix} \in \mathbb{R}^{(1+d_x)\times(1+d_x)}$$

As $\Omega(z_i)$ is otherwise left unrestricted, the errors are allowed to be heteroskedastic. All results in this paper hold conditionally on a realization of the instruments $\boldsymbol{z} := (z_1', \ldots, z_n') \in \mathbb{R}^{n\times d_z}$ so from this point forth they are treated as fixed and all expectations can be understood as conditional on the instruments.

Under this setup, the researcher wishes to test a two-sided restriction on the structural parameter:

$$H_0 : \beta = \beta_0 \text{ vs. } H_1 : \beta \neq \beta_0$$

I am interested in constructing powerful tests for this null-alternate pair that are asymptotically valid under arbitrarily weak identification and with minimal restrictions on the number of instruments d_z. To this end, define the null errors $\varepsilon_i(\beta_0) := y_i - x_i'\beta_0$. Using these, I construct a variable, r_i, that is a "partialed-out" version of the endogenous variable satisfying $\text{Cov}(r_i, \epsilon_i(\beta_0)) = 0$:

$$\begin{aligned} r_i := x_i - \rho(z_i)\epsilon_i(\beta_0), \quad \rho(z_i) &:= \frac{\text{Cov}(\epsilon_i(\beta_0), x_i)}{\text{Var}(\epsilon_i(\beta_0))} \in \mathbb{R}^{d_x} \\ &= \frac{\Sigma_{v\epsilon}(z_i) + \Sigma_{vv}(z_i)(\beta - \beta_0)}{(1, \beta - \beta_0)'\Omega(z_i)'(1, \beta - \beta_0)}. \end{aligned}$$

Each element of the nuisance parameter $\rho(z_i)$, $\rho_\ell(z_i)$ for $\ell = 1, \ldots, d_x$, can be interpreted as the (conditional) slope coefficient from a simple linear regression of $x_{\ell i}$ on $\epsilon_i(\beta_0)$. Thus, if $\rho_\ell(\cdot)$ falls in some function class Φ it can be estimated directly under H_0 by solving empirical analogs of:[2]

$$\rho_\ell(z_i) = \arg\min_{\varphi\in\Phi} \bar{\mathbb{E}}[(x_{\ell i} - \epsilon_i(\beta_0)\varphi(z_i))^2].$$

Though other estimators of the conditional slope parameter are possible, I propose an ℓ_1-penalized estimate which is consistent under the assumption that $\rho_\ell(z_i)$ has an approximately sparse representation in some (growing) basis $b(z_i) := (b_1(z_i), \ldots, b_{d_b}(z_i))' \in \mathbb{R}^{d_b}$. That is, $\rho_\ell(z_i) = b(z_i)'\phi_\ell + \xi_{\ell i}$ where $\xi_{\ell i}$ represents an approximation error that tends to zero with the sample size and ϕ_ℓ is sparse in the sense that many of its coefficients are zero. Under homoskedasticity, $\rho_\ell(z_i)$ is a constant function and thus has a spare representation in any basis that contains a constant term. In general, the aproximate sparsity assumption can either be interpreted as an assumption that there are only a few instruments that are important for explaining variation in the covariance matrix $\Omega(z_i)$ or as an assumption that the function $\rho(z_i)$ can be accurately approximated using a small set of basis terms in $b(z_i)$.

The parameter ϕ_ℓ can be estimated via LASSO:

$$\hat{\phi}_\ell = \arg\min_{\phi\in\mathbb{R}^{d_b}} \mathbb{E}_n[(x_{\ell i} - \epsilon_i(\beta_0)b(z_i)'\phi)^2] + \lambda\|\phi\|_1, \tag{2.2}$$

or via post-LASSO, refitting an unpenalized version of (2.2) using only the basis terms associated

[2]Under H_1, $\rho_\ell(z_i)$ can be estimated directly by solving empirical analogs of $\rho_\ell(z_i) = \arg\min_{\phi\in\Phi} \mathbb{E}[(x_{\ell i} - \eta_i(\beta_0)\phi(z_i))^2]$ where $\eta_i(\beta_0) = \epsilon_i(\beta_0) - \mathbb{E}[\epsilon_i(\beta_0)|z_i]$. This requires an initial estimate of $\mathbb{E}[\epsilon_i(\beta_0)|z_i]$, however.

with nonzero coeffecients in the inital LASSO regression. The estimating procedure in (2.2) is a simple ℓ_1-penalized regression of $x_{\ell i}$ against $\epsilon_i(\beta_0)b(z_i)$ and can be easily implemented using out-of-the-box software available on most platforms. Under standard conditions, this leads to a consistent estimate of $\rho_\ell(z_i)$ as long as the sparsity condition $s^2\log^M(d_b n)/n \to 0$ where s is the number of nonzero elements of ϕ_ℓ and M is a positive constant that depends on the moment bounds imposed. With $\hat{\rho}(z_i) \coloneqq b(z_i)'\hat{\phi}_\ell$, I construct the estimated version of $r_{\ell i}$, $\hat{r}_{\ell i} \coloneqq x_i - \hat{\rho}(z_i)\epsilon_i(\beta_0)$ for each $\ell \in [d_x]$.

Remark 2.1. Approximate sparsity of $\rho(z_i)$ may be a particularly palatable assumption in cases where the instrument set is generated by functions of a smaller initial set of instruments, as in Angrist and Krueger (1991), Paravisini et al. (2014), Gilchrist and Sands (2016), and Derenoncourt (2022). In these cases, the dimensionality of the basis, d_b, need not be much larger than the dimensionality of the instruments, d_z, to provide a good approximation of $\rho(z_i)$. If taking $b(z_i) = z_i$ provides a good approximation of $\rho_\ell(z_i)$, consistency of $\hat{\rho}(\cdot)$ is achievable under $d_z^2\log^M(d_z n)/n \to 0$ even if ϕ_ℓ is fully dense. This requirement is weaker than the $d_z^3/n \to 0$ requirement of the standard K-statistic.

2.1. Test Statistic

The test statistic is based on an arbitrary jackknife-linear estimate of the first stage,

$$\widehat{\Pi}_{\ell i} = \sum_{j\neq i} h_{ij}\hat{r}_{\ell j}, \quad \ell \in [d_x]$$

where the weights h_{ij} derive from a hat matrix $H = [h_{ij}] \in \mathbb{R}^{n\times n}$ that may depend only on the instruments $\boldsymbol{z}$. The phrase "hat matrix" is borrowed from ordinary least squares (OLS) where the projection matrix, $\boldsymbol{z}(\boldsymbol{z}'\boldsymbol{z})^{-1}\boldsymbol{z}'$, is sometimes referred to as the hat matrix in the sense that $\hat{\boldsymbol{x}} = \boldsymbol{z}(\boldsymbol{z}'\boldsymbol{z})^{-1}\boldsymbol{z}'\boldsymbol{x}$. It is important to note that while $\widehat{\Pi}_{\ell i}$ does not depend on $\hat{r}_{\ell i}$, it may depend on z_i through the hat matrix H.

Formally, the only structure I require on the hat matrix H is a balanced-design condition described in Section 3. However, for reasons explained in Section 4 it may be optimal to introduce some regularization in estimating the first-stage models $\widehat{\Pi}_{\ell i}$ so I suggest using the deleted diagonal ridge-regression hat matrix $H(\lambda^\star)$:

$$[H(\lambda^\star)]_{ij} = \begin{cases} [\boldsymbol{z}(\boldsymbol{z}'\boldsymbol{z} + \lambda^\star I_{d_z})^{-1}\boldsymbol{z}']_{ij} & \text{if } i \neq j \\ 0 & \text{otherwise} \end{cases} \tag{2.3}$$

where, following recommendations in Harrell (2015), the penalty parameter $\lambda^\star$ is set so that the effective degrees of freedom of the resulting hat matrix is no more than a fraction of the sample size:

$$\lambda^\star = \inf\{\lambda \geq 0 : \operatorname{trace}(\boldsymbol{z}(\boldsymbol{z}'\boldsymbol{z} + \lambda I_{d_z})^{-1}\boldsymbol{z}') \leq n/5\}$$

The ridge hat matrix has the benefit of being well defined even when the number of instruments is larger than the sample size. I stress, though, that the $\widehat{\Pi}_{\ell i}$ estimators are not required to be consistent and the researcher may use any other hat matrix that she believes will lead to plausible estimates of Π_i, for example the jackknife OLS of Angrist et al. (1999).

For each $i \in [n]$, define $\widehat{\Pi}_i \coloneqq (\widehat{\Pi}_{1i}, \ldots, \widehat{\Pi}_{d_x i}) \in \mathbb{R}^{d_x}$, $\widehat{\Pi}_{\epsilon i} \coloneqq \epsilon_i(\beta_0)\widehat{\Pi}_i$, and

$$\begin{aligned} \varepsilon(\beta_0) &\coloneqq \big(\varepsilon_1(\beta_0), \ldots, \varepsilon_n(\beta_0)\big)' \in \mathbb{R}^n \\ \widehat{\Pi} &\coloneqq \big(\widehat{\Pi}_1', \ldots, \widehat{\Pi}_n'\big)' \in \mathbb{R}^{n\times d_x} \\ \widehat{\Pi}_\varepsilon &\coloneqq \big(\widehat{\Pi}_{\epsilon 1}', \ldots, \widehat{\Pi}_{\epsilon n}'\big)' \in \mathbb{R}^{n\times d_x} \end{aligned} \tag{2.4}$$

The new test statistic, named the jackknife K-statistic for its inspiration by the work of Kleibergen (2002, 2005), can then be defined

$$JK(\beta_0) := \epsilon(\beta_0)'\widehat{\Pi}\big(\widehat{\Pi}_\epsilon'\widehat{\Pi}_\epsilon\big)^{-1}\widehat{\Pi}'\epsilon(\beta_0) \times \mathbf{1}\{\lambda_{\min}(\widehat{\Pi}_\epsilon'\widehat{\Pi}_\epsilon) > 0\} \tag{2.5}$$

I will show that, under appropriate conditions, the limiting distribution of $JK(\beta_0)$ under H_0 is $\chi^2_{d_x}$. For exposition, I will largely focus on the case where $d_x = 1$, in which case I can take advantage of the simplified form of the test statistic, $JK(\beta_0) = \big(\sum_{i=1}^n \epsilon_i(\beta_0)\widehat{\Pi}_i\big)^2 / \sum_{i=1}^n \epsilon_i^2(\beta_0)\widehat{\Pi}_i^2$. The extension to $d_x > 1$ is not immediate but is possible under strengthened moment conditions and is deferred to Section 5.

3. Limiting Behavior with a Single Endogenous Variable

The limiting behavior of the test statistic is analyzed via a direct Gaussian approximation technique. When there is a single endogenous variable this approach can be considerably simplified. For each $i \in [n]$, let $(\tilde{\epsilon}_i(\beta_0), \tilde{r}_i)'$ be jointly Gaussian random variables generated (i) independently of each other and the data and (ii) with the same mean and covariance matrix as $(\epsilon_i(\beta_0), r_i)'$. In addition, define $\tilde{\Pi}_i := \sum_{j\neq i} h_{ij}\tilde{r}_j$. The goal is to show that the quantiles of $JK(\beta_0)$ can be approximated by corresponding quantiles of the Gaussian statistic,

$$JK_G(\beta_0) := \frac{(\sum_{i=1}^n \tilde{\epsilon}_i(\beta_0)\tilde{\Pi}_i)^2}{\sum_{i=1}^n \mathbb{E}[\epsilon_i^2(\beta_0)]\tilde{\Pi}_i^2} \tag{3.1}$$

Since uncorrelated jointly Gaussian random variables are independent, under H_0 the vector $(\tilde{\epsilon}_1(\beta_0), \ldots, \tilde{\epsilon}_n(\beta_0))'$ is mean zero and independent of $(\tilde{r}_1, \ldots, \tilde{r}_n)'$. The null distribution of $JK_G(\beta_0)$ conditional on any realization of $(\tilde{r}_1, \ldots, \tilde{r}_n)'$ is then χ^2_1 and so its unconditional null distribution is also χ^2_1.

3.1. Interpolation Approach

Error arising from estimation of $\rho(z_i)$ prevents immediate comparison of the distribution of $JK(\beta_0)$ to the distribution of $JK_G(\beta_0)$. As such, I begin by considering the distribution of an infeasible statistic, $JK_I(\beta_0)$, which could be constructed if $\rho(z_i)$ were known to the researcher:

$$JK_I(\beta_0) := \frac{\big(\sum_{i=1}^n \epsilon_i(\beta_0)\widehat{\Pi}_i^I\big)^2}{\sum_{i=1}^n \epsilon_i^2(\beta_0)\big(\widehat{\Pi}_i^I\big)^2} \times \mathbf{1}\Big\{\sum_{i=1}^n \epsilon_i^2(\beta_0)\big(\widehat{\Pi}_i^I\big)^2 > 0\Big\}$$

where $\widehat{\Pi}_i^I = \sum_{j\neq i} h_{ij}r_j$. To show that the distribution of $JK_I(\beta_0)$ can be approximated by the distribution of $JK_G(\beta_0)$, I adapt Lindeberg's interpolation method, first introduced by Lindeberg (1922) in an elegant proof of the central limit theorem. This method consists of one-by-one replacment of the terms $(\epsilon_i(\beta_0), r_i)$ in the expression of $JK_I(\beta_0)$ with their Gaussian analogs, $(\tilde{\epsilon}_i(\beta_0), \tilde{r}_i)$, and bounding the resulting one-step distributional changes.

Applying the interpolation method directly on the statistics $JK_I(\beta_0)$ and $JK_G(\beta_0)$, however, is not tractable as it requires bounding expectations of derivatives with respect to terms in the denominator. When identification is weak, the denominators of $JK_I(\beta_0)$ and $JK_G(\beta_0)$ may both be arbitrarily close to zero with positive probability. Derivatives with respect to terms in the denominators thus may not have finite expectations. Instead, I consider a different approach.

For a scaling factor s_n, introduced below, define the scaled numerators and denominators

$$N := \left(\frac{s_n}{\sqrt{n}}\sum_{i=1}^{n}\epsilon_i(\beta_0)\widehat{\Pi}_i^I\right)^2 \qquad \tilde{N} := \left(\frac{s_n}{\sqrt{n}}\sum_{i=1}^{n}\tilde{\epsilon}_i(\beta_0)\tilde{\Pi}_i\right)^2$$

$$D := \frac{s_n^2}{n}\sum_{i=1}^{n}\epsilon_i^2(\beta_0)\left(\widehat{\Pi}_i^I\right)^2 \qquad \tilde{D} := \frac{s_n^2}{n}\sum_{i=1}^{n}\mathbb{E}[\epsilon_i^2(\beta_0)]\left(\tilde{\Pi}_i\right)^2$$

and for any $a \geq 0$, define the decomposed statistics

$$JK_I^a(\beta_0) := N - aD \qquad JK_G^a(\beta_0) := \tilde{N} - a\tilde{D}$$

Since $D = 0$ implies $N = 0$ and since $\tilde{D} \neq 0$ almost surely, the event $\{JK_V(\beta_0) \leq a\}$ is almost surely equivalent to the event $\{JK_V^a(\beta_0) \leq 0\}$ for $V = I, G$. The decomposed statistics no longer have denominators to be dealt with and are tractable for the interpolation argument. In local neighborhoods of H_0 and uniformly over $a \in \mathbb{R}$, I show that the probability that the infeasible decomposed statistic is less than zero can be approximated by the probaiblity that the gaussian decomposed statistic is less than or equal to zero.

I now detail the assumptions needed for the argument. Define $\eta_i := (\beta - \beta_0)v_i + \epsilon_i$ and $\zeta_i := v_i - \rho(z_i)\eta_i$, noting $\eta_i = \epsilon_i(\beta_0) - \mathbb{E}[\epsilon_i(\beta_0)]$ and $\zeta_i = r_i - \mathbb{E}[r_i]$. In what comes below $c > 1$ can be considered an arbitrary constant that may be updated upon each use but that does not depend on sample size n.

Assumption 3.1 (Moment Conditions)**.** *There is a fixed constant $c > 1$ such that (i) $\{|\Pi_i| + |(\beta - \beta_0)| + |\rho(z_i)|\} \leq c$, and (ii) for any $l, k \in \mathbb{N} \cup \{0\}$ such that $l + k \leq 6$, $c^{-1} \leq \mathbb{E}[|\eta_i|^l|\zeta_i|^k] \leq c$.*

Assumption 3.2 (Balanced Design)**.** *(i) For $s_n^{-2} = \max_i \mathbb{E}[(\widehat{\Pi}_i^I)^2]$ the following is bounded away from zero, $c^{-1} \leq \mathbb{E}[\frac{s_n^2}{n}\sum_{i=1}^n(\widehat{\Pi}_i^I)^2]$; (ii) $\max_i s_n^2\sum_{j\neq i}h_{ji}^2 \leq c$; and (iii) the following ratio is bounded away from zero: $\frac{\sum_{k=2}^n\lambda_k^2(HH')}{\sum_{k=1}^n\lambda_k^2(HH')} \geq c^{-1}$ where $\lambda_k(HH')$ represents the k^{th} largest eigenvalue of the matrix HH'.*

Assumption 3.1 imposes light moment conditions on the random variables η_i and ζ_i, which in turn imply restrictions on $\epsilon_i(\beta_0)$ and r_i. In particular, Assumption 3.1(i) imposes that $\epsilon_i(\beta_0)$ and r_i have finite means while Assumption 3.1(ii) bounds, both from above and away from zero, the first through sixth central moments of the random variables.

Assumption 3.2 imposes conditions on the hat matrix H. Assumption 3.2(i) requires that the average second moment of the infeasible first-stage estimators be on the same order as the maximum first-stage estimator second moment. This is imposed to rule out hat matrices that are close to zero and ensure that the effective number of observations used to test the null is growing with the sample size. Assumption 3.2(ii) requires that the maximum leverage of any observation be bounded. When H is symmetric, it is automatically satisfied under Assumption 3.1(i) and the definition of s_n.[1] Assumption 3.2(iii) can be viewed as a technical requirement that there be more than one "effective" instrument in the hat matrix.[2] Remark 3.1 below discusses how the validity of Assumption 3.2 may be assesed in practice.

The scaling factor s_n captures both the "size" of the elements in the hat matrix H and the strength of identification. If elements of the hat matrix are on the same order as a constant,

[1] To see this, notice that $s_n^{-2} = \max_i \mathbb{E}[(\widehat{\Pi}_i^I)^2] \geq \max_i \text{Var}(\widehat{\Pi}_i^I) = \max_i \sum_{j\neq i} h_{ij}^2 \text{Var}(r_j)$. By Assumption 3.1, $\text{Var}(r_j)$ is bounded from below by c^{-1}. Inverting this chain of inequalities yields that $s_n^2\sum_{j\neq i}h_{ij}^2$ is bounded from above uniformly over all $i \in [n]$.

[2] In the case of a standard projection matrix (no deleted diagonal), Assumption 3.2(iii) would be satisfied whenever $\text{rank}(z(z'z)^{-1}z) > 1$.

one would expect $s_n = O(n^{-1})$ under strong identification ($\Pi_i \propto 1$) while $s_n = O(n^{-1/2})$ under weak identification ($\Pi_i \lesssim n^{-1/2}$).

The local neighborhoods of H_0 considered are characterized by the local power index P, defined below, as well as an additional regularity condition. The local power index is a measure of the association between the true first stage Π_i and the first-stage estimates $\widehat{\Pi}_i$. Section 4, discusses discuss how the strength of this association is related to the power of the test under local alternatives.

$$P := (\beta - \beta_0)^2 \mathbb{E}\left[\left(\frac{s_n}{\sqrt{n}} \sum_{i=1}^{n} \Pi_i \widehat{\Pi}_i^I\right)^2\right]$$

Assumption 3.3 (Local Identification)**.** *(i) The local power index P is bounded, $P \leq c$; and (ii)* $\max_i \mathbb{E}[(s_n \sum_{j \neq i} h_{ji} \epsilon_j(\beta_0))^2] \leq c$.

Under H_0, Assumption 3.3 is trivially satisfied since $(\beta - \beta_0) = 0$ and $\sum_{j \neq i} s_n^2 h_{ji}^2 \leq c$. Assumption 3.3(ii) is an additional technical condition that requires that the maximum value of $\mathbb{E}[(\sum_{j \neq i} h_{ji} \epsilon_j(\beta_0))^2]$ be on the same or lesser order than the maximum value of $\mathbb{E}[(\sum_{j \neq i} h_{ij} r_j)^2]$. It is always satisfied whenever $\mathbb{E}[\epsilon_i(\beta_0)] = \Pi_i(\beta - \beta_0)$ is in a $\sqrt{n}$-neighborhood of zero in the sense that $|\Pi_i(\beta - \beta_0)| \leq C/\sqrt{n}$ for all $i \in [n]$ and some constant C. In general, Assumption 3.3(ii) can be roughly interpreted as requiring the local neighborhoods of H_0 considered to be those in which the means of $\epsilon_i(\beta_0)$ are of the same or lesser order than the means of r_i for $i \in [n]$.

Under Assumptions 3.1–3.3, I establish that the CDF of the infeasible statistic, $JK_I(\beta_0)$, can be uniformly approximated by the CDF of the Gaussian statistic, $JK_G(\beta_0)$. This result does not require $JK_G(\beta_0)$ to have a fixed limiting distribution.

Lemma 3.1 (Infeasible Uniform Approximation)**.** *Suppose that Assumptions 3.1–3.3 hold. Then,*

$$\sup_{a \in \mathbb{R}} \left| \Pr(JK_I(\beta_0) \leq a) - \Pr(JK_G(\beta_0) \leq a) \right| \to 0$$

I additionally show that the test based on the $JK_I(\beta_0)$ statistic is consistent whenever the power index diverges, $P \to \infty$, and Assumption 3.3(ii) holds.

Proposition 3.1 (Consistency)**.** *Suppose that Assumptions 3.1, 3.2, and 3.3(ii) hold. Then if $P \to \infty$ the test based on $JK_I(\beta_0)$ is consistent.*

The dependence of the consistency result on Assumption 3.3(ii) is a nontrivial restriction due to bias taken on in constructing r_i. In particular, against certain alternatives it is possible that $\mathbb{E}[\widehat{\Pi}_i^I] = 0$ for all $i \in [n]$ even under strong identification. In general, however, bias in $\mathbb{E}[r_i]$ does not imply a violation of Assumption 3.3(ii), which requires only that the *size* of $\mathbb{E}[r_i]$ be of a weakly greater order than that of $\mathbb{E}[\epsilon_i(\beta_0)]$. Moreover, Proposition 3.1 does not necessarily rule out consistency when $P \to \infty$ but Assumption 3.3(ii) fails. In Section 4, I propose a combination with the sup-score test to combat a power decline against the alternatives mentioned above.

3.2. Limiting Behavior of Test Statistic

The final step in characterizing the limiting behavior of the feasible test statistic is to show that the difference between the infeasible and feasible statistics is negligible. I begin with a technical lemma stating that the difference between $JK(\beta_0)$ and $JK_I(\beta_0)$ is asymptotically negligible whenever the differences between the scaled numerators and the scaled denominators

are asymptotically negligible. Define these differences:

$$\Delta_N := \frac{s_n}{\sqrt{n}} \sum_{i=1}^{n} \epsilon_i(\beta_0)(\widehat{\Pi}_i - \widehat{\Pi}_i^I)$$
$$\Delta_D := \frac{s_n^2}{n} \sum_{i=1}^{n} \epsilon_i^2(\beta_0)(\widehat{\Pi}_i^2 - (\widehat{\Pi}_i^I)^2)$$

Lemma 3.2. *Suppose Assumptions 3.1–3.3 hold and* $(\Delta_N, \Delta_D)' \to_p 0$. *Then* $|JK(\beta_0) - JK_I(\beta_0)| \to_p 0$.

While Lemma 3.2 is a simple statement, it is not obvious. In particular, showing that the difference between the infeasible and feasible statistics is negligible requires showing that $1/(D + \Delta_D)$ is bounded in probability, where D represents the scaled denominator of $JK_I(\beta_0)$. In a standard analysis, this would be done by arguing that D converges in distribution and applying the continuous mapping theorem. This approach is not applicable here as D may not have a fixed limiting distribution. Instead, I directly show that $1/(D + \Delta_D)$ is bounded in probability by showing $\Pr(D \leq \delta_n) \to 0$ for any sequence $\delta_n \to 0$.

Lemma 3.2 allows the researcher to use alternate choices of estimators for $\rho(z_i)$, so long as they can verify that $(\Delta_N, \Delta_D)' \to_p 0$. Below, I verify that this condition can be satisfied for the ℓ_1-penalized estimation procedure proposed in (2.2). This requires a strengthened moment condition on η_i. Given a random variable X and $v > 0$ the Orlicz (quasi-)norm is defined

$$\|X\|_{\psi_v} := \inf\{t > 0 : \mathbb{E}\exp(|X|^v/t^v) \leq 2\}$$

Random variables with a finite Orlicz norm for some $v \in (0, 1] \cup \{2\}$ are termed α-sub-exponential random variables (Gotze et al., 2021). This class encompasses a wide range of potential distributions including all bounded and sub-Gaussian random random variables (with $v = 2$), all sub-exponential random variables such as Poisson or noncentral χ^2 random variables (with $v = 1$), as well as random variables with "fatter" tails such as Weibull distributed random variables with shape parameter $v \in (0, 1]$.

Assumption 3.4 (Estimation Error). *(i) There is a fixed constant* $v \in (0, 1] \cup \{2\}$ *such that* $\|\eta_i\|_{\psi_v} \leq c$; *(ii) The basis terms* $b(z_i)$ *are bounded,* $\|b(z_i)\|_\infty \leq C$ *for all* $i = 1, \ldots, n$; *(iii) the approximation error satisfies* $(\mathbb{E}_n[\xi_i^2])^{1/2} = o(n^{-1/2})$; *(iv) the researcher has access to an estimator* $\widehat{\phi}$ *of* ϕ *that satisfies* $\log(d_b n)^{2/(v \wedge 1)}\|\widehat{\phi} - \phi\|_1 \to_p 0$; *(v) the following moment bounds hold*

(va) $\max_{1 \leq \ell \leq d_b} \left|\mathbb{E}\left[\frac{s_n}{\sqrt{n}} \sum_{i=1}^n \sum_{j \neq i} h_{ij}\epsilon_i(\beta_0)b_\ell(z_j)\epsilon_j(\beta_0)\right]\right| \leq c$

(vb) $\max_{\substack{1 \leq i \leq n \\ 1 \leq \ell \leq d_b}} |\mathbb{E}[s_n \sum_{j \neq i} h_{ij} b_\ell(z_j)\epsilon_j(\beta_0)]| \leq c.$

Assumption 3.4(i) strengthens the moment condition on η_i to require that η_i be in the class of α-sub-exponential random variables. Assumption 3.4(ii) is a standard condition in ℓ_1-penalized estimation. At the cost of extra notation, it can be relaxed and the sup-norm of the basis terms can be allowed to grow slowly with the sample size to accommodate normalized b-splines or wavelets. Assumption 3.4(iii) is a bound on the rate of decay of the approximation error, similar to the approximate sparsity condition of Belloni et al. (2012). Assumption 3.4(v) is a strengthening of the definition of local neighborhoods and can be interpreted similarly to Assumption 3.3(ii). It is satisfied under H_0.

Assumption 3.4(iv) is a high-level condition on the rate of consistency of the parameter estimate $\hat{\phi}$ in the ℓ_1 norm. This can be verified under approximate sparsity for both the LASSO estimator in (2.2) or post-LASSO procedures based on refitting an unpenalized version of (2.2) only

using the basis terms selected in a LASSO first stage.[3] Under appropriate choice of penalty parameter, this is satisfied when $s^2 \log^{2(v+1)/v}(d_b n)/n \to 0$, where s denotes the number of nonzero elements of ϕ. This condition allows for the dimensionality of the basis terms, d_b, to grow near exponentially as a function of the sample size.

Under Assumptions 3.1–3.4, I establish that the difference between the infeasible and feasible statistics can be treated as negligible when the estimation procedure proposed in (2.2) is used. In combination with Lemma 3.1 this yields the main result.

Theorem 3.1 (Uniform Approximation)**.** *Suppose that Assumptions 3.1–3.4 hold. Then*

$$\sup_{a\in\mathbb{R}} \left|\Pr(JK(\beta_0) \le a) - \Pr(JK_G(\beta_0) \le a)\right| \to 0$$

Corollary 3.1 (Size Control)**.** *Suppose that Assumptions 3.1, 3.2 and 3.4 hold. Then, under H_0, $JK(\beta_0) \rightsquigarrow \chi^2_1$.*

While $JK_G(\beta_0)$ does not have a fixed distribution, examining its behavior is still tractable and allows for insight into the power properties of the jackknife K-test. In the next section, I use this result to analyze the local power of the proposed test. To improve power against certain alternatives, I suggest a combination with the sup-score statistic of Belloni et al. (2012).

Remark 3.1. A sufficient condition for Assumption 3.2(i) is that there is some fixed quantile $q \in (0, 100)$ such that $(cq)^{-1} \le \frac{q^{\text{th}}\text{-quantile of } \mathbb{E}[(\widehat{\Pi}_i^l)^2]}{\max_i \mathbb{E}[(\widehat{\Pi}_i^l)^2]}$. In practice this can be verified by checking that there is some quantile q such that both

$$\frac{q^{\text{th}}\text{-quantile of } \sum_{j\neq i} h_{ij}^2}{\max_i \sum_{j\neq i} h_{ij}^2} \text{ and } \frac{q^{\text{th}}\text{-quantile of } (\sum_{j\neq i} h_{ij}\hat{r}_j)^2}{\max_i (\sum_{j\neq i} h_{ij}\hat{r}_j)^2} \tag{3.2}$$

are bounded away from zero. Similarly, Assumption 3.2(ii) can be verified by checking that $\max_i \sum_{j\neq i} h_{ji}^2 / \max_i \sum_{j\neq i} h_{ij}^2$ is bounded from above while Assumption 3.2(iii) can be verified by checking the eigenvalues of HH'.

4. Improving Power against Certain Alternatives

Using the characterization of the limiting behavior of the test statistic derived in Section 3, I analyze the local power properties of the test. Unfortunately, against certain alternatives the test statistic may have trivial power, a deficiency shared with the K-statistics of Kleibergen (2002, 2005). To combat this, I propose a simple combination with the sup-score statistic of Belloni et al. (2012) based on a thresholding rule.

4.1. Local Power Properties

In local neighborhoods of H_0, as defined in Assumptions 3.3 and 3.4, Theorem 3.1 implies that the limiting behavior of $JK(\beta_0)$ can be analyzed by examining the behavior of the Gaussian analog statistic, $JK_G(\beta_0)$. Under local alternatives, $\Pi_i^2(\beta - \beta_0)^2 \to 0$, the distribution of $JK_G(\beta_0)$ conditional on $\tilde{r} = (\tilde{r}_1, \ldots, \tilde{r}_n)$ is nearly noncentral χ^2_1, with noncentrality parameter

$$\mu^2_\infty(\tilde{r}) = (\beta - \beta_0)^2 \frac{\left(\sum_{i=1}^n \Pi_i \tilde{\Pi}_i\right)^2}{\sum_{i=1}^n \text{Var}(\eta_i)\tilde{\Pi}_i^2}. \tag{4.1}$$

[3] See Belloni et al. (2012), van der Greer (2016), Tan (2017), and Chetverikov and Sørensen (2021) for references under various choices of penalty parameter.

The numerator of $\mu^2_\infty(\tilde{r})$ suggests that power is maximized when the first-stage estimate $\tilde{\Pi}_i$ is close to the true first stage value Π_i. Indeed, when errors are homoskedastic $\mu^2_\infty(\tilde{r})$ is maximized by setting $\tilde{\Pi}_i = \Pi_i$ reflecting the classical result of Chamberlain (1987). The denominator of $\mu^2_\infty(\tilde{r})$ suggests that having first-stage estimates $\tilde{\Pi}_i$ with low second moments may increase power. This guides the recommendation for the use of ℓ_2-regularization in constructing the hat matrix, H.

Unfortunately, estimators of Π_i based on $r_i = x_i - \rho(z_i)\epsilon_i(\beta_0)$ may not be close to Π_i under H_1. This is because the mean of r_i will in general differ from Π_i

$$\mathbb{E}[r_i] = \Pi_i - \rho(z_i)\Pi_i(\beta - \beta_0)$$

This deficiency is inherited from the similarity of the $JK(\beta_0)$ statistic to the K-statistic. As pointed out by Moreira (2001), this need not be an issue as long as there is a fixed constant $C \neq 0$ such that $\mathbb{E}[r_i] = C\Pi_i$ for all $i \in [n]$. However, in general, this will introduce bias into the first-stage estimates $\widehat{\Pi}_i$ under H_1. The power implications of this bias are particularly pronounced when $\rho(z_i)$ is a constant $(\beta - \beta_0) = 1/\rho(z_i)$. In this case, $\mathbb{E}[r_i]$, and thus $\mathbb{E}[\tilde{\Pi}_i]$, will equal zero for each $i \in [n]$, and the $JK(\beta_0)$ statistic will select a direction completely at random to direct power into.[1]

4.2. A Simple Combination Test

To combat this loss of power for tests based on the K-statistic, a common strategy is to combine the K-statistic with the Anderson-Rubin statistic based on a conditioning statistic. While the Anderson-Rubin statistic does not have optimal power on its own, it has the benefit of directing power equally in all directions avoiding the pitfalls of the K-statistic which lacks power in certain directions. Prominent examples of such tests are the conditional likelihood ratio test of Moreira (2003), the GMM-M test of Kleibergen (2005), and the minimax regret tests of Andrews (2016). These combinations make use of the fact that the Anderson-Rubin statistic is asymptotically independent of both the K-statistic and the conditioning statistic.

Following this approach, I consider a simple combination of my proposed test statistic with the sup-score statistic of Belloni et al. (2012). The test based on the sup-score statistic (4.2) is similar in spirit to the Anderson-Rubin test but controls size even when d_z grows near exponentially as a function of the sample size.

$$S(\beta_0) := \sup_{1\leq\ell\leq d_z} \left| \frac{\sum_{i=1}^n \epsilon_i(\beta_0) z_{\ell i}}{(\sum_{i=1}^n z_{\ell i}^2)^{1/2}} \right| \tag{4.2}$$

A size $\theta \in (0,1)$ test based on the sup-score statistic rejects whenever $S(\beta_0) > c^S_{1-\theta}$ where, for $e_1, \ldots, e_n$ iid standard normal and generated independently of the data, $c^S_{1-\theta}$ is the simulated multiplier bootstrap critical value:

$$c^S_{1-\theta} := (1-\theta) \text{ quantile of } \sup_{1\leq\ell\leq d_z} \left| \frac{\sum_{i=1}^n e_i \epsilon_i(\beta_0) z_{\ell i}}{(\sum_{i=1}^n z_{\ell i}^2)^{1/2}} \right| \text{ conditional on } \{(y_i, x_i, z_i)\}_{i=1}^n.$$

As with the Anderson-Rubin test, tests based on the sup-score statistic may have suboptimal power properties in overidentified models as they do not incorporate first-stage information. However, the sup-score statistic does retain the benefit of directing power evenly in all directions, avoiding pitfalls of tests based on $JK(\beta_0)$ against certain alternatives.

[1] Andrews et al. (2006) and Andrews (2016) point out this deficiency in the context of the K-statistics of Kleibergen (2002, 2005).

The combination test will be based on an attempt to detect whether the alternative β is such that $\mathbb{E}[\widehat{\Pi}_i^I] = 0$ for all $i = 1, \ldots, n$. When this is the case, the test based on $JK(\beta_0)$ will choose a direction completely at random to direct power into. It would then be optimal for the researcher to test the null hypothesis using the sup-score statistic. Detection of whether $\mathbb{E}[\widehat{\Pi}_i^I] = 0$ is based on the conditioning statistic:

$$C = \max_{1 \leq i \leq n} \left| \frac{\sum_{j \neq i} h_{ij} \hat{r}_j}{(\sum_{j \neq i} h_{ij}^2)^{1/2}} \right|. \tag{4.3}$$

Under the assumption that $\mathbb{E}[\widehat{\Pi}_i^I] = 0$ for all $i \in [n]$, quantiles of the conditioning statistic can be simulated analogously to the sup-score critical value. For a new set of $e_1, \ldots, e_n$ iid standard normal and generated independently of the data, and for any $\theta \in (0, 1)$, define the conditional quantile

$$c_{1-\theta}^C := (1-\theta) \text{ quantile of } \max_{1 \leq i \leq n} \left| \frac{\sum_{j \neq i} e_i h_{ij} \hat{r}_j}{(\sum_{j \neq i} h_{ij}^2)^{1/2}} \right| \text{ conditional on } \{(y_i, x_i, z_i)\}_{i=1}^n \tag{4.4}$$

Depending on the value of the conditioning statistic, the thresholding test decides whether the test based on $JK(\beta_0)$ or one based on $S(\beta_0)$ should be run.

$$T(\beta_0; \tau) = \begin{cases} \mathbf{1}\{JK(\beta_0) > \chi^2_{1;1-\alpha}\} & \text{if } C \geq \tau \\ \mathbf{1}\{S(\beta_0) > c^S_{1-\alpha}\} & \text{if } C < \tau \end{cases} \tag{4.5}$$

for some cutoff τ, which I take in the simulation study and empirical exercise to be the 75^{th} quantile of the distribution of C under the assumption that $\mathbb{E}[\widehat{\Pi}_i^I] = 0, \forall i \in [n]$.

To show that the thresholding test controls size, I show that the joint distribution of $JK(\beta_0)$, $S(\beta_0)$, and C can be uniformly approximated by the joint distribution of Gaussian analog statistics. In the limiting Gaussian regime the jackknife K-statistic and sup-score stastic are each marginally independent of the conditioning statistic, implying that asymptotic size control is not affected by looking at the conditioning statistic before deciding which test to run. It is noteable that this independence is only marginal, the joint distribution jackknife K-statistic and sup-score statistic may depend on the conditioning statistic even in the limiting Gaussian regime. This prevents more sophisticated combinations like those of Moreira (2003) or Andrews (2016). However, the simple combination test proposed here appears to preform well both in empirical practice and simulation study.

To establish that the asymptotic validity of the thresholding test, I rely on the following assumption:

Assumption 4.1 (Combination Conditions)**.** *Assume that (i) there is a* $v \in (0, 1] \cup \{2\}$ *such that* $\|\zeta_i\|_{\psi_v} \leq c$*; (ii)* $\max_{i,j} |\frac{h_{ij}}{(\mathbb{E}_n[h_{ij}^2])^{1/2}}| + \max_{l,i} |\frac{z_{li}}{(\mathbb{E}_n[z_{li}^2])^{1/2}}| \leq c$*; and (iii)* $\log^{7+4/v}(d_z n)/n \to 0$.

Assumption 4.1(i) is a strengthening of the moment bound on r_i similar to that of Assumption 3.4(i). As discussed, while more restrictive than the condition in Assumption 3.1, this still allows for a wide range of potential distributions for r_i. Assumption 4.1(ii) requires that the number of observations used to test $\mathbb{E}[\widehat{\Pi}_i] = 0$ via the conditioning statistic and the number of observations used to test the null hypothesis via the sup-score test are both growing with the sample size. It can be verified by looking at the hat matrix H and the instruments. Finally, Assumption 4.1(iii) is a light requirement on the number of instruments d_z needed for the validity of the sup-score test. It allows the number of instruments to grow near exponentially as a function of sample size.

Theorem 4.1. *Suppose Assumptions 3.1–3.4 and 4.1 hold. Then,*

1. *the test based on $T(\beta_0;\tau)$ has asymptotic size α for any choice of cutoff τ, and*
2. *if $\mathbb{E}[\widehat{\Pi}_i^I] = 0$ for all $i \in [n]$, there exist sequences $\delta_n \searrow 0$ and $\beta_n \searrow 0$ such that with probability at least $1-\delta_n$,*

$$\sup_{\theta\in(0,1)} \left|\Pr_e(C \le c^C_{1-\theta}) - (1-\theta)\right| \le \beta_n,$$

where $\Pr_e(\cdot)$ denotes the probability with respect to only the variables $e_1,\dots,e_n$.

The first part of Theorem 4.1 establishes the asymptotic validity of the thresholding test $T(\beta_0;\tau)$ for any choice of cutoff τ. The proof of this statement follows the logic outlined above. The second part of Theorem 4.1 establishes the validity of the multiplier bootstrap procedure to approximate quantiles of the conditioning statistic. It follows directly from results in Belloni et al. (2018) after verifying that error taken on from estimation of $\rho(z_i)$ can be treated as negligible.

Remark 4.1. As mentioned by Andrews (2016) in the context of the standard K-statistic, this attempt to rectify the power deficiency via this particular conditioning statistic is not perfect. In particular, under heteroskedasticity, the means of the partialed-out endogenous variables, $\mathbb{E}[r_i]$, may not be scaled versions of the true first stages. However, as long as $\mathbb{E}[r_i] \neq 0$, one can still expect $\mathbb{E}[\widehat{\Pi}_i^I] = \sum_{j\neq i} h_{ij}\Pi_i + (\beta-\beta_0)\sum_{j\neq i} h_{ij}\rho(z_i)\Pi_i$ to be related to the true fist stage Π_i and for the test to have nontrivial power. Moreover, in light of the dependence of the consistency result in Proposition 3.1 on Assumption 3.3(ii), in the case where $\mathbb{E}[\widehat{\Pi}_i] = 0$ for all $i \in [n]$ it may be particularly important to avoid using the jackknife K-statistic to test H_0.

5. Analysis with Multiple Endogenous Variables

To analyze the limiting behavior of the test statistic when $d_x > 1$, I follow the basic idea of Section 3, which is to show that quantiles of the jackknife K-statistic can be approximated by analogous quantiles of the Gaussian statistic:

$$JK_G(\beta_0) := \tilde{\epsilon}(\beta_0)\tilde{\Pi}(\tilde{\Pi}_\epsilon'\tilde{\Pi}_\epsilon)^{-1}\tilde{\Pi}'\tilde{\epsilon}(\beta_0);$$

where $(\tilde{\epsilon}_i(\beta_0), \tilde{r}_i)'$ are Gaussian with the same mean and covariance matrix as $(\epsilon_i(\beta_0), r_i)'$ and for $\tilde{\Pi}_{\ell i} = \sum_{j\neq i} h_{ij}\tilde{r}_{\ell j}$ define $\tilde{\Pi}_i := (\tilde{\Pi}_{1i},\dots,\tilde{\Pi}_{d_x i})' \in \mathbb{R}^{d_x}$, $\tilde{\Pi}_{\epsilon i} := (\mathbb{E}[\epsilon_i^2(\beta_0)])^{1/2}\tilde{\Pi}_i$,

$$\begin{aligned}\tilde{\epsilon}(\beta_0) &:= (\tilde{\epsilon}_1(\beta_0),\dots,\tilde{\epsilon}_n(\beta_0))' \in \mathbb{R}^n,\\ \tilde{\Pi} &:= (\tilde{\Pi}_1,\dots,\tilde{\Pi}_n)' \in \mathbb{R}^{n\times d_x},\\ \text{and}\quad \tilde{\Pi}_\epsilon &:= (\tilde{\Pi}_{\epsilon 1},\dots,\tilde{\Pi}_{\epsilon n})' \in \mathbb{R}^{n\times d_x}.\end{aligned}$$

As in Section 3, notice that, since uncorrelated random variables are independent, under H_0 the vector $\tilde{\epsilon}(\beta_0)$ is mean zero and independent of $(\tilde{\Pi}, \tilde{\Pi}_\epsilon)$. Conditional on any realization of $(\tilde{\Pi}, \tilde{\Pi}_\epsilon)$ the $JK_G(\beta_0)$ statistic then follows a $\chi^2_{d_x}$ distribution, and thus, its unconditional distribution is also $\chi^2_{d_x}$. As before, error taken on from the estimation of $\rho(z_i)$ prevents immediate comparison of $JK(\beta_0)$ to $JK_G(\beta_0)$ so I begin by showing that the quantiles of an infeasible statistic, $JK_I(\beta_0)$, can be approximated by corresponding quantiles of $JK_G(\beta_0)$ where:

$$JK_I(\beta_0) := \epsilon(\beta_0)(\widehat{\Pi}^I)((\widehat{\Pi}^I_\epsilon)'(\widehat{\Pi}^I_\epsilon))^{-1}(\widehat{\Pi}^I)'\epsilon(\beta_0),$$

for $\widehat{\Pi}^I$ and $\widehat{\Pi}^I_\epsilon$ defined the same way as $\widehat{\Pi}$ and $\widehat{\Pi}_\epsilon$ in (2.4), respectively, but using the true values $(r_1,\dots,r_n)'$ in place of their estimates $(\hat{r}_1,\dots,\hat{r}_n)'$.

When there are multiple endogenous variables, $d_x > 1$, I cannot take advantage of the simplified form of the test statistic to establish this approximation as in Section 3. Instead I deal directly with the test statistics themselves. Consider functions $\varphi_\gamma(\cdot) \in C_b^3(\mathbb{R})$ that approximate the indicators $\mathbf{1}\{\cdot \leq a\}$, where $a \in \mathbb{R}$ is arbitrary and γ is a scaling factor inversely proportional to the quality of the approximation but positively proportional to the derivatives of φ_γ. The goal is to show, for a sequence γ_n tending to zero, that

$$\mathbb{E}[\varphi_{\gamma_n}(JK_I(\beta_0)) - \varphi_{\gamma_n}(JK_G(\beta_0))] \to 0 \tag{5.1}$$

As mentioned in Section 3, the Lindeberg's original approach cannot be applied the derivative of the test statistics with respect to terms in the denominator matrix, $\widehat{\Pi}_\epsilon'\widehat{\Pi}_\epsilon$, may be as large as the inverse of the minimum eigenvalue of the denominator matrix. When identification is sufficiently weak, the denominator matrix will have a nonnegligible distribution and the inverse of its minimum eigenvalue may not have finite moments.

To get around this, I modify the argument by considering a "data-dependent" choice of approximation parameter γ_n. This choice of approximation parameter inversely scales with the determinant of the denominator matrix and thus, since the determinant is the product of the eigenvalues, inversely scales with the minimum eigenvalue.[1] Geometrically, this approach can be thought of as "stretching out" the function $\varphi_{\gamma_n}(\cdot)$ in directions where the minimum eigenvalue of the denominator matrix is close to zero. Through the chain rule, this allows for control of the overall derivative of $\varphi_{\gamma_n}(JK_I(\beta_0))$ with respective to an individual observation.

This approach relies on stronger moment conditions, needed mainly needed to bound moments of the determinant of the denominator matrix. For all $\ell = 1, \ldots, d_x$ let $\zeta_{\ell i} := v_i - \rho_\ell(z_i)\eta_i$ and $\eta_i = \epsilon_i - v_i'(\beta - \beta_0)$, the demeaned versions of $r_{\ell i}$ and $\epsilon_i(\beta_0)$, respectively.

Assumption 5.1 (Moment Conditions)**.** *Assume (i) there are constants $c > 1$ and $\upsilon \in (0, 1] \cup \{2\}$ such that $\|\epsilon_i\|_{\psi_a} \leq c$ and $\|\zeta_{\ell i}\|_{\psi_\upsilon} \leq c$, and (ii) $c^{-1} \leq \lambda_{\min}(\mathbb{E}[\eta_i\eta_i']) \leq \lambda_{\max}(\mathbb{E}[\eta_i\eta_i']) \leq c$.*

Assumption 5.2 (Balanced Design)**.** *(i) Let $s_{\ell,n}^{-2} = \max_{1\leq i\leq n} \mathbb{E}[(\widehat{\Pi}_{\ell i}^I)^2]$ for each $\ell \in [d_x]$; then, the minimum eigenvalue of the following matrix is bounded away from zero:*

$$c^{-1} \leq \lambda_{\min}\mathbb{E}\left(\frac{s_{\ell,n}s_{k,n}}{n}\sum_{i=1}^n(\widehat{\Pi}_{\ell i}^I)(\widehat{\Pi}_{ki}^I)\right)_{\substack{1\leq\ell\leq d_x\\ 1\leq k\leq d_x}}$$

(ii) $\max_i s_n \sum_{j\neq i} h_{ji}^2 \leq c$; and (iii) the following ratio is bounded away from zero: $\frac{\sum_{k=2}^n \lambda_k^2(HH')}{\sum_{k=1}^n \lambda_k^2(HH')} \geq c^{-1}$ where $\lambda_k(HH')$ represents the k^{th} largest eigenvalue of the matrix HH'.

Assumption 5.1(i) strengthens Assumption 3.1 to require that the random variables (η_i, ζ_i), and thus, by extension, $(\epsilon_i(\beta_0), r_i)$ are υ-sub-exponential. As discussed below Assumption 3.4 this is more restrictive than the finite sixth moments needed to establish Lemma 3.1 but still allows for a wide range of possible distributions. Assumption 5.1(ii) is a light regularity condition requiring that the random variables $(\eta_{1i}, \ldots, \eta_{d_x i})$ be linearly independent.

Assumption 5.2(i) is a natural extension of Assumption 3.2(i) to the setting where $d_x > 1$. It requires that the average second moment of any linear combination of the first-stage estimates is proportional to the maximum second moment of the same linear combination. Assumption 5.2(ii,iii) are the same conditions as Assumption 3.2(ii,iii) and can again be implicitly thought of as requiring that the maximum leverage of any one observation be bounded and

[1] The determinant has the benefit of being a smooth function of elements of the matrix. This makes it nicer to work with than the minimum eigenvalue itself, which loses differentiability when the dimension of its eigenspace is larger than one.

there be than two effective instruments in the hat matrix. Assumption 5.2 thus reduces to Assumption 3.2 when $d_x = 1$.

Finally, two assumptions are needed to define the local neighborhoods considered and to show that estimation error can be treated as negligible when using the ℓ_1-penalized procedure detailed in Section 2. These assumptions are the natural extensions Assumption 3.3 and Assumption 3.4, respectively, and the reader is thus referred to Section 3 for a discussion of their interpretations.

Assumption 5.3 (Local Identification). *(i) The local power index is bounded $P \leq c$ for*

$$P = \sum_{\ell=1}^{d_x} \mathbb{E}\left[\left(\frac{s_{\ell,n}}{\sqrt{n}} \sum_{i=1}^{n} \widehat{\Pi}_{\ell i}^{I} \Pi_i'(\beta - \beta_0)\right)^2\right]$$

(ii) $\mathbb{E}[(s_{n,\ell} \sum_{j \neq i} h_{ji} \epsilon_j(\beta_0))^2] \leq c$ for all $\ell = 1, \ldots, d_x$.

Assumption 5.4 (Estimation Error). *(i) The basis terms $b(z_i)$ are bounded, $\|b(z_i)\|_\infty \leq C$ for all $i = 1, \ldots, n$; (ii) the approximation error satisfies $(\mathbb{E}_n[\xi_{\ell i}^2])^{1/2} = o(n^{-1/2})$; (iii) the researcher has access to estimators $\widehat{\phi}_\ell$ of ϕ_ℓ that satisfy $\log(d_b n)^{2/(\upsilon \wedge 1)} \|\widehat{\phi}_\ell - \phi_\ell\|_1 \to_p 0$ for each $\ell \in [d_x]$; and (iv) locally identified in the sense that*

(iva) $\max_{\substack{1 \leq \ell \leq d_x \\ 1 \leq k \leq d_b}} \left|\mathbb{E}\left[\frac{s_{n,\ell}}{\sqrt{n}} \sum_{i=1}^{n} \sum_{j \neq i} h_{ij} \epsilon_i(\beta_0) b_k(z_j) \epsilon_j(\beta_0)\right]\right| \leq c$

(ivb) $\max_{\substack{1 \leq i \leq n \\ 1 \leq \ell \leq d_b}} |\mathbb{E}[s_{n,\ell} \sum_{j \neq i} h_{ij} b_\ell(z_j) \epsilon_j(\beta_0)]| \leq c.$

With these stated, I show that the results of Section 3 can be extended to the general case.

Theorem 5.1 (Uniform Approximation). *Suppose that Assumptions 5.1–5.4 hold. Then,*

$$\sup_{a \in \mathbb{R}} |\Pr(JK(\beta_0) \leq a) - \Pr(JK_G(\beta_0) \leq a)| \to 0$$

In particular, under H_0, $JK(\beta_0) \rightsquigarrow \chi^2_{d_x}$.

As discussed in Section 4.1, tests based on the jackknife K-statistic may suffer from suboptimal power properties against certain alternatives. The power decline is particularly prnounced whenever $\mathbb{E}[\widehat{\Pi}_{\ell i}] = 0$ for some $\ell \in [d_x]$ and all $i \in [n]$. To improve power in this direction, I propose a generalization of the thresholding test in Section 4.2 based on the conditioning statistic C

$$C := \min_{1 \leq \ell \leq d_x} \max_{1 \leq i \leq n} \left|\frac{\sum_{j \neq i} h_{ij} \hat{r}_{\ell j}}{(\sum_{j \neq i} h_{ij}^2)^{1/2}}\right| \tag{5.2}$$

The conditioning statistic C attempts to detect whether, for *some* $\ell \in [d_x]$, $\mathbb{E}[\widehat{\Pi}_{\ell i}^{I}] = 0$ for all $i \in [n]$. Under the assumption that $\mathbb{E}[\widehat{\Pi}_{\ell i}^{I}] = 0, \forall i \in [n], \ell \in [d_x]$, quantiles of C can be simulated by multiplier bootstrap. Let $e_1, \ldots, e_n$ be generated iid standard normal independent of the data and for any $\theta \in (0, 1)$, define the conditional bootstrap quantile:

$$c_{1-\theta}^{C} := (1-\theta) \text{ quantile of } \min_{1 \leq \ell \leq d_x} \max_{1 \leq i \leq n} \left|\frac{\sum_{j \neq i} e_j h_{ij} \hat{r}_j}{(\sum_{j \neq i} h_{ij}^2)^{1/2}}\right| \text{ conditional on } \{(y_i, x_i, z_i)\}_{i=1}^{n}$$

Based on the value of the conditioning statistic the researcher can decide whether to run a test

based on $JK(\beta_0)$ or a test based on the sup-score statistic $S(\beta_0)$.

$$T(\beta_0;\tau) := \begin{cases} \mathbf{1}\{JK(\beta_0) > \chi^2_{d_x;1-\alpha}\} & \text{if } C > \tau \\ \mathbf{1}\{S(\beta_0) > c^S_{1-\alpha}\} & \text{if } C \leq \tau \end{cases} \tag{5.3}$$

As with Theorem 4.1, I show the asymptotic validity of the thresholding test by first establishing that quantiles of $(JK(\beta_0), C)$ and $(S(\beta_0), C)$ can jointly be approximated by Gaussian analogs and then using the marginal independence of the Gaussian analog testing and conditioning statistics under the null; $(JK(\beta_0) \perp C)$ and $(S\beta_0) \perp C)$ under H_0.

Theorem 5.2. *Suppose that Assumptions 4.1(ii,iii), 5.1, 5.2, and 5.4 hold. Then,*

1. *the test based on $T(\beta_0;\tau)$ has asymptotic size α for any choice of cutoff τ, and*
2. *if $\mathbb{E}[\widehat{\Pi}^I_{\ell i}] = 0$ for all $i \in [n]$ and $\ell \in [d_x]$, there exist sequences $\delta_n \searrow 0$ and $\beta_n \searrow 0$ such that with probability at least $1-\delta_n$,*

$$\sup_{\theta\in(0,1)} |\Pr_e(C \leq c^C_{1-\theta}) - (1-\theta)| \leq \beta_n$$

where $\Pr_e(\cdot)$ denotes the probability with respect to only the variables $e_1, \ldots, e_n$.

The first part of Theorem 5.2 establishes the validity of the test based on the thresholding statistic for any choice of cutoff τ. In practice, I recommend taking the cutoff, τ, to be the 75^{th} quantile of the distribution of C under the assumption that $\mathbb{E}[\widehat{\Pi}^I_{\ell i}] = 0$ for all $\ell \in [d_x]$ and $i \in [n]$. The second part of Theorem 5.2 establishes that this quantile can be simulated via the multiplier bootstrap procedure described above.

6. Empirical Application

I apply the testing procedures proposed in this paper to the data of Gilchrist and Sands (2016), who examine the effect of social spillovers in movie consumption. The sample consists of 1,671 opening weekend days[1] between January 1, 2002 and January 1, 2012. For each opening weekend, the authors observe gross ticket sales for movies wide released in theaters in the United States with a run in theaters of at least six weeks.[2] The data are obtained through Box Office Mojo, a subsidiary of the Internet Movie Database (IMDb).

The outcome variables of interest are gross ticket sales of movies that opened in a given weekend in the second through sixth weeks of their run, while the endogenous variable is the gross ticket sales of a movie in its opening weekend. To control for seasonal periodicity in both the supply of and demand for movies, a vector of date controls are included. Formally, the authors are interested in the parameters β_w, $w = 2, \ldots, 7$ from the linear IV model(s):

$$\text{Sales}^\perp_{wi} = \beta_w \text{Sales}^\perp_{1i} + \epsilon_{wi} \tag{6.1}$$

where, for $w = 1, \ldots, 6$, $\text{Sales}^\perp_{wi}$ represents gross national ticket sales, after the partialing out of date controls and a constant, $7(w-1)$ days after day i, of movies that opened on the opening weekend of i. The variable $\text{Sales}^\perp_{7i} = \sum_{w=1}^6 \text{Sales}^\perp_{wi}$ denotes the cumulative national ticket sales in the second through sixth running weekends of movies who opened in weekend i, after the partialing out of date controls and a constant. The parameter β_w represents the social spillover effect of strong opening weekend sales on sales in later weeks.

[1] An opening weekend day is a Friday, Saturday, or Sunday of opening weekend.
[2] A wide released movie is any movie that ever shows on 600 or more screens.

	Selected Instruments		*Oracle Estimator*	
Number of Instruments	F-stat.	Coverage Prob.	F-stat.	Coverage Prob.
One Instrument	12.539	0.302	4.911	0.904
Two Instruments	11.185	0.150	5.040	0.830
Three Instruments	10.060	0.070	4.820	0.810

Table 6.1: Comparasions of first-stage F-statistics and 95% confidence interval coverage Probability using selected and oracle instruments

To instrument for sales on opening weekend the authors employ a vector of nationally aggregated weather measures. These weather measures include the proportion of movie theaters experiencing maximum temperatures in 5° Fahrenheit bins on the interval $[10°, 100°]$, the proportion of movie theaters experiencing precipitation levels in 0.25 inch per hour increments on the interval $[0, 1.5]$, and the proportions of theaters experiencing any type of snow and of theaters experiencing any type of rain. Since unusually poor weather may cause people to substitute away from outdoor activities and into watching a movie, these measures provide a source of exogenous variation in opening weekend sales that can be used to identify the effect of social spillovers.

Putting together the nationally aggregated weather measures leaves Gilchrist and Sands (2016) with 48 linearly independent instrumental variables. To handle the large number of instruments, the authors employ a post-LASSO estimate of the first stage (Belloni et al., 2012); they set the first-stage penalty parameter so that the number of instrument selected is one, two, or three. The resulting first-stage F-statistics using the selected instrument(s), 38.80, 25.86, and 20.95, respectively, seem to indicate strong identification. However, the first-stage F-statistic on the full set of instrumental variables is only 3.80. Moreover, since the LASSO objective is an ℓ_1 penalized version of the OLS loss, using the variables selected by LASSO may mechanically lead to higher F-statistics even if the underlying relationship between the instruments and the endogenous variables is weak.

Table 6.1 provides evidence from a simple simulation experiment to demonstrate this. For the simulation experiment I generate an iid sample of size $n = 1000$. For each $i \in [n]$, I generate 10 mutually independent instruments $Z_{ki} \sim N(0, 1)$ for $1 \leq k \leq 10$. The endogenous variable is generated to only have a weak relationship with the instruments, $X_i = \frac{2}{\sqrt{n}} \sum_{\ell=1}^{10} Z_{\ell i} + v_i$, and the outcome is generated $Y_i = X_i + \epsilon_i$ where (ϵ_i, v_i) are independent standard normals. From the initial set of 10 instrumental variables I generate an additional 55 technical instruments by squaring and taking all interactions between variables in the initial set. These generated instruments are correlated with the initial instruments but do not directly enter the first stage.

I then set the LASSO penalty so that only a certain number of instruments are chosen and report the resulting average first stage F-statistics and 95% confidence interval coverage over one thousand simulations. As a comparasion I also report the average first-stage F-statistics and 95% confidence interval coverage from the oracle estimator, which only uses the relevant 10 initial instruments. Despite the fact that the first-stage F-statstic on selected instruments is more than double the first-stage F-statistic using the oracle first stage estimator, the coverage rate of 95% confidence intervals based on LASSO selected instruments is significantly degraded compared to both the nominal coverage probability and the coverage probability using the oracle first-stage estimator.

Given a lack of clarity on the strength of identification, I seek to validate the results of Gilchrist and Sands (2016) using the weak identification testing procedures proposed in this paper. The setting is particularly suitable for weak IV testing using the jackknife K-statistic. With 48 instruments and a sample size of 1671, $d_z^3 = 110{,}592 \gg n$, making the tests of Moreira (2003,

2009), Kleibergen (2005), and Andrews (2016) inapplicable. On the other hand, it is unclear whether asymptotic approximations based on $d_z \to \infty$ will accurately describe the finite-sample distribution of test statistics with 48 instruments. Moreover, since fluctuations in movie theater attendance seem to be largely driven by either particularly cold or particularly hot weather (see Figure 4 in Gilchrist and Sands (2016)), the nuisance parameter $\rho(z_i)$ is plausibly approximately sparse.

Tables 6.2 reports the 95% confidence intervals for $\beta_1, \ldots, \beta_7$ generated by weak-instrument robust confidence intervals for three sets of instruments: the first is the initial set of 48 instruments in Gilchrist and Sands (2016), the second set includes only the temperature instruments for $d_z = 36$, and the final includes the initial instruments as well as all interactions between the temperature instruments and the remaining instruments for $d_z = 524$. For reference, I also provide point estimates and standard errors for $\beta_1, \ldots, \beta_7$ from Gilchrist and Sands (2016), Table 2. To facilitate comparison, these point estimates and standard errors come from a specification that uses all the instruments in the first stage of a 2SLS procedure.

I compare the 95% confidence intervals based on the jackknife K-statistic to those based on the jackknife Lagrange-Multiplier (JLM) statistic of Matsushita and Otsu (2022) and the sup-score statistic of Belloni et al. (2012). Confidence intervals based on the jackknife AR statistic of Mikusheva and Sun (2021) are not reported as they were empty for all specifications, a result that could indicate misspecification of the linear model or be related to a tendency of the jackknife AR test to over reject documented in Section 7. Similarly, confidence intervals based on the thresholding statistic, implemented as recommended in Section 4, are also not reported as they always align with the those of jackknife K-statistic.

While the Gilchrist and Sands (2016) point estimates are always in the 95% confidence intervals generated by the $JK(\beta_0)$ and JLM tests, the confidence intervals from the identification-robust procedures are somewhat wider than those generated with the 2SLS standard errors. Interestingly, the confidence intervals from inverting the jackknife K-test tend to be quite similar to the confidence intervals from the JLM test. This is surprising given the distinct forms of the $JK(\beta_0)$ and the JLM test statistics.

For many of the parameters of interest, the confidence interval based on the sup-score statistic are either empty or nearly empty. As with the empty jackknife AR confidence intervals, this could suggest misspecification of the linear IV model. However, in specifications where the sup-score confidence interval is not empty or nearly empty, the sup-score confidence interval is always wider than that of the jackknife K-statistic. For example, when using the original instrument set the sup-score confidence interval for the cumulative effect of social spillovers, β_7, is nearly 40% larger than that of the jackknife K-statistic. This suggests the jackknife K-test may be more powerful in this setting and is aligned with the fact that the combination test reccomends using a test based on the jackknife K-statistic.

Parameter	β_2	β_3	β_4	β_5	β_6	β_7
Estimate (s.e.)	0.475 (0.024)	0.269 (0.023)	0.164 (0.017)	0.121 (0.013)	0.093 (0.010)	1.222 (0.074)
	Initial instrument set, $d_z = 48$					
$JK(\beta_0)$	$[0.436, 0.557]$	$[0.227, 0.334]$	$[0.134, 0.214]$	$[0.100, 0.167]$	$[0.080, 0.134]$	$[1.003, 1.391]$
$S(\beta_0)$	$\emptyset$	$[0.294, 0.334]$	$[0.087, 0.094]$	$\emptyset$	$\emptyset$	$[0.990, 1.518]$
JLM	$[0.436, 0.557]$	$[0.227, 0.334]$	$[0.134, 0.214]$	$[0.107, 0.167]$	$[0.087, 0.134]$	$[1.010, 1.384]$
	Temperature instruments only, $d_z = 36$					
$JK(\beta_0)$	$[0.449, 0.597]$	$[0.255, 0.389]$	$[0.148, 0.248]$	$[0.114, 0.194]$	$[0.094, 0.154]$	$[1.086, 1.555]$
$S(\beta_0)$	$\emptyset$	$[0.302, 0.329]$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$
JLM	$[0.449, 0.597]$	$[0.255, 0.389]$	$[0.154, 0.248]$	$[0.114, 0.194]$	$[0.094, 0.154]$	$[1.092, 1.555]$
	Initial instruments plus all interactions with temp. instruments, $d_z = 524$					
$JK(\beta_0)$	$[0.443, 0.604]$	$[0.215, 0.342]$	$[0.094, 0.228]$	$[0.087, 0.154]$	$[0.054, 0.121]$	$[0.916, 1.435]$
$S(\beta_0)$	$[0.416, 0.477]$	$\emptyset$	$\emptyset$	$[0.034, 0.121]$	$[0.121, 0.208]$	$[0.918, 1.562]$
JLM	$[0.463, 0.497]$	$[0.268, 0.282]$	$[0.161, 0.174]$	$[0.101, 0.107]$	$[0.063, 0.084]$	$[1.059, 1.137]$

Table 6.2: 95% Confidence Intervals in the data of Gilchrist and Sands (2016).

7. Simulation Study

In this simulation study, I examine the performance of tests based on the $JK(\beta_0)$ statistic and compare it with that of other tests that may be used in settings where the number of instruments is nonnegligible as a fraction of sample size. I consider a reduced-form data-generating process (DGP) similar to that of Matsushita and Otsu (2022). The outcome variable, y_i, and endogenous variable, x_i, are generated according to

$$\begin{aligned} y_i &= x_i + \epsilon_i \\ x_i &= \Pi_i + v_i \end{aligned} \tag{7.1}$$

where $\Pi_i = \frac{1}{r_n}\sum_{k=1}^{5} \frac{3}{4}\bar{z}_{ki} + \frac{1}{4}\bar{z}_{ki}^2 + \frac{1}{4}\bar{z}_{ki}^3$ is a transformation of an initial set of instruments $\bar{z}_i \in \mathbb{R}^{10}$ generated as described below. The value of r_n varies depending on the strength of identification considered; for strong identification, $r_n = 1$, while under weak identification, $r_n = 1/\sqrt{n}$. To model heteroskedasticity, the errors (e_i, v_i) are generated $\epsilon_i = (1 + \varrho_1(\bar{z}_{1i}^2 + \bar{z}_{2i}^2 + \bar{z}_{2i}\bar{z}_{3i}))e_{1i}$, and $v_i = \varrho_2(1 + \bar{z}_{1i})\epsilon_i + (1 - \varrho_2)^2 e_{2i}$ where e_{1i} and e_{2i} are generated independently of each other and other variables in the model according to a Laplace distribution with location parameter $\mu = 0$ and scale parameter $b = 1$. Since jackknife K-statistic has an exact χ^2 distribution when the errors are jointly Gaussian and $\rho(z_i)$ is known, I purposefully avoid normally distributed errors to investigate the quality of asymptotic approximations. The parameters ϱ_1 and ϱ_2 control the degree of heteroskedasticity and endogeneity, respectively.

In addition to considering the behavior of tests under both weak and strong identification, I examine the size of the test under three different instrument regimes. In all three regimes, I begin with an initial set of instruments $\bar{z}_i = (\bar{z}_{1i}, \ldots, \bar{z}_{10i})'$ generated independently across indices according to a multivariate Gaussian distribution with Toeplitz covariance structure, $\text{Cov}(\bar{z}_{\ell i}, \bar{z}_{ki}) = 2^{-|\ell - k|}$. In the first regime, the instruments only include the initial set, $z_i = \bar{z}_i$ so that $d_z = 10$. In the second regime, the full set of instruments z_i additionally includes all quadratic and cubic terms, $(z_{\ell i}^2, z_{\ell i}^3)$, $\ell = 1, \ldots, 10$ so that in total $d_z = 30$. In the third regime, the full set of instrument includes the initial set of instruments, $\bar{z}_i$, and all quadratic terms (10 additional terms) and interactions of the initial set of instruments ($\binom{10}{2} = 45$ additional terms), so that in total $d_z = 65$. Under each regime, the full set of instruments is passed to the test statistics with no indication about which instruments correspond to the initial set, and thus no indication about which instruments are relevant to the DGP.

I compare the simulated size of the jackknife K test and to the performance of the sup-score test, $S(\beta_0)$, of Belloni et al. (2012), the thresholding test introduced in Section 4.2, the standard

DGP				Testing Procedure						
n	d_z	ϱ_1	ϱ_2	$JK(\beta_0)$	$S(\beta_0)$	$T(\beta_0;\tau_{0.3})$	$T(\beta_0;\tau_{0.75})$	A.Rbn.	JAR	JLM
200	10	0.2	0.3	0.0516	0.0352	0.0406	0.0406	0.0296	0.0766	0.0502
		0.2	0.6	0.0542	0.0306	0.0442	0.0384	0.0258	0.0748	0.0400
		0.5	0.3	0.0470	0.0338	0.0416	0.0418	0.0238	0.0784	0.0460
		0.5	0.6	0.0506	0.0350	0.0416	0.0390	0.0280	0.0676	0.0384
	30	0.2	0.3	0.0570	0.0124	0.0422	0.0200	0.0088	0.1000	0.0382
		0.2	0.6	0.0564	0.0126	0.0408	0.0208	0.0124	0.0962	0.0322
		0.5	0.3	0.0498	0.0100	0.0366	0.0190	0.0096	0.1090	0.0318
		0.5	0.6	0.0562	0.0118	0.0420	0.0216	0.0088	0.1104	0.0292
	65	0.2	0.3	0.0542	0.0316	0.0428	0.0370	0.0314	0.0764	0.0420
		0.2	0.6	0.0532	0.0366	0.0418	0.0398	0.0250	0.0780	0.0376
		0.5	0.3	0.0474	0.0308	0.0388	0.0362	0.0244	0.0748	0.0354
		0.5	0.6	0.0484	0.0324	0.0366	0.0388	0.0282	0.0708	0.0402
500	10	0.2	0.3	0.0590	0.0468	0.0478	0.0516	0.0376	0.0652	0.0452
		0.2	0.6	0.0530	0.0420	0.0460	0.0466	0.0366	0.0692	0.0434
		0.5	0.3	0.0496	0.0370	0.0408	0.0368	0.0338	0.0710	0.0464
		0.5	0.6	0.0512	0.0426	0.0456	0.0438	0.0334	0.0696	0.0404
	30	0.2	0.3	0.0522	0.0202	0.0386	0.0278	0.0238	0.0818	0.0322
		0.2	0.6	0.0558	0.0208	0.0408	0.0310	0.0266	0.0888	0.0342
		0.5	0.3	0.0554	0.0178	0.0392	0.0280	0.0174	0.0940	0.0272
		0.5	0.6	0.0570	0.0156	0.0426	0.0236	0.0206	0.0984	0.0280
	65	0.2	0.3	0.0542	0.0372	0.0434	0.0432	0.0384	0.0754	0.0464
		0.2	0.6	0.0584	0.0442	0.0482	0.0470	0.0334	0.0676	0.0438
		0.5	0.3	0.0614	0.0460	0.0504	0.0496	0.0316	0.0708	0.0434
		0.5	0.6	0.0526	0.0378	0.0434	0.0420	0.0298	0.0692	0.0358

Table 7.1: Simulated Size of Identification and Heteroskedasticity Robust Tests under Weak Identification. Each DGP is simulated 5000 times.

Anderson-Rubin (A.Rbn.) test of Anderson and Rubin (1949) and Staiger and Stock (1997), the jackknife AR test (JAR) of Crudu et al. (2021) and Mikusheva and Sun (2021), and the jackknife LM test (JLM) of Matsushita and Otsu (2022). To estimate the parameter $\rho(z_i)$, I implement the ℓ_1-penalized procedure of (2.2) via the `glmnet` package in `R` (Friedman et al., 2010). The penalty parameter λ is selected via tenfold cross-validation. I use the full vector of instruments as the basis to approximate $\rho(z_i)$. For the jackknife AR test I use cross-fit estimates of test statistic variances proposed and shown to improve power by Mikusheva and Sun (2021). Critical values of the sup-score and conditioning statistic are simulated with the procedures described in Section 4 with 1000 bootstrap replications. For the combination test cutoff, I report results using two different quantiles of the conditioning statistic under the assumption that $\mathbb{E}[\widehat{\Pi}_i^l] = 0$ for all $i \in [n]$; $\tau_{0.3}$ corresponding to the 30^{th} quantile and $\tau_{0.75}$ corresponding to the 75^{th} quantile.

Tables 7.1 and 7.2 report the simulated size for all tests under weak and strong identification, respectively. One can see that the $JK(\beta_0)$ statistic has nearly exact size in almost all the setups considered. In contrast, the jackknife AR test seems to overreject in nearly all the simulation setups considered. This is also the case in the simulation study of Matsushita and Otsu (2022) and so may be an artifact of the similarity of my simulation design to theirs.

The sup-score, jackknife AR, and jackknife LM test all seem to have particularly poor perfor-

DGP				Testing Procedure						
n	d_z	ϱ_1	ϱ_2	$JK(\beta_0)$	$S(\beta_0)$	$T(\beta_0;\tau_{0.3})$	$T(\beta_0;\tau_{0.75})$	A.Rbn.	JAR	JLM
200	10	0.2	0.3	0.0474	0.0420	0.0474	0.0468	0.0308	0.0728	0.0424
		0.2	0.6	0.0512	0.0386	0.0512	0.0506	0.0304	0.0764	0.0378
		0.5	0.3	0.0416	0.0318	0.0414	0.0414	0.0248	0.0794	0.0428
		0.5	0.6	0.0446	0.0342	0.0446	0.0442	0.0244	0.0806	0.0384
	30	0.2	0.3	0.0482	0.0122	0.0448	0.0264	0.0110	0.1048	0.0370
		0.2	0.6	0.0498	0.0120	0.0480	0.0312	0.0118	0.0980	0.0378
		0.5	0.3	0.0456	0.0126	0.0410	0.0262	0.0082	0.1146	0.0268
		0.5	0.6	0.0482	0.0110	0.0474	0.0308	0.0094	0.1090	0.0302
	65	0.2	0.3	0.0528	0.0380	0.0526	0.0510	0.0276	0.0696	0.0460
		0.2	0.6	0.0464	0.0360	0.0464	0.0468	0.0302	0.0728	0.0416
		0.5	0.3	0.0482	0.0298	0.0480	0.0466	0.0246	0.0738	0.0412
		0.5	0.6	0.0396	0.0320	0.0390	0.0386	0.0258	0.0748	0.0356
500	10	0.2	0.3	0.0524	0.0444	0.0524	0.0524	0.0394	0.0684	0.0472
		0.2	0.6	0.0476	0.0430	0.0476	0.0476	0.0400	0.0644	0.0490
		0.5	0.3	0.0434	0.0410	0.0434	0.0434	0.0340	0.0702	0.0404
		0.5	0.6	0.0448	0.0382	0.0448	0.0448	0.0350	0.0736	0.0432
	30	0.2	0.3	0.0502	0.0214	0.0502	0.0498	0.0240	0.0854	0.0368
		0.2	0.6	0.0522	0.0208	0.0522	0.0524	0.0224	0.0858	0.0392
		0.5	0.3	0.0456	0.0202	0.0456	0.0434	0.0220	0.0918	0.0264
		0.5	0.6	0.0500	0.0186	0.0500	0.0498	0.0204	0.0924	0.0268
	65	0.2	0.3	0.0490	0.0426	0.0490	0.0490	0.0350	0.0742	0.0472
		0.2	0.6	0.0522	0.0458	0.0522	0.0522	0.0436	0.0652	0.0442
		0.5	0.3	0.0542	0.0476	0.0542	0.0542	0.0294	0.0712	0.0446
		0.5	0.6	0.0438	0.0420	0.0438	0.0438	0.0306	0.0666	0.0500

Table 7.2: Simulated Size of Identification and Heteroskedasticity Robust Tests under Strong Identification. Each DGP is simulated 5000 times.

mance under both weak and strong identification when $d_z = 30$. This is the setup with the most correlation between the instruments. While tests based on the jackknife AR statistic can have a simulated size that is nearly double the nominal size in this setting, both the sup-score and and jackknife LM tests appear to be conservative. Notably, the size properties of the sup-score test do seem to improve under both weak and strong identification when the sample size increases from $n = 200$ to $n = 500$. In contrast, the size properties of the jackknife LM test do not seem to improve when the sample size increases and indeed worsen for three out of the four DGPs considered under both weak and strong identification. This suggests that the requirement of $d_z \to \infty$ is important for the quality of finite-sample approximation by its limiting distribution.

The thresholding test seems to control size in all the setups considered. However, under weak identification the thresholding test appears to inherit the conservative nature of the sup-score test, even in the "large" sample size regime of $n = 500$. This is not the case under strong identification, suggesting that the thresholding-test is choosing to run tests based on the $JK(\beta_0)$ with high probability in this regime. This behavior is similar to the conditional combination tests of Moreira (2003), Andrews (2016) which place heigher weight on the K-statistic under strong identification.

Figure 7.1 plots calibrated local power curves under an intermediate identification strength where the first stage is in a $n^{-1/3}$ neighborhood of zero, $d_z = 65$, $\varrho_1 \in \{0.2, 0.5\}$ and $\varrho_2 \in$

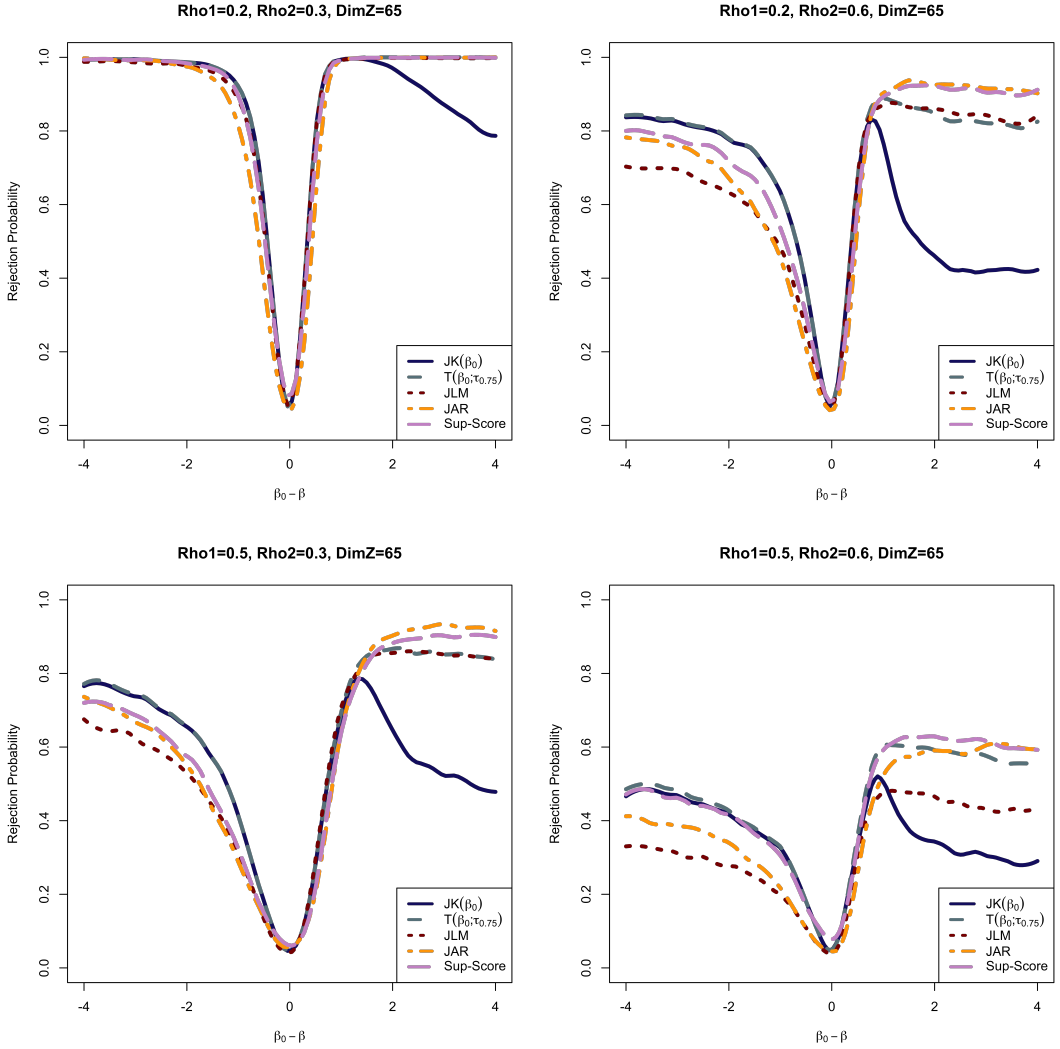

Figure 7.1: Calibrated Local Power Curves under Intermediate Identification Strength and 65 Instruments. Sample size is 500 and rejection probability is calculated on a grid of 100 $(\beta_0 - \beta)$ points between -4 and 4. At each point the DGP is simulated 2000 times.

$\{0.3, 0.6\}$. The critical value of each test is set to simulated 95^{th} quantile of the distribution of the corresponding test-statistic under H_0. I compare the calibrated local power curves of the $JK(\beta_0)$ test, the combination test with cutoff $\tau_{0.75}$, the jackknife AR test, the Jackknife LM test, and the sup-score test. The jackknife K-test has stronger power than the jackknife AR, jackknife LM, and sup-score tests in local neighborhoods of the null as well as for negative values of $(\beta_0 - \beta)$. For values of $(\beta_0 - \beta)$ larger than 1.5, tests based on the jackknife K-statistic suffer from a loss of power as described in Section 4. This power decline is largerly ameliorated by combining the jackknife K-statistic with the sup-score statistic and the thresholding test has good power properties over all alternatives considered. However, tests based on the jackknife AR or jackknife LM statistic still provide better power than the threshholding test for very positive values of $(\beta_0 - \beta)$.

To investigate the effect of correlated instruments on power properties in a setting with plausibly many instruments, I additionally examine local power under a fourth instrument regime. This setup adds the ten cubic terms $z_{\ell i}^3, \ell \in [10]$ to the interactions and quadratic terms of the third instrument regime for a total of 75 instruments. Figure 7.2 plots calibrated local power curves under this fourth instrument regime. While all tests have lower power in this regime than

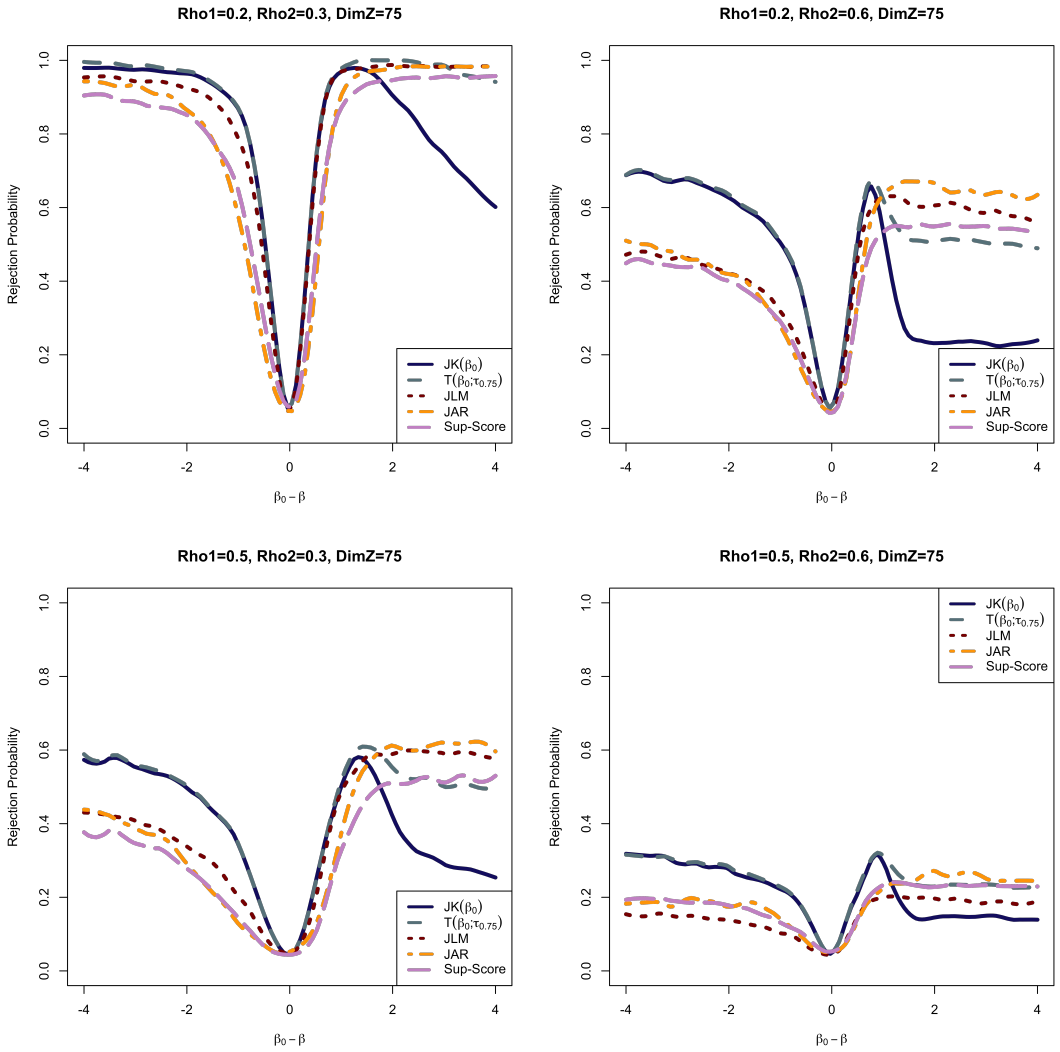

Figure 7.2: Calibrated Local Power Curves under Intermediate identification Strength and 75 Instruments. Sample size is 500 and rejection power is calculated on a grid of 100 $(\beta_0 - \beta)$ points between -4 and 4. At each point the DGP is simulated 2000 times.

in the regime considered in Figure 7.1, the many instrument jackknife AR and jackknife LM tests appear to face a steeper power decline than tests based on the jackknife K-statistic or the thresholding statistic.

These results should not be interpreted as critiques of the benchmark testing procedures of Anderson and Rubin (1949), Staiger and Stock (1997), Belloni et al. (2012), Crudu et al. (2021), Mikusheva and Sun (2021), and Matsushita and Otsu (2022), whose work I rely on and was inspired by.

A. Appendix

I provide a proof of the main results in Section 3, concluding with a proof of Theorem 3.1. The proof of Proposition 3.1 as well as of supporting lemmas can be found in the online appendix. The proofs for Section 4 follow a similar strategy, but combines this proof strategy with that of Chernozhukov et al. (2013) to establish joint Gaussian approximation of the jackknife K-statisic, sup-score statistics, and conditioning statistics. The full proofs for Sections 4 and 5 can be found in Navjeevan (2023). .

Before proceeding, we will introduce some notation. Let $\tilde{H} = s_n H$ and $\tilde{h}_{ij} = s_n h_{ij}$, where s_n is as in Assumption 3.2. WLOG let $\tilde{h}_{ii} = h_{ii} = 0$ and define

$$N := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \epsilon_i(\beta_0) \sum_{j=1}^{n} \tilde{h}_{ij} r_j \qquad \tilde{N} := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{\epsilon}_i(\beta_0) \sum_{j=1}^{n} \tilde{h}_{ij} \tilde{r}_j$$

$$D := \frac{1}{n} \sum_{i=1}^{n} \epsilon_i^2(\beta_0) \Big(\sum_{j=1}^{n} \tilde{h}_{ij} r_j \Big)^2 \qquad \tilde{D} := \frac{1}{n} \sum_{i=1}^{n} \kappa_i^2(\beta_0) \Big(\sum_{j=1}^{n} \tilde{h}_{ij} \tilde{r}_j \Big)^2$$

where $(\tilde{\epsilon}_i(\beta_0), \tilde{r}_i)$ are jointly Gaussian with the same mean and covariance matrix as $(\epsilon_i(\beta_0), r_i)$ and $\kappa_i^2(\beta_0) = \mathbb{E}[\epsilon_i^2(\beta_0)]$. In addition define the decomposed statistics

$$JK^a := N^2 - aD \;\text{ and }\; J\tilde{K}^a := \tilde{N}^2 - a\tilde{D}$$

Noting that $\{JK_I(\beta_0) \le a\} = \{JK^a \le 0\}$ and $\{JK_G(\beta_0) \le a\} \overset{\text{a.s}}{=} \{J\tilde{K}^a \le 0\}$.

Proof of Lemma 3.1. For any $a \in \mathbb{R}$, begin by defining

$$N_{-i} := \frac{1}{\sqrt{n}} \sum_{j \neq i} \dot{\epsilon}_j(\beta_0) \sum_{\ell \neq i} \tilde{h}_{j\ell} \dot{r}_\ell \qquad D_{-i} := \frac{1}{n} \sum_{j \neq i} \ddot{\epsilon}_j^2(\beta_0) \Big(\sum_{\ell \neq i} \tilde{h}_{j\ell} \dot{r}_\ell \Big)^2$$

$$JK_{-i} := N_{-i}^2 - aD_{-i}$$

where for each $\ell \in [n]$, $\dot{\epsilon}_\ell(\beta_0)$ is equal to $\epsilon_\ell(\beta_0)$ if $\ell > i$ and $\tilde{\epsilon}_\ell(\beta_0)$ if $\ell < i$, $\dot{r}_\ell$ is equal to r_ℓ if $\ell > i$ and $\tilde{r}_\ell$ if $\ell < i$, and $\ddot{\epsilon}_\ell^2(\beta_0)$ is equal to $\kappa_\ell^2(\beta_0)$ if $\ell < i$ and $\epsilon_\ell^2(\beta_0)$ if $\ell > i$. While the definitions of $\dot{\epsilon}_\ell, \dot{r}_\ell$, and $\ddot{\epsilon}_\ell$ depend on i because we will be considering only one deviation at a time, we will supress this dependence to simplify notation.

Next, define the one-step deviations

$$\Delta_{1i} := \epsilon_i(\beta_0) \sum_{j=1}^{n} \tilde{h}_{ij} \dot{r}_j + r_i \sum_{j=1}^{n} \tilde{h}_{ji} \dot{\epsilon}_j(\beta_0)$$

$$\tilde{\Delta}_{1i} := \tilde{\epsilon}_i(\beta_0) \sum_{j=1}^{n} \tilde{h}_{ij} \dot{r}_j + \tilde{r}_i \sum_{j=1}^{n} \tilde{h}_{ji} \dot{\epsilon}_j(\beta_0)$$

$$\Delta_{2i} := \underbrace{a\epsilon_i^2(\beta_0)\Big(\sum_{j=1}^{n} \tilde{h}_{ij} \dot{r}_j\Big)^2 + a r_i^2 \sum_{j=1}^{n} \tilde{h}_{ji}^2 \ddot{\epsilon}_j^2(\beta_0)}_{\Delta_{2i}^a} + \underbrace{2ar_i \sum_{j=1}^{n} \ddot{\epsilon}_j^2(\beta_0) \sum_{\ell \neq i} \tilde{h}_{j\ell} \tilde{h}_{ji} \dot{r}_\ell}_{\Delta_{2i}^b}$$

$$\tilde{\Delta}_{2i} := \underbrace{a\kappa_i^2(\beta_0)\Big(\sum_{j=1}^{n} \tilde{h}_{ij} \dot{r}_j\Big)^2 + a \tilde{r}_i^2 \sum_{j=1}^{n} \tilde{h}_{ji}^2 \ddot{\epsilon}_j^2(\beta_0)}_{\Delta_{2i}^a} + \underbrace{2a\tilde{r}_i \sum_{j=1}^{n} \ddot{\epsilon}_j^2(\beta_0) \sum_{\ell \neq i} \tilde{h}_{j\ell} \tilde{h}_{ji} \dot{r}_\ell}_{\tilde{\Delta}_{2i}^b}$$

Using the one-step deviations, write the difference $\mathbb{E}[\varphi(K^a) - \varphi(\tilde{K}^a)]$ as a telescoping sum, one by one replacing $(\Delta_{1i}, \Delta_{2i})$ with $(\tilde{\Delta}_{1i}, \tilde{\Delta}_{2i})$ in the expressions of $JK^a = N^2 - aD$ until we arrive at

$J\tilde{K}^a = \tilde{N}^2 - a\tilde{D}$.

$$\begin{aligned}\mathbb{E}[\varphi(JK^a) - \varphi(J\tilde{K}^a)] = \sum_{i=1}^{n} &\mathbb{E}[\varphi(JK_{-i} + n^{-1/2}N_{-i}\Delta_{1i} + n^{-1}\Delta_{1i}^2 - n^{-1}\Delta_{2i})] \\ &- \mathbb{E}[\varphi(JK_{-i} + n^{-1/2}N_{-i}\tilde{\Delta}_{1i} + n^{-1}\tilde{\Delta}_{1i}^2 - n^{-1}\tilde{\Delta}_{2i})]\end{aligned} \tag{A.1}$$

Via a second-order Taylor expansion, we can write each term inside the summand

$$\begin{aligned}\mathbb{E}[\text{Term}_i] = &\mathbb{E}[\varphi'(JK_{-i})\{2n^{-1/2}N_{-i}(\Delta_{1i} - \tilde{\Delta}_{1i}) + n^{-1}(\Delta_{1i}^2 - \Delta_{1i}^2) - n^{-1}(\Delta_{2i} - \tilde{\Delta}_{2i})\}] \\ &+ \mathbb{E}[\varphi''(JK_{-i})\{4n^{-1}N_{-i}^2(\Delta_{1i}^2 - \tilde{\Delta}_{1i}^2) + n^{-2}(\Delta_{1i}^4 - \tilde{\Delta}_{1i}^4) - n^{-2}(\Delta_{2i}^2 - \Delta_{2i}^2)\}] \\ &+ \mathbb{E}[\varphi''(JK_{-i})\{4n^{-3/2}N_{-i}(\Delta_{1i}^3 - \tilde{\Delta}_{1i}^3) + 4n^{-3/2}N_{-i}(\Delta_{1i}\Delta_{2i} - \tilde{\Delta}_{1i}\tilde{\Delta}_{2i})\}] \\ &+ \mathbb{E}[\varphi''(JK_{-i})\{2n^{-2}(\Delta_{1i}^2\Delta_{2i} - \tilde{\Delta}_{1i}^2\tilde{\Delta}_{2i})\}] + R_i + \tilde{R}_i\end{aligned}$$

where R_i and $\tilde{R}_i$ are remainder terms to be examined later. Let $\mathcal{F}_{-i}$ denote the sigma algebra generated by all random variables whose index is not equal to i. Since (a) for each $i \in [n]$ the mean and covariance matrix of $(\epsilon_i(\beta_0), r_i)$ is the same as the mean and covariance matrix of $(\tilde{\epsilon}_i(\beta_0), \tilde{r}_i)$, (b) $\mathbb{E}[\epsilon_i^2(\beta_0)] = \kappa_i^2(\beta_0)$, and (c) random variables are independent across indices, we have that

$$\begin{aligned}\mathbb{E}[\Delta_{1i} - \tilde{\Delta}_{1i}|\mathcal{F}_{-i}] &= \mathbb{E}[\Delta_{1i}^2 - \tilde{\Delta}_{1i}^2|\mathcal{F}_{-i}] = \mathbb{E}[\Delta_{2i} - \tilde{\Delta}_{2i}|\mathcal{F}_{-i}] \\ &= \mathbb{E}[\Delta_{2i}^b - \tilde{\Delta}_{2i}^b|\mathcal{F}_{-i}] = \mathbb{E}[\Delta_{1i}\Delta_{2i}^b - \tilde{\Delta}_{1i}\tilde{\Delta}_{2i}^b|\mathcal{F}_{-i}] = 0\end{aligned} \tag{A.2}$$

Using this we can simplify the prior display

$$\begin{aligned}\mathbb{E}[\text{Term}_i] = &\underbrace{n^{-2}\mathbb{E}[\varphi''(JK_{-i})(\Delta_{1i}^4 - \Delta_{1i}^4)]}_{\mathbf{A}_i} - \underbrace{n^{-2}\mathbb{E}[\varphi''(JK_{-i})((\Delta_{2i}^a)^2 - (\tilde{\Delta}_{2i}^a)^2)]}_{\mathbf{B}_i} \\ &- \underbrace{2n^{-2}\mathbb{E}[\varphi''(JK_{-i})(\Delta_{2i}^a\Delta_{2i}^b - \tilde{\Delta}_{2i}^a\tilde{\Delta}_{2i}^b)]}_{\mathbf{C}_i} + \underbrace{4n^{-3/2}\mathbb{E}[\varphi''(JK_{-i})N_{-i}(\Delta_{1i}^3 - \tilde{\Delta}_{1i}^3)]}_{\mathbf{D}_i} \\ &+ \underbrace{4n^{-3/2}\mathbb{E}[\varphi''(JK_{-i})N_{-i}(\Delta_{1i}\Delta_{2i}^a - \tilde{\Delta}_{1i}\tilde{\Delta}_{2i}^a)]}_{\mathbf{E}_i} + \underbrace{2n^{-2}\mathbb{E}[\varphi''(JK_{-i})(\Delta_{1i}^2\Delta_{2i} - \tilde{\Delta}_{1i}^2\tilde{\Delta}_{2i})}_{\mathbf{F}_i} \\ &+ R_i + \tilde{R}_i\end{aligned}$$

where for some $J\bar{K}_{1i}$ and $J\bar{K}_{2i}$ we can express

$$\begin{aligned}R_i &= \mathbb{E}[\varphi'''(J\bar{K}_{1i})\{n^{-1/2}N_{-i}\Delta_{1i} + n^{-1}\Delta_{1i}^2 + n^{-1}\Delta_{2i}\}^3] \\ \tilde{R}_i &= \mathbb{E}[\varphi'''(J\bar{K}_{2i})\{n^{-1/2}N_{-i}\tilde{\Delta}_{1i} + n^{-1}\tilde{\Delta}_{1i}^2 + n^{-1}\tilde{\Delta}_{2i}\}^3]\end{aligned}$$

Applications of Lemmas A.5 and A.6, Cauchy-Schwarz, and generalized Holder's inequalty, will allow us to bound for a fixed constant M that depends only on c,

$$\begin{aligned}|\mathbf{A}_i| &\le \frac{M}{n^2}L_2(\varphi) & |\mathbf{B}_i| &\le \frac{Ma^2}{n^2}L_2(\varphi) & |\mathbf{C}_i| &\le \frac{Ma^2}{n^{3/2}}L_2(\varphi) \\ |\mathbf{D}_i| &\le \frac{M}{n^{3/2}}L_2(\varphi) & |\mathbf{E}_i| &\le \frac{M(a \vee 1)}{n^{3/2}}L_2(\varphi) & |\mathbf{F}_i| &\le \frac{Ma^3}{n^{3/2}}L_2(\varphi)\end{aligned}$$

and

$$|R_i| + |\tilde{R}_i| \le \frac{M}{n^{3/2}}L_3(\varphi) + \frac{Ma^3}{n^3}L_3(\varphi)$$

Combining these bounds and summing over n gives that there is a constant M that depends only on the constant c such that:

$$|\mathbb{E}[\varphi(JK^a) - \varphi(J\tilde{K}^a)]| \leq \frac{M(a^3 \vee 1)}{\sqrt{n}}(L_2(\varphi) + L_3(\varphi)) \tag{A.3}$$

Next, by Assumption 3.1, we know that $\kappa_i^2(\beta_0) \in [c^{-1}, c]$ for all $i = 1, \ldots, n$ so that $\tilde{D} \geq \frac{c^{-1}}{n}\sum_{i=1}^n(\sum_{j=1}^n \tilde{h}_{ij}r_j)^2$. Then for any sequence $\delta_n \searrow 0$;

$$\begin{aligned}\Pr(\tilde{D} \leq \delta_n) &\leq \Pr\Big(\frac{1}{cn}\sum_{i=1}^n\Big(\sum_{j=1}^n \tilde{h}_{ij}\tilde{r}_j\Big)^2 \leq \delta_n\Big) \\ &= \Pr\big(\|\tilde{r}'\bar{H}^{1/2}\|^2 \leq \delta_n\big)\end{aligned} \tag{A.4}$$

where $\tilde{r} := (\tilde{r}_1, \ldots, \tilde{r}_n)' \in \mathbb{R}^n$ and $\bar{H} := \frac{1}{cn}\tilde{H}\tilde{H}' \in \mathbb{R}^{n\times n}$. $\bar{H}$ is symmetric and positive semidefinite so we can take $\bar{H}^{1/2}$ to be its symmetric square root, which will also be symmetric and positive semidefinite (and thus not necessarily equal to $\sqrt{\frac{c}{n}}\tilde{H}$). I provide two bounds on (A.4), the first of which corresponds to the strong identification setting while the second corresponds to weak identification.

First Bound. Since $\delta_n \searrow 0$ we will eventually have that $\delta_n < c^{-1}/2$. When this happens we can bound using Chebyshev's inequality and $c^{-1} < \mathbb{E}[r'\bar{H}r] < c$:

$$\begin{aligned}\Pr(\tilde{r}'\bar{H}\tilde{r} \leq \delta_n) &= \Pr(\tilde{r}'\bar{H}\tilde{r} - \mathbb{E}[\tilde{r}'\bar{H}\tilde{r}] \leq \delta_n - \mathbb{E}[\tilde{r}'\bar{H}\tilde{r}]) \\ &\leq \Pr(\tilde{r}'\bar{H}\tilde{r} - \mathbb{E}[r'\bar{H}r] \geq \mathbb{E}[\tilde{r}'\bar{H}\tilde{r}] - \delta_n) \\ &\leq \Pr(|\tilde{r}'\bar{H}\tilde{r} - \mathbb{E}[r'\bar{H}r]| \geq \frac{1}{2c}) \\ &\leq 2c\,\mathrm{Var}(r'\bar{H}r)\end{aligned} \tag{A.5}$$

Under strong identification we will expect $\mathrm{Var}(r'\bar{H}r) \to 0$.

Second Bound. For the second bound, we will directly use bounds on the density of Gaussian quadratic forms from Götze et al. (2019). The vector $r'\bar{H}^{1/2}$ is Gaussian with covariance matrix $\Sigma_r = \bar{H}^{1/2}\boldsymbol{R}\bar{H}^{1/2}$ where $\boldsymbol{R} = \mathrm{diag}(\mathrm{Var}(r_1), \ldots, \mathrm{Var}(r_n))$. Let $\Lambda_1 = \sum_{k=1}^n \lambda_k^2(\Sigma_r)$ and $\Lambda_2 = \sum_{k=2}^n \lambda_k^2(\Sigma_r)$. By Assumption 3.2 and Lemma A.4, Λ_2/Λ_1 is bounded away from zero. Using Theorem A.1 we can then bound for some constant $C > 0$

$$\Pr(\|r'H\|^{1/2} \leq \delta_n) \leq C\delta_n\Lambda_1^{-1} \tag{A.6}$$

To combine the bounds in (A.5) and (A.6), first write

$$\mathrm{Var}(\tilde{r}'\bar{H}\tilde{r}) = 2\mathrm{trace}(\boldsymbol{R}\bar{H}\boldsymbol{R}\bar{H}) + 4\mu_r\bar{H}\boldsymbol{R}\bar{H}\mu_r$$

for $\mu_r = \mathbb{E}[r]$. Using the fact that $\bar{H}^{1/2}\boldsymbol{R}\bar{H}^{1/2}$ is symmetric positive definite we can bound:

$$\begin{aligned}\mu_r'\bar{H}\boldsymbol{R}\bar{H}\mu_r &= (\mu_r'\bar{H}^{1/2})'(\bar{H}^{1/2}\boldsymbol{R}\bar{H}^{1/2})(\bar{H}^{1/2}\mu_r) \\ &\leq \lambda_1(\bar{H}^{1/2}\boldsymbol{R}\bar{H}^{1/2})\|\mu_r'\bar{H}^{1/2}\|^2 \\ &= \sqrt{\lambda_1^2(\bar{H}^{1/2}\boldsymbol{R}\bar{H}^{1/2})}\|\mu_r'\bar{H}^{1/2}\|^2 \\ &= \sqrt{\lambda_1(\bar{H}^{1/2}\boldsymbol{R}\bar{H}\boldsymbol{R}\bar{H}^{1/2})}\|\mu_r'\bar{H}^{1/2}\|^2 \\ &\leq \sqrt{\mathrm{trace}(\bar{H}^{1/2}\boldsymbol{R}\bar{H}\boldsymbol{R}\bar{H}^{1/2})}\|\mu_r'\bar{H}^{1/2}\|^2\end{aligned}$$

$$= \sqrt{\text{trace}(\boldsymbol{R}\bar{H}\boldsymbol{R}\bar{H})}\|\mu_r'\bar{H}\|^2 \leq c^2\Lambda_1^{1/2} \tag{A.7}$$

where the first equality uses the symmetric square root of $\bar{H}$, the first inequality comes from Courant-Fischer minmax principle and the third equality uses the fact that the eigenvalues of A^2 are the squares of the eigenvalues of A, for any generic symmetric matrix A. The second inequality comes from the fact that a matrix times its transpose is always positive semidefinite and that for M psd, $\lambda_1(M) \leq \sqrt{\text{trace}(M^2)}$ since the trace is the sum of the (weakly positive) eigenvalues. The final inequality uses $\mu_r'\bar{H}\mu_r = \frac{c}{n}\sum_{i=1}^n(\mathbb{E}[\tilde{\Pi}_i])^2 \leq \frac{c}{n}\sum_{i=1}^n \mathbb{E}[(\tilde{\Pi}_i)^2] \leq c^2$.

Combining (A.5), (A.6), and (A.7) gives us

$$\Pr(\tilde{D} \leq \delta_n) \leq C \min\left\{\Lambda_1 + \Lambda_1^{1/2}, \delta_n\Lambda_1^{-1}\right\} \tag{A.8}$$

Regardless of the behavior of Λ_1, this tends to zero as $\delta_n \to 0$.

Now fix a $\Delta \geq 0$ and consider any $0 < a \leq \Delta$. Let $\tilde{\varphi}(\cdot) : \mathbb{R} \to \mathbb{R}$ be three times continuously differentiable with bounded derivatives up to the third order such that $\tilde{\varphi}(x)$ is 1 if $x \leq 0$, $\tilde{\varphi}(x)$ is decreasing if $x \in (0,1)$, and $\tilde{\varphi}(x)$ is zero if $x \geq 1$. Consider a sequence $\gamma_n \searrow 0$ slowly enough such that $(\gamma_n^{-2} + \gamma_n^{-3})/\sqrt{n} \to 0$ and define $\varphi_n(x) = \tilde{\varphi}(\frac{x}{\gamma_n})$.

By (A.3) we can write for some constant M that depends only on Δ:

$$\begin{aligned}
\Pr(JK_I(\beta_0) \leq a) = \Pr(JK^a \leq 0) &\leq \mathbb{E}[\varphi_n(JK^a)] \\
&\leq \mathbb{E}[\varphi_n(J\tilde{K}^a)] + \frac{M}{\sqrt{n}}(\gamma_n^2 + \gamma_n^{-3}) \\
&\leq \Pr(J\tilde{K}^a \leq 0) + \Pr(0 \leq \tilde{N}^2 - a\tilde{D} \leq \gamma_n) + \frac{M}{\sqrt{n}}(\gamma_n^2 + \gamma_n^{-3})
\end{aligned}$$

Applying Lemma A.1 and $\{J\tilde{K}^a \leq 0\} = \{JK_G(\beta_0) \leq a\}$ gives:

$$\begin{aligned}
&\leq \Pr(JK_G(\beta_0) \leq a) + \underbrace{\Pr(a \leq \tilde{N}^2/\tilde{D} \leq a + \gamma_n^{1/2})}_{\mathbf{A}} \\
&\quad + \underbrace{\Pr(\tilde{D} \leq \gamma_n^{1/2})}_{\mathbf{B}} + \frac{M}{\sqrt{n}}(\gamma_n^{-2} + \gamma_n^{-3})
\end{aligned}$$

By Lemma A.3, we can bound $\mathbf{A} \leq M\gamma_n^{1/2}$ while by Equation (A.8), $\mathbf{B} \leq M\gamma_n^{1/4}$. Since γ_n is chosen such that $\frac{M}{\sqrt{n}}(\gamma_n^{-2} + \gamma_n^{-3}) \to 0$ we can conclude that $\Pr(JK_I(\beta_0) \leq a) \leq \Pr(JK_G(\beta_0) \leq a) + o(1)$. A symmetric argument with $\varphi_n(x) = \tilde{\varphi}(1 - \frac{x}{\gamma_n})$ gives a lower bound so that, in total

$$\Pr(JK_G(\beta_0) \leq a) - o(1) \leq \Pr(JK_I(\beta_0) \leq a) \leq \Pr(JK_G(\beta_0) \leq a) + o(1)$$

where the $o(1)$ is uniform for all $a \leq \Delta$. This yields

$$\sup_{a \leq \Delta}\left|\Pr(JK_I(\beta_0) \leq a) - \Pr(JK_G(\beta_0) \leq a)\right| = o(1) \tag{A.9}$$

Noting that the numerator $JK_G(\beta_0)$ is $O_p(1)$ under Assumption 3.3 while the inverse of the denominator of $JK_G(\beta_0)$ is $O_p(1)$ by (A.8), we can apply Lemma A.2 to conclude. □

Proof of Lemma 3.2. Using the definitions at the top of this appendix, we can write

$$JK_I(\beta_0) - JK(\beta_0) = \frac{2ND\Delta_N + D\Delta_N - N^2\Delta_D}{D^2 + D\Delta_D}$$

Apply Lemma A.6 to see that $N^2 = O_p(1)$ while under Assumption 3.2, $D = O_p(1)$. Thus, $2ND\Delta_n + D\Delta_n - N^2\Delta_D = o_p(1)$. To show that $(D^2 + D\Delta_D)^{-1} = O_p(1)$ it suffices to show that $\Pr(D \leq \delta_n) \to 0$ for any sequence $\delta_n \searrow 0$. To do so, we can follow the same steps as in the proof of Lemma 3.1 to show that $\Pr(D \leq \delta_n) \leq \Pr(\tilde{D} \leq \delta_n) + o(1)$. An application of (A.8) shows that $\Pr(\tilde{D} \leq \delta_n) = o(1)$ allowing us to conclude. □

Proof of Theorem 3.1. By Lemma 3.2 and Lemma A.7 we have that $|JK(\beta_0) - JK_I(\beta_0)| \to_p 0$. Next for any $a \in \mathbb{R}$ and $\epsilon > 0$ we have that $\{JK \leq a\} \subseteq \{JK_I \leq a + \epsilon\} \cup \{|JK_I - JK| > \epsilon\}$; thus, by applying union bound and rearranging we obtain:

$$\begin{aligned}\Pr(JK \leq a) - \Pr(JK_G \leq a) \leq \Pr(a < JK_G \leq a + \epsilon) \\ + |\Pr(JK_I \leq a + \epsilon) - \Pr(JK_G \leq a + \epsilon)| \\ + \Pr(|JK_I - JK| > \epsilon)\end{aligned}$$

Take a sequence $\epsilon_n \searrow 0$ such that $\Pr(|JK - JK_I| > \epsilon_n) \to 0$. By the above

$$\begin{aligned}\sup_{a\in\mathbb{R}} \Pr(JK \leq a) - \Pr(JK_G \leq a) \leq \sup_{a\in\mathbb{R}} \Pr(a < JK_G \leq a + \epsilon_n) \\ + \sup_{a\in\mathbb{R}} |\Pr(JK_I \leq a + \epsilon_n) - \Pr(JK_G \leq a + \epsilon_n)| \\ + \Pr(|JK_I - JK| > \epsilon_n)\end{aligned}$$

The first term goes to zero as $\epsilon_n \to 0$ via Lemma A.3; the second term goes to zero by Lemma 3.1, and the third term goes to zero by Lemma 3.2. A symmetric argument shows that $\sup_{a\in\mathbb{R}} \Pr(JK_G \leq a) - \Pr(JK \leq a) \leq o(1)$ which completes the proof. □

Supporting Results

Theorem A.1 (Götze et al. (2019), Theorem 2.6)**.** *Let ξ be a gaussian element with zero mean and covariance Σ_ξ. Then it holds for any $\boldsymbol{a} \in \mathbb{R}^n$ that*

$$\sup_{x\geq 0} p_\xi(x, \boldsymbol{a}) \lesssim (\Lambda_{1\xi}\Lambda_{2\xi})^{-1/2}$$

where $p_\xi(x, a)$ denotes the p.df of $\|\xi - \boldsymbol{a}\|^2$.

Theorem A.2 (Gotze et al. (2021), Theorem 1.2)**.** *Let $X_1, \ldots, X_n$ be independent random variables satisfying $\|X_i\|_{\Psi_a} \leq M$ for some $a \in (0, 1] \cup \{2\}$ and let $f : \mathbb{R}^n \to \mathbb{R}$ be a polynomial of total degree $D \in \mathbb{N}$. Then for all $t > 0$;*

$$\Pr(|f(X) - \mathbb{E}[f(X)]| \geq t) \leq 2\exp\left(-\frac{1}{C_{D,a}} \min_{1\leq d\leq D} \left(\frac{t}{M^d \|\mathbb{E}f^{(d)}(X)\|_{HS}}\right)^{a/d}\right)$$

Lemma A.1. *Let X_n and Y_n be two sequences of random variables and let $W_n = X_n/Y_n$. Then for any $c \in \mathbb{R}$ and any $\delta > 0$:*

$$\Pr(0 \leq X_n - cY_n \leq \delta) \leq \Pr(c \leq W_n \leq \delta^{1/2} + c) + \Pr(Y_n \leq \delta^{1/2})$$

and

$$\Pr(-\delta \leq X_n - cY_n \leq 0) \leq \Pr(c - \delta^{1/2} \leq W_n \leq c) + \Pr(Y_n \leq \delta^{1/2})$$

Lemma A.2. *Suppose that X_n and Y_n are sequences of (real-valued) random variables such that $Y_n = O_p(1)$ and for any $\Delta \in \mathbb{R}$*

$$\sup_{x \leq \Delta} |\Pr(X_n \leq x) - \Pr(Y_n \leq x)| \to 0$$

Then $\sup_{x \in \mathbb{R}} |\Pr(X_n \leq x) - \Pr(Y_n \leq x)| \to 0$.

Lemma A.3. *Let $f(\cdot, \tilde{r})$ be the density function of $\frac{\tilde{N}}{\tilde{D}^{1/2}}|\tilde{r}$. Under Assumptions 3.1 and 3.3 there is a constant $M > 0$ such that $\sup_x |f(x, \tilde{r})| \leq M$ for almost all $\tilde{r}$.*

Lemma A.4. *Let $D \in \mathbb{R}^{n \times n}$ be a diagonal real matrix such that $d_{ii} \in [u, U]$ for all $i = 1, \ldots, n$. Let $A \in \mathbb{R}^{n \times n}$ be a symmetric real matrix. For an arbitrary square matrix M, let $\lambda_k(M)$ denote the k^{th} largest eigenvalue of M. Then for any $k = 1, \ldots, n$:*

$$u\lambda_k(A^2) \leq \lambda_k(ADA) \leq U\lambda_k(A^2)$$

Lemma A.5. *Let $\Delta_{1i}, \tilde{\Delta}_{1i}, \Delta^a_{2i}, \tilde{\Delta}^a_{2i}, \Delta^b_{2i}, \tilde{\Delta}^b_{2i}$ be as in the proof of Lemma 3.1 . Then under Assumptions 3.1 and 3.2 there is a constant $M > 0$ such that for any $k = 1, \ldots, 6$:*

$$\mathbb{E}[|\Delta_{1i}|^k] \leq M \qquad \mathbb{E}[|\tilde{\Delta}_{1i}|^k] \leq M$$

and for any $k = 1, \ldots, 3$:

$$\mathbb{E}[|\Delta^a_{2i}|^k] \leq M\alpha^k \qquad \mathbb{E}[|\tilde{\Delta}^k_{2i}|] \leq M\alpha^k$$
$$\mathbb{E}[|\Delta^b_{2i}/\sqrt{n}|^k] \leq M\alpha^k \qquad \mathbb{E}[|\tilde{\Delta}^b_{2i}/\sqrt{n}|^k] \leq M\alpha^k$$

Lemma A.6. *Under Assumptions 3.1–3.3 there is a fixed constant M such that for all $i = 1, \ldots, n$ and any $k = 1, \ldots, 6$,*

$$\mathbb{E}[|N|^k] + \mathbb{E}[|N_{-i}|^k] \leq M$$

Lemma A.7. *Suppose that Assumptions 3.1–3.4 hold. Then $(\Delta_N, \Delta_D)' \to_p 0$.*

References

Anderson, T. W. and H. Rubin (1949). Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations. *The Annals of Mathematical Statistics 20*(1), 46 – 63.

Andrews, D. W. and J. H. Stock (2007). Testing with many weak instruments. *Journal of Econometrics 138*(1), 24–46. 50th Anniversary Econometric Institute.

Andrews, D. W. K., M. J. Moreira, and J. H. Stock (2006). Optimal two-sided invariant similar tests for instrumental variables regression. *Econometrica 74*(3), 715–752.

Andrews, I. (2016). Conditional linear combination tests for weakly identified models. *Econometrica 84*(6), 2155–2182.

Angrist, J. D., G. W. Imbens, and A. B. Krueger (1999). Jackknife instrumental variables estimation. *Journal of Applied Econometrics 14*(1), 57–67.

Angrist, J. D. and A. B. Krueger (1991). Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics 106*(4), 979–1014.

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

Belloni, A., V. Chernozhukov, D. Chetverikov, C. Hansen, and K. Kato (2018). High-dimensional econometrics and regularized gmm.

Celentano, M., A. Montanari, and Y. Wu (2020, 09–12 Jul). The estimation error of general first order methods. In J. Abernethy and S. Agarwal (Eds.), *Proceedings of Thirty Third Conference on Learning Theory*, Volume 125 of *Proceedings of Machine Learning Research*, pp. 1078–1141. PMLR.

Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics 34*(3), 305–334.

Chatterjee, S. (2006). A generalization of the Lindeberg principle. *The Annals of Probability 34*(6), 2061 – 2076.

Chernozhukov, V., D. Chetverikov, and K. Kato (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics 41*(6), 2786 – 2819.

Chetverikov, D. and J. R.-V. Sørensen (2021). Analytic and bootstrap-after-cross-validation methods for selecting penalty parameters of high-dimensional m-estimators. *ArXiv NA*, 1–50.

Crudu, F., G. Mellace, and Z. Sándor (2021). Inference in instrumental variable models with heteroskedasticity and many instruments. *Econometric Theory 37*(2), 281–310.

Derenoncourt, E. (2022, February). Can you move to opportunity? evidence from the great migration. *American Economic Review 112*(2), 369–408.

Friedman, J., R. Tibshirani, and T. Hastie (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software 33*(1), 1–22.

Gautier, E. and C. Rose (2021). High-dimensional instrumental variables regression and confidence sets.

Gilchrist, D. S. and E. G. Sands (2016). Something to talk about: Social spillovers in movie consumption. *Journal of Political Economy 124*(5), 1339–1382.

Götze, F., A. Naumov, V. Spokoiny, and V. Ulyanov (2019). Large ball probabilities, Gaussian comparison and anti-concentration. *Bernoulli 25*(4A), 2538 – 2563.

Gotze, F., H. Sambale, and A. Sinulis (2021). Concentration inequalities for polynomials in alpha-sub-exponential random variables. *Electronic Journal of Probability 26*(none), 1 – 22.

Harrell, F. E. (2015). *Regression Modeling Strategies*. Spring Series in Statistics. Springer Cham.

Kleibergen, F. (2002, 02). Pivotal statistics for testing structural parameters in instrumental variables regression. *Econometrica 70*, 1781–1803.

Kleibergen, F. (2005). Testing parameters in gmm without assuming that they are identified. *Econometrica 73*(4), 1103–1123.

Lim, D., W. Wang, and Y. Zhang (2022). A conditional linear combination test with many weak instruments.

Lindeberg, J. W. (1922). Eine neue herleitung des exponentialgesetzes in der wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift 15*, 211–225.

Matsushita, Y. and T. Otsu (2022). A jackknife lagrange multiplier test with many weak instruments. *Econometric Theory*, 1–24.

Mikusheva, A. (2023). Many weak instruments in time series econometrics. *Working Paper*.

Mikusheva, A. and L. Sun (2021, 12). Inference with many weak instruments. *The Review of Economic Studies 89*(5), 2663–2686.

Moreira, M. (2009, 10). Tests with correct size when instruments can be arbitrarily weak. *Journal of Econometrics 152*, 131–140.

Moreira, M. J. (2001). *Tests with correct size when instruments can be arbitrarily weak*. Citeseer.

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

Navjeevan, M. (2023). An identification and dimensionality robust test for instrumental variables models.

Paravisini, D., V. Rappoport, P. Schnabl, and D. Wolfenzon (2014, 09). Dissecting the Effect of Credit Supply on Trade: Evidence from Matched Credit-Export Data. *The Review of Economic Studies 82*(1), 333–359.

Pouzo, D. (2015). Bootstrap consistency for quadratic forms of sample averages with increasing dimension. *Electronic Journal of Statistics 9*(2), 3046 – 3097.

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

Tan, Z. (2017). Regularized calibrated estimation of propensity scores with model misspecification and high-dimensional data. *ArXiv NA*, 1–60.

van der Greer, S. (2016). *Estimation and Testing under Sparsity*. Lecture Notes in Mathematics. Springer, New York, NY.