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

Manu Navjeevan*

University of California, Los Angeles

Revised: November 13, 2023

Abstract

I propose a new identification-robust test for the structural parameter in a heteroskedastic linear instrumental variables model. The proposed test statistic is similar in spirit to a jackknife version of the K-statistic and the resulting test has exact asymptotic size so long as an auxiliary parameter can be consistently estimated. This is possible under approximate sparsity even when the number of instruments is potentially much larger than the sample size. As the number of instruments is allowed, but not required, to be large, the limiting behavior of the test statistic is difficult to examine via existing central limit theorems. Instead, I derive the asymptotic chi-squared distribution of the test statistic using a direct Gaussian approximation technique. To improve power against certain alternatives, I propose a simple combination with the sup-score statistic of Belloni et al. (2012) based on a thresholding rule. I demonstrate favorable size control and power properties in a simulation study and apply the new methods to revisit the effect of social spillovers in movie consumption.

Keywords: Instrumental Variables, Weak Identification, High-Dimensional

JEL Codes: C12, C36, C55

*Email: mnavjeevan@ucla.edu. 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.

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. The variable $z_i = (z_{1i}, z_{2i})' \in \mathbb{R}^{d_z}$ 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 identification-robust testing procedures all require some conditions on the rate of growth of the number of instruments, d_z, in relation to the sample size, n. The tests of Anderson and Rubin (1949), Staiger and Stock (1997), Moreira (2001, 2003), and Kleibergen (2002, 2005) are shown by Andrews and Stock (2007) to control size under heteroskedasticity when $d_z^3/n \to 0$. Recent tests developed in Mikusheva and Sun (2021), Crudu et al. (2021), Matsushita and Otsu (2022), and Lim et al. (2022) allow the number of instruments to be proportional to sample size, $d_z/n \to \varrho \in [0, 1)$, but require that the number of instruments be large, $d_z \to \infty$. Moreover, while the tests proposed by Belloni et al. (2012) and Mikusheva (2023) require only $\log^M(d_z)/n \to 0$ for some constant $M > 1$, they either rely on sample splitting or fail to incorporate first-stage information, both of which may reduce power in overidentified models. These conditions on the growth rate of d_z can be difficult to interpret and the variety of tests available under alternate regimes may make it unclear to the researcher which test should be applied in her particular setting.[1]

In contrast, the test considered in this paper can work in any of the settings described above. The proposed test statistic is similar in spirit to a jackknife version of the K-statistic and makes use of an auxiliary conditional slope parameter to construct first-stage estimates that are uncorrelated with the structural errors under the null hypothesis. So long as this auxiliary parameter can be consistently estimated, the proposed test statistic has a limiting chi-squared distribution with degrees of freedom equal to the number of structural parameters. The conditional slope parameter is simple to estimate with out-of-the-box methods, and consistency is achievable under approximate sparsity even when the number of instruments is much larger than the sample size. This approximate sparsity assumption is trivially satisfied when the first- and second-stage errors are homoskedastic.

As the number of instruments is allowed to be larger than the sample size, traditional central limit theorems are insufficient to derive the limiting distribution of the test statistic. Instead, the limiting behavior of the test statistic is examined directly with a variation of Lindeberg's interpolation method (Lindeberg, 1922). The direct approach allows me to characterize the limiting behavior of the test statistic in both high- and low-dimensional settings under a unified argument. I show that, in local neighborhoods of the null, the distribution of the test statistic can be uniformly approximated by the distribution of an analog statistic that replaces each observation in the expression of the test statistic with a Gaussian version that has the same mean and covariance matrix as the original. These local neighborhoods of the null are characterized

[1] For example, consider a setting similar to that of Derenoncourt (2022) where the researcher has a dozen or so instruments and a sample size of a few hundred. The number of instruments cubed is larger than sample size, but asymptotic approximations based on $d_z \to \infty$ seem unlikely to resemble the finite sample distribution.

by a local power index which I introduce in Section 3. An interesting feature of the direct Gaussian approximation argument is that, while the numerator and denominator the jackknife K-statistic may both have nontrivial distributions under weak identification, neither needs to converge on its own to a weak limit in order to characterize the limiting distribution of the jackknife K-statistic.

When there is a single endogenous variable, a leading case in empirical applications, analysis of limiting behavior can be considerably simplified by taking advantage of the particular form of the test statistic. In this case I show that, under an additional regularity condition, an infeasible version of the test that could be constructed if the auxiliary parameter was known to the researcher is consistent whenever the local power index diverges. When the local power index is bounded, I examine the limiting power of the test by examining the behavior of the analog statistic. Under the alternative hypothesis the analog statistic has a nearly non-central χ^2 distribution conditional on the first-stage estimates. The noncentrality parameter is proportional to the correlation between the true first-stage model and the first-stage estimates. Unfortunately, partialling out the structural parameter may introduce bias into the first-stage estimates under the alternate hypothesis. This issue is pointed out by Moreira (2001), Andrews et al. (2006), and Andrews (2016) in the context of the non-jackknife K-statistic. Against certain alternatives this bias can be particularly pronounced and the additional regularity condition needed for the consistency result may fail to hold.[2]

To address this, I propose a simple combination of the jackknife K-statistic with the sup-score statistic of Belloni et al. (2012) based on a thresholding rule. As with the Anderson-Rubin statistic, while the sup-score statistic does not yield efficient inference under strong identification, it does not suffer from a loss of power against any particular alternative when identification is weak (Andrews et al., 2006; Andrews, 2016). The combination test decides whether the jackknife K-test or the sup-score test should be run by comparing the value of a conditioning statistic to a predetermined cutoff value. In the approximating Gaussian regime, this conditioning statistic is marginally independent of both the jackknife K-statistic and the sup-score statistic. This allows me to show that the combination test controls size under the null without having to require that the conditioning statistic converges in distribution to a stable limit. In a simulation study, I find that taking this cutoff value to be the 75^{th} quantile of the distribution of the conditioning statistic delivers a reasonable balance of power against local and distant alternatives. Using results in Chernozhukov et al. (2017) and Belloni et al. (2018) this quantile can be simulated via a multiplier bootstrap procedure.

When there are multiple endogenous variables, I cannot take advantage of the simplified form of the test statistic. Instead, I use a more involved interpolation argument that relies on strengthened moment conditions. Under these strengthened conditions I derive the limiting chi-squared distribution of the jackknife K-statistic in the larger context and propose a generalization of the thresholding test to improve power properties.

I apply the proposed testing procedures to the data of Gilchrist and Sands (2016) to generate weak instrument-robust confidence intervals for the effect of social spillovers in movie consumption. Following Belloni et al. (2012), the authors' initial analysis uses conventional heteroskedasticity-robust standard errors after estimating the first-stage via post-LASSO. The validity of this analysis depends on the structural parameter being strongly identified. Using a simple numerical demonstration, I argue that the first-stage F-statistics reported by the authors may not be reliable indicators of identification strength when LASSO is used to select instruments. The identification-robust confidence intervals generated by inverting the jackknife K-statistic are larger than those implied by the initial analysis but do not rule out the authors' point estimates. Moreover, these confidence intervals are typically smaller than those

[2]This does not necessarily mean, however, that the test is not consistent.

implied by inverting the sup-score statistic in specifications for which the sup-score confidence interval is nonempty.

Finally, I examine the applicability of the theoretical results in this paper through a simulation study. Tests based on the jackknife K-statistic are shown to have nearly exact size in a variety of settings. While the jackknife K-statistic may have diminished power against certain alternatives, this deficiency seems to be ameliorated by combining the jackknife K-statistic with the sup-score statistic via the thresholding test. Compared to the many-instrument tests of Mikusheva and Sun (2021) and Matsushita and Otsu (2022), the tests proposed in this paper appear to have favorable size control and power properties.

The outline of this paper is as follows. Section 2 formally defines the model considered and introduces the jackknife K-statistic. Section 3 provides an overview of the Gaussian approximation approach with a single endogenous variable and characterizes the limiting behavior of the test statistic in this setting. Section 4 uses this characterization to examine the power properties of the test and introduces the combination test to address power deficiencies against certain alternatives. Section 5 extends the analyses of Sections 3 and 4 to the case of multiple endogenous variables and outlines the Gaussian approximation argument in this setting. Section 6 contains the empirical application while Section 7 provides evidence from simulation study. Proofs of the main results are deferred to Appendices A–C.

Notation. For any $n \in \mathbb{N}$ let $[n]$ denote the set $\{1, \dots, n\}$. I work with a sequence of probability measures P_n on the data $\{(y_i, x_i, z_i) : i \in [n]\}$. To accommodate independent but not identically distributed observations, let $\mathbb{E}_n[f_i] = n^{-1} \sum_{i=1}^n f_i$ denote the empirical expectation and $\bar{\mathbb{E}}[f] = \mathbb{E}_n[\mathbb{E}[f_i]]$ denote the average expectation operator.

1.1 Prior Literature and Empirical Practice

When the first-stage F-statistic is small, standard asymptotic approximations may fail to accurately describe the finite-sample behavior of IV estimates. This was first pointed out by Nelson and Startz (1990) and Bound et al. (1995) who consider the finite-sample behavior of two-stage least squares (2SLS) in alternate settings where the IV is only weakly correlated with the endogenous variable. In a seminal paper, Staiger and Stock (1997) capture this phenomena in an asymptotic framework by considering a sequence of first-stage models that shrink to zero with the sample size. Under this framework, standard IV estimates are no longer consistent and inference procedures based on these statistics fail to control size. Because of these results, there has been a large interest in developing tests for the structural parameter that control size regardless of identification strength.

To test hypotheses about the structural parameter when instruments are suspected to be weak, Staiger and Stock (1997) propose the use of the Anderson-Rubin statistic, which does not require any assumptions about identification strength to control size. Noting that the Anderson-Rubin test is inefficient in overidentified models, Moreira (2001) and Kleibergen (2002, 2005) propose the use of the (non-jackknife) K-statistic, which has a limiting null distribution that does not depend on the number of instruments. Compared to the Anderson-Rubin statistic, these tests have improved power in local neighborhoods of the null but can perform poorly against certain alternatives. To address this, Moreira (2003) and Kleibergen (2005) suggest combinations of the K-statistic and Anderson-Rubin statistic based on a conditioning statistic that is independent of them both under the null. Andrews et al. (2006) characterize the power envelope in a homoskedastic weakly identified IV model and show that the test based on the conditional likelihood ratio statistic of Moreira (2003) has nearly optimal power in this setting. When errors are heteroskedastic, Andrews (2016) proposes alternate combinations of the K-statistic and AR-statistic based on a minimax regret criterion.

These initial tests are developed under asymptotic frameworks that treat the number of instruments as fixed or growing slowly relative to the sample size (Han and Phillips, 2006; Newey and Windmeijer, 2009; Andrews and Stock, 2007). However, with the emergence of large datasets and more sophisticated research designs, researchers may encounter scenarios where the number of instruments may not be negligible as a ratio of sample size. A prominent example of this is in judge-design settings where the number of instruments is equal to the number of judges to whom an individual can be assigned to (Maestas et al., 2013; Sampat and Williams, 2019; Dobbie et al., 2018). Since each judge can handle only a finite number of cases the number of instruments is proportional to the sample size. Moreover, to flexibly model the first-stage, researchers may generate a large number of instruments by enriching a "small" initial set of instruments via polynomial (or other) transformations. Angrist and Krueger (1991) famously interact quarter-of-birth, state-of-birth, and year-of-birth dummies to construct a total of 180 instruments. Belloni et al. (2012) show that, when identification is strong, researchers can use a potentially high-dimensional, $d_z \gg n$, set of first-stage instrument basis terms in conjunction with a LASSO or post-LASSO estimate of the first-stage. This strategy has been successfully employed in practice by Paravisini et al. (2014), Gilchrist and Sands (2016), Derenoncourt (2022), and Jou and Morgan (2023).

To address these settings, there has been recent interest in developing weak instrument-robust tests under asymptotic frameworks that do not require that the ratio of instruments to sample size tends to zero. Crudu et al. (2021), Mikusheva and Sun (2021), and Matsushita and Otsu (2022) take advantage of a new central limit theorem for quadratic forms developed in Chao et al. (2012) and propose weak identification-robust tests that are valid even when the number of instruments is proportional to sample size; $d_z/n \to \varrho \in [0, 1)$. Following the many instruments asymptotic framework first introduced by Bekker (1994), the analyses in these papers rely on the number of instruments diverging. When the number of instruments is fixed or diverges slowly to infinity, these asymptotic approximations may provide poor characterizations of the proposed test statistics' finite sample distribution.

Limited identification-robust testing procedures exist for the high-dimensional case, $d_z \gg n$. To my knowledge, the only two options available are the sup-score test of Belloni et al. (2012) and the split-sample optimal instrument AR test developed in Mikusheva (2023). The sup-score test makes use of Gaussian approximations for maxima of high-dimensional vectors developed in Chernozhukov et al. (2013) but suffers from the same issue as the Anderson-Rubin test in that its critical value is increasing with the number of instruments. The spilt sample optimal instrument AR test splits the dataset into two parts and uses one split to estimate an optimal instrument and the other to test the null hypothesis. This may lead to a loss of power as only half of the sample is being effectively used to test the null hypothesis.

Weak instrument-robust tests may be particularly interesting in high-dimensional and heteroskedastic settings because of a lack of clarity on how to pretest for identification strength. When the number of instruments is fixed and errors are homoskedastic, Stock and Yogo (2005) propose pretesting for the strength of identification via thte first-stage F-statistic. Based on their results, common practice in empirical settings has been to use standard Wald tests whenever the first-stage F-statistic exceeds 10. Lee et al. (2022) point out this recommendation is not applicable in heteroskedastic models and update the recommended F-statistic cutoff. To pretest for weak identification in the many-instruments asymptotic framework, $d_z \to \infty$, Mikusheva and Sun (2021) propose a new $\tilde{F}$-statistic and suggest using identification-robust procedures when $\tilde{F} < 4.14$. When the number of instruments is larger than sample size there is no accepted full-sample pretest for identification strength.[3] In particular, I argue in Section 6 that first-stage

[3] Mikusheva (2023) suggests a split-sample pretest in the same spirit as the split-sample optimal instrument AR test.

F-statistics resulting from first-stage post-LASSO procedures can be misleading even if they are much larger the standard cutoff of 10.

Asymptotic Regime	Main Tests
Low-Dimensional: $d_z^3/n \to 0$	Anderson-Rubin K/Lagrange Multiplier Conditional Linear Combination
Many-Instruments: $d_z/n \to \phi \in [0,1)$ $d_z \to \infty$	Jackknife-AR Jackknife-LM Conditional Linear Combination
High-Dimensional: $\log^M(d_z n)/n \to 0$	Sup-Score Test Split-Sample AR

Table 1: Existing Identification and Heteroskedasticity Robust Tests for Linear IV models.

I contribute to these literatures by proposing a new identification-robust test for the structural parameter that can work in potentially high-dimensional settings ($d_z \gg n$) without requiring that the number of instruments diverges. The testing procedures in this paper may be particularly applicable in intermediate cases where the number of instruments cubed may not be negligible relative to sample size but it is unclear whether asymptotic approximations based on $d_z \to \infty$ will accurately describe finite sample behavior. Examples of such intermediate cases include the analyses of Derenoncourt (2022), where $d_z = 9$ and $n = 239$, Paravisini et al. (2014), where $d_z = 10$ and $n = 5995$, and Gilchrist and Sands (2016), where $d_z = 52$ and $n = 1671$.

In addition to the literature on weak-instrument robust testing, I contribute to a growing literature on direct gaussian approximation and interpolation techniques (Chatterjee, 2006, 2010; Pouzo, 2015; Chernozhukov et al., 2013, 2017; Celentano et al., 2020). These techniques have proven useful to approximate the behaviors of statistics in a variety of nonstandard settings, such as high-dimensional random vectors or spectral analysis of random matrices. Prior analysis of statistics via interpolation techniques has relied on the boundedness of the derivatives of these statistics with respect to individual observations. This condition does not hold in my setting as the derivative of the jackknife K-statistic with respect to terms in the denominator may be unbounded. This poses a number of technical challenges for my interpolation argument that must be overcome in order to characterize the limiting behavior of the jackknife K-statistic, particularly when $d_x > 1$. I contribute to this literature by proposing modifications of the original Lindeberg (1922) interpolation technique that can accommodate statistics with unbounded derivatives.

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] Along with the structural equation in (1.1), the

[1] For discussion refer to Appendix E.

IV model can then be written with the first stage 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 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]

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

I will largely work under the assumption that $\rho(z_i)$ 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. This allows for nesting of the low-dimensional case, where the number of instruments is fixed, and the high dimensional case, where the number of instruments is potentially much larger than the sample size, under

[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.

a unified estimation procedure. Under homoskedasticity, $\rho_\ell(z_i)$ is a constant function and thus has a spare representation in any basis that contains a constant term.

As in Chernozhukov et al. (2022), 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 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 $\epsilon_i(\beta_0) b(z_i)$. It 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. The estimation procedure is discussed in more detail in Section 3.2. With $\hat{\rho}(z_i) := b(z_i)'\hat{\phi}_\ell$, I construct the estimated version of $r_{\ell i}$, $\hat{r}_{\ell i} := x_i - \hat{\rho}(z_i)\epsilon_i(\beta_0)$ for each $\ell \in [d_x]$.

2.1 Test Statistic

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

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

for some "hat" matrix $H = [h_{ij}] \in \mathbb{R}^{n \times n}$. The phrase "hat matrix" is borrowed from ordinary least squares (OLS) where the projection matrix, $z(z'z)^{-1}z'$, is sometimes referred to as the hat matrix in the sense that $\hat{x} = z(z'z)^{-1}z'x$. In practice, the hat matrix, H, can be any matrix that depends only on z. It is important to note that while $\widehat{\Pi}_{\ell i}$ does not depend on the observation $r_{\ell i}$, it may depend on z_i through the hat matrix H. This gives the test power against alternatives where $\mathbb{E}[\epsilon_i(\beta_0) z_i] \neq 0$. For technical reasons, I will assume that $h_{ii} = 0$ for each $i \in [n]$ so that $\widehat{\Pi}_{\ell i}$ can be written as $\widehat{\Pi}_{\ell i} = \sum_{j=1}^n h_{ij} r_{\ell j}$.

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 $H(\lambda^\star)$:

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

where, following recommendations in Harrell (2015) and van Wieringen (2023), 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:

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

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 researcher may use any other hat matrix that she believes will lead to plausible first-stage estimates. Other possible choices of hat matrix include the jackknife OLS hat matrix of Angrist et al. (1999), the deleted diagonal projection matrix introduced in Chao et al. (2012) and successfully used in Kline et al. (2020), Crudu et al. (2021), Mikusheva and Sun (2021), and Matsushita and Otsu (2022), or hat matrices based on selecting instruments via some preliminary unsupervised technique such as principal

component analysis (PCA). Remark 3.1 below discusses how the balanced-design condition may be verified for arbitrary choices of hat matrices.

For each $i = 1, \ldots, n$, define $\widehat{\Pi}_i = (\widehat{\Pi}_{1i}, \ldots, \widehat{\Pi}_{d_x i}) \in \mathbb{R}^{d_x}$ and $\widehat{\Pi}_{\epsilon i} = \epsilon_i(\beta_0)\widehat{\Pi}_i$. Collect these in the matrices

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

The jackknife K-statistic 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 on the hat matrix, H, and moment bounds, I will show that 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 the form of the test statistic simplifies to $JK(\beta_0) = \big(\sum_{i=1}^n \epsilon_i(\beta_0)\widehat{\Pi}_i\big)^2 / \sum_{i=1}^n \epsilon_i^2(\beta_0)\widehat{\Pi}_i^2$. The extension to $d_x > 1$ is not immediate but is possible under strengthened moment conditions and is explored in 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. In this section, I detail the approach and take advantage of the simplified analysis to characterize the limiting behavior of the test statistic under local alternatives to H_0. This direct approach has the advantage of not relying on any particular central limit theorem, which allows a great deal of flexibility in the choice of hat matrix H.

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 will be 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 deviations.

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 events $(\{JK_I(\beta_0) \leq a\}, \{JK_G(\beta_0) \leq a\})$ are almost surely equivalent to the events $(\{JK_I^a(\beta_0) \leq 0\}, \{JK_G^a(\beta_0) \leq 0\})$. The decomposed statistics no longer have denominators to be dealt with and are tractable for the interpolation argument. I show for any $\varphi(\cdot) \in C_b^3(\mathbb{R})$, the space of all thrice continuously differentiable functions with bounded derivatives up to the third order, that there is a fixed constant $M > 0$ such that

$$|\mathbb{E}[\varphi(JK_I^a) - \varphi(JK_G^a)]| \leq \frac{M(a^3 \vee 1)}{\sqrt{n}}(L_2(\varphi) + L_3(\varphi)) \tag{3.2}$$

where $L_2(\varphi) := \sup_x |\varphi''(x)|$ and $L_3(\varphi) := \sup_x |\varphi'''(x)|$. By taking $\varphi(\cdot)$ to be a sequence of functions approximating the indicator function, $\mathbf{1}\{x \leq 0\}$, the result in (3.2) can be used to show that the cumulative distribution function (CDF) of the infeasible statistic $JK_I(\beta_0)$ can be approximated by the CDF of the Gaussian statistic $JK_G(\beta_0)$ at each point $a \in \mathbb{R}$. A Glivenko-Cantelli type argument is then be applied to show the approximation holds uniformly over all points on the real line. The Lindeberg interpolation argument on the decomposed test statistics makes use of the fact that the numerator and denominator of the Gaussian test statistic are functions of quadratic forms in the random vectors $\epsilon(\beta_0) := (\epsilon_1(\beta_0), \ldots, \epsilon_n(\beta_0))'$ and $r := (r_1, \ldots, r_n)'$.[1]

Moving from approximation of expectations of smooth functions to approximation of the CDF relies on a particular anticoncentration bound on $\tilde{D}$. I show that that this bound can be established under either weak or strong identification. This allows for the limiting null distribution of the test statistic under various identification regimes to be derived via a unifying argument. Additionally, even though $(N, D, \tilde{N}, \tilde{D})$ may all have nonnegligible distributions when identification is weak, the interpolation argument does not require any of these to individually converge in distribution or probability anywhere stable. This allows for a wide

[1] See Pouzo (2015) for another example of the Lindeberg interpolation method applied to approximate the distribution of quadratic forms.

range of possible hat matrices H to be used in constructing the first stage estimates, $(\widehat{\Pi}_1, \ldots, \widehat{\Pi}_n)$. In particular, no assumption need be made on the number of instruments used to construct H nor any requirement imposed that the first-stage estimates $(\widehat{\Pi}_1, \ldots, \widehat{\Pi}_n)$ are consistent.

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'.*

Assumptions 3.1 and 3.2 allow characterization of the null distribution of $JK(\beta_0)$. 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(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 mainly to rule out hat matrices that are all zeroes or nearly all zeros so that the effective number of observations used to test the null is growing with the sample size. Remark 3.1 below discusses how this assumption and Assumption 3.2(ii) may be verified in practice. Remark 3.2 compares this balanced design assumption to that in the many-instruments literature (Crudu et al., 2021; Mikusheva and Sun, 2021; Matsushita and Otsu, 2022; Lim et al., 2022), noting that their balanced design neither implies nor is implied by the one in this paper.

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.[2] 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, 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 ($\Pi_i \lesssim n^{-1/2}$). Assumption 3.2(iii) can be viewed as a technical requirement that there be more than one "effective" instrument in the hat matrix.[3] This condition can be easily verified in practice by examining the eigenvalues of HH'.

In addition to characterizing the limiting distribution of $JK(\beta_0)$ under H_0, I also examine the behavior of $JK(\beta_0)$ in local neighborhoods of the null. These local neighborhoods are characterized by the local power index P, defined below, as well as an additional regularity condition that restricts the size of $\mathbb{E}[\epsilon_i(\beta_0)]$ relative to $\mathbb{E}[r_i]$.

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

[2] To see this, notice that $s_n^{-2} = \max_i \mathbb{E}[(\widehat{\Pi}_i^I)^2] \geq \max_i \mathrm{Var}(\widehat{\Pi}_i^I) = \max_i \sum_{j \neq i} h_{ij}^2 \mathrm{Var}(r_j)$. By Assumption 3.1, $\mathrm{Var}(r_j)$ 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 $i \in [n]$.

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

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 $\sum_{j\neq i} s_n^2 h_{ji}^2 \leq c$. The local power index is the second moment of the scaled numerator, N and is a measure of the association between the true first stage Π_i and the first-stage estimates $\widehat{\Pi}_i$. In Section 4, I discuss how the strength of this association is related to the power of the test under local alternatives. Proposition 3.1 below shows that when Assumption 3.3(ii) holds, $P \to \infty$ implies that the test based on the infeasible statistic $JK_I(\beta_0)$ is consistent.

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 $\mathbb{E}[(\sum_{j\neq i} h_{ij}r_j)^2]$. Using the moment bounds in Assumption 3.1 and Assumption 3.2(ii) one can verify that Assumption 3.3(ii) is equivalent to the existence of constants $C_1, C_2 > 0$ such that

$$\begin{aligned}\max_i \Big(\sum_{j\neq i} h_{ji}\mathbb{E}[\epsilon_j(\beta_0)]\Big)^2 &\leq C_1 \max_i \mathbb{E}\Big[(\sum_{j\neq i} h_{ij}r_j)^2\Big] + C_2 \\ &= C_1 \max_i \left\{ \sum_{j\neq i} h_{ij}^2 \operatorname{Var}(r_j) + \Big(\sum_{j\neq i} h_{ij}\mathbb{E}[r_j]\Big)^2 \right\} + C_2\end{aligned}$$

for all $i \in [n]$. 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 $(\epsilon_1(\beta_0), \ldots, \epsilon_n(\beta_0))$ are of the same or lesser order than the means of $(r_1, \ldots, r_n)$.

Under Assumptions 3.1–3.3, I establish a main technical lemma stating 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,*

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

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; i.e for any fixed $a \in \mathbb{R}$, $\Pr(JK_I(\beta_0) \leq a) \to 0$.*

The dependence of the consistency result on Assumption 3.3(ii) is a nontrivial restriction because of the 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. This is an extreme case, however. In general, 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, as discussed in Remark 3.5, Proposition 3.1 does not necessarily rule out consistency when $P \to \infty$ but Assumption 3.3(ii) fails.

Regardless, bias taken on in constructing r_i has consequences for the power of the test in finite samples. This is particularly true when the mean of r_i is of a lesser order than that of $\epsilon_i(\beta_0)$ as will be discussed in Section 4. To rectify this deficiency in tests based on the jackknife K-statistic, I suggest a thresholding test that decides whether to use the jackknife K-statistic or the sup-score Belloni et al. (2012) statistic based on the value of the conditioning statistic. This

conditioning statistic, in turn, is based on a test statistic for the null hypothesis that $\mathbb{E}[\widehat{\Pi}_i^I] = 0$ for all $i \in [n]$.

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 to a stable limit and then applying the continuous mapping theorem.[4] This approach is not applicable here as neither the scaled numerator nor the scaled denominator has a 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 $\delta_n \to 0$. This is done by first showing that quantiles of D can be approximated by quantiles of $\tilde{D}$, the scaled denominator of $JK_G(\beta_0)$. If the variance of $\tilde{D}$ is bounded away from zero, its density can also be bounded with new bounds on Gaussian quadratic form densities from Götze et al. (2019), which yields the result. Otherwise, if $\text{Var}(\tilde{D}) \to 0$, the result holds by an application of Chebyshev's inequality and $\mathbb{E}[D] > c^{-1}$ from Assumption 3.2(i). This particular anticoncentration bound for $\tilde{D}$ is also important in the proof of Lemma 3.1 to establish anticoncentration for the decomposed Gaussian test statistic.

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; Sambale, 2022). 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*

[4] This is the approach taken by Kleibergen (2002, 2005)

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* $\log(d_b n)^{2/(\upsilon \wedge 1)} \|\widehat{\phi} - \phi\|_1 \to_p 0$; *(v) the following moment bounds hold*

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

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

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. While this condition is more restrictive than the moment condition in Assumption 3.1, as discussed above, it still allows for a wide range of potential distributions. 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 bases such as 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(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. See Belloni et al. (2012), van der Greer (2016), Tan (2017), and Chetverikov and Sørensen (2021) for references under various choices of penalty parameter. This condition allows for the dimensionality of the basis terms, d_b, to grow near exponentially as a function of the sample size. Following the analysis of Tan (2017) one can see that, under appropriate choice of penalty parameter, this may be satisfied as long as $s^2 \log^{2(\upsilon+1)/\upsilon}(d_b n)/n \to 0$, where the sparsity index s denotes the number of nonzero elements of ϕ.

Assumption 3.4(v) is a strengthening of the definition of local neighborhoods and can be interpreted similarly to Assumption 3.3(ii). Since the moment conditions in Assumption 3.4(va,vb) hold with $b_\ell(z_j)\epsilon_j(\beta_0)$ replaced with r_j, Assumption 3.4(v) can be interpreted as requiring that $|\mathbb{E}[\sum_{j \neq i} h_{ij} b_\ell(z_j)\epsilon_j(\beta_0)]|$ is on the same order as $|\mathbb{E}[\sum_{j \neq i} h_{ij} r_j]|$ for all $i = 1, \ldots, n$ and $\ell = 1, \ldots, d_b$. As with Assumption 3.3(ii), it is trivially satisfied under H_0 or, using the fact that $\max_i \sum_{j \neq i} s_n^2 h_{ij}^2 \le c$, whenever $\mathbb{E}[\epsilon_i(\beta_0)] = \Pi_i(\beta - \beta_0)$ is in a $\sqrt{n}$-neighborhood of zero.

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.

Lemma 3.3. *Suppose that Assumptions 3.1–3.4 hold. Then* $(\Delta_N, \Delta_D)' \to_p 0$.

Lemmas 3.1–3.3 are combined for the main result, local approximation of the distribution of the feasible test statistic, $JK(\beta_0)$, by the distribution of the Gaussian statistic, $JK_G(\beta_0)$. An immediate corollary is that the limiting null distribution of $JK(\beta_0)$ is χ_1^2.

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, $JK(\beta_0) \rightsquigarrow \chi_1^2$.

If the limiting $JK_G(\beta_0)$ had a fixed distribution under H_1, Theorem 3.1 would follow immediately from Lemmas 3.1–3.3, and an application of Slutsky's lemma. However, under H_1, there is nothing preventing the distribution of $JK_G(\beta_0)$ changing with the sample size. Instead I

establish Theorem 3.1 directly using the fact that both $JK(\beta_0)$ and $JK_G(\beta_0)$ are bounded in probability and that $JK_G(\beta_0)$ has a density that is bounded uniformly over n.

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 $(cq)^{-1} \leq \frac{q^{\text{th}}\text{-quantile of } \mathbb{E}[(\widehat{\Pi}_i^I)^2]}{\max_i \mathbb{E}[(\widehat{\Pi}_i^I)^2]}$. In practice this can be verified by checking that there is some quantile q such that both

$$\frac{q^{\text{th}}\text{-quantile of } \sum_{j\neq i} h_{ij}^2}{\max_i \sum_{j\neq i} h_{ij}^2} \text{ and } \frac{q^{\text{th}}\text{-quantile of } (\sum_{j\neq i} h_{ij}\hat{r}_j)^2}{\max_i (\sum_{j\neq i} h_{ij}\hat{r}_j)^2} \tag{3.3}$$

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.

Remark 3.2. The balanced-design condition in Assumption 3.2(i) is neither weaker nor stronger than that in the many instruments literature (Crudu et al., 2021; Mikusheva and Sun, 2021; Matsushita and Otsu, 2022; Lim et al., 2022). These papers require that the projection matrix $P = z(z'z)^{-1}z'$ satisfies $[P]_{ii} \leq \delta \leq 1$ for some value δ and all $i \in [n]$. Since P is idempotent, $[P]_{ii} = 1$ for some $i \in [n]$ implies that $[P]_{ij} = 0$ for $j \neq i$.[5] This would not violate Assumption 3.2 if one were to take H such that $h_{ij} = [P]_{ij}\mathbf{1}\{i \neq j\}$; $\mathbb{E}[(\widehat{\Pi}_i^I)^2] = 0$ is allowed for a constant share of $i \in [n]$. Conversely, if the instruments are fixed or grow slowly, it is possible to construct a projection matrix P of rank d_z where $[P]_{ii}$ is bounded away from one for all $i \in [n]$, but "most" of the rows are zero. I view this as a theoretical edge case, however, that seems unlikely to result from real data.

Remark 3.3. The Lindeberg interpolation method allows me to give a nearly uniform explicit bound on the Gaussian approximation error. In particular, using the bound in (3.2), I show that for any fixed value $\Delta > 0$;

$$\sup_{a\leq\Delta} \left|\Pr(JK_I(\beta_0) \leq a) - \Pr(JK_G(\beta_0) \leq a)\right| \leq Cn^{-2/13}$$

where C is a constant that depends only on (c, Δ). Lemma 3.1 makes use of the fact that the limiting statistic $JK_G(\beta_0)$ is bounded in probability and extends this result to show that the approximation error tends to zero uniformly over the real line. While it does not account for estimation error in $\hat{\rho}(\cdot)$, obtaining an explicit bound reflects an improvement over the original analyses of K-statistics in Kleibergen (2002, 2005). These original studies rely on continuous mapping theorems to obtain the limiting chi-squared distributions, making the rate of decay of the approximation error difficult to analyze.

Remark 3.4. The interpolation argument relies on the fact that the first and second moments of $(\tilde{\epsilon}_i(\beta_0), \tilde{r}_i)$ are the same as the first and second moments of $(\epsilon_i(\beta_0), r_i)$ to match the first and moments of one-step deviations with Gaussian analogs. Without the jackknife form of $\widehat{\Pi}_i^I$, these one step deviations would additionally contain cross-terms such as $h_{ii}r_i\epsilon_i(\beta_0)$, for $i \in [n]$. While the first moment of this cross-term is matched by the first moment of the Gaussian analog, $h_{ii}\tilde{\epsilon}_i(\beta_0)\tilde{r}_i$, the second moment is not matched. This is manageable, however, so long as the terms h_{ii} are "small." An example of when the h_{ii} terms are small is when H is taken to be

[5] Since P is idempotent, $[P]_{ii} = \sum_{i=1}^n [P]_{ij}^2 = [P]_{ii}^2 + \sum_{j\neq i} [P]_{ij}^2$.

the OLS projection matrix, $H = z(z'z)^{-1}z$, and the number of instruments satisfies $d_z^3/n \to 0$. See Appendices A and E for details.

Remark 3.5. Proposition 3.1 does not necessarily rule out that a test based on $JK_I(\beta_0)$ is consistent when $P \to \infty$ but Assumption 3.3(ii) fails to hold. The proof of Proposition 3.1 relies on showing that, when $P \to \infty$ and Assumption 3.3(ii) holds, $\mathbb{E}[|N|] \to \infty$ while $\text{Var}(|N|)$ and $\mathbb{E}[D]$ are bounded. These facts can be combined to show that $\Pr(N^2 - aD \leq 0) \to 0$ for any fixed $a \in \mathbb{R}$. When Assumption 3.3(ii) fails, $P \to \infty$ may imply that $\text{Var}(|N|) \to \infty$ as well, making the limiting behavior of the test difficult to analyze. There is reason to believe that this issue can be overcome, Andrews et al. (2004) show that the K-statistic of Kleibergen (2002) is consistent against fixed alternatives under strong identification. However, a full consistency result is not pursued here and left to future work.

Remark 3.6. 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, may not need to be much larger than the dimensionality of the instruments, d_z, to provide a good approximation of $\rho(z_i)$. Interestingly, if taking $b(z_i) = z_i$ provides a good approximation of $\rho(z_i)$, the Tan (2017) result suggests that consistency of $\hat{\rho}(\cdot)$ is achievable under $d_z^2 \log^{2(v+1)/v}(d_z n)/n \to 0$ even if ϕ is fully dense. This requirement is weaker than the $d_z^3/n \to 0$ requirement of the standard K-statistic.

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, $JK_G(\beta_0)$. Conditional on the vector $\tilde{r} = (\tilde{r}_1, \ldots, \tilde{r}_n)$, the distribution of $JK_G(\beta_0)$ is nearly non-central χ_1^2 with noncentrality parameter $\mu(\tilde{r})$, $JK_G(\beta_0)|\tilde{r} \sim A^2(\tilde{r}) \cdot \chi_1^2(\mu(\tilde{r}))$:

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

Under local alternatives, the terms $\Pi_i^2(\beta - \beta_0)^2 \to 0$ so that $A(\tilde{r}) \to 1$ and $|\mu^2(\tilde{r}) - \mu_\infty^2(\tilde{r})| \to 0$, where

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

The numerator of $\mu_\infty^2(\tilde{r})$ 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 $\mu_\infty^2(\tilde{r})$ is maximized by setting $\tilde{\Pi}_i = \Pi_i$ reflecting the classical result of Chamberlain (1987). The denominator of $\mu_\infty^2(\tilde{r})$ 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 AR-statistic based on a conditioning statistic. While the AR-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 AR-statistic is asymptotically independent of both the K-statistic and the conditioning statistic.

Unfortunately, the asymptotic validity of these tests under heteroskedasticity is based on the assumption that $d_z^3/n \to 0$, which may not reasonably describe many settings discussed above. Instead, to improve the power of tests based on the jackknife K-statistic, I consider a simple combination with the sup-score statistic of Belloni et al. (2012). The test based on the sup-score statistic 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 $S(\beta_0) > c_{1-\theta}^S$ where, for $e_1, \ldots, e_n$ iid standard normal and generated independently of the data, $c_{1-\theta}^S$ is the simulated multiplier bootstrap critical value:

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

As with the Anderson-Rubin test, tests based on the sup-score statistic may have suboptimal power properties in overidentified models as it does 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.

The combination test will be based on an attempt to detect whether the alternative β is such that $\mathbb{E}[\widehat{\Pi}_i^J] = 0$ for all $i = 1, \ldots, n$. When this is the case, the test based on $JK(\beta_0)$ will choose a

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

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:

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

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

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

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.

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

for some cutoff τ, which I take in the simulation study and empirical exercise to be the 75$^{\text{th}}$ 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 compare the rejection probability to that of a Gaussian analog. In addition to $JK_G(\beta_0)$, defined in (3.1), define the Gaussian analogs of $S(\beta_0)$ and the conditioning statistic C:

$$S_G(\beta_0) := \sup_{1\le \ell\le d_z} \left| \frac{\sum_{i=1}^n \tilde{\epsilon}_i(\beta_0) z_{\ell i}}{(\sum_{i=1}^n z_{\ell i}^2)^{1/2}} \right| \qquad C_G := \sup_{1\le i\le n} \left| \frac{\sum_{j\neq i} h_{ij}\tilde{r}_j}{(\sum_{j\neq i} h_{ij}^2)^{1/2}} \right|$$

where, as in Section 3, $(\tilde{\epsilon}_i(\beta_0), \tilde{r}_i)'$ are generated independently of each other and the data following a Gaussian distribution with the same mean and covariance matrix as $(\epsilon_i(\beta_0), r_i)$. Since $\text{Cov}(\tilde{\epsilon}_i(\beta_0), \tilde{r}_i) = 0$ under H_0, the statistics C_G and $S_G(\beta_0)$ are independent under the null. Similarly, the null distribution of $JK_G(\beta_0)$ is the same conditional on any realization of $(\tilde{r}_1, \ldots, r_n)$; it is also independent of C_G under the null. The Gaussian analog thresholding test decides whether the researcher should run a test based on $S_G(\beta_0)$ or $JK_G(\beta_0)$ depending on the value of C_G as in (4.5).

The test statistics $JK_G(\beta_0)$ and $S_G(\beta_0)$ are only marginally independent of the conditioning statistic C_G under the null. This limits the ways in which the test statistics can be combined using the conditioning statistic while still controlling size. This marginal independence in the Gaussian limit is enough, however, for the asymptotic validity of the thresholding test, $T(\beta_0;\tau)$. To establish that the behavior of the pairs $(C, JK(\beta_0))$ and $(C, S(\beta_0))$ can be approximated by the behavior of $(C_G, JK_G(\beta_0))$ and $(C_G, S_G(\beta_0))$, respectively, 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* $\|\zeta_i\|_{\psi_v} \le c$*; (ii)* $\max_{i,j} |\frac{h_{ij}}{(\mathbb{E}_n[h_{ij}^2])^{1/2}}| + \max_{l,i} |\frac{z_{li}}{(\mathbb{E}_n[z_{li}^2])^{1/2}}| \le c$*; 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 the conditions needed for error taken on from estimation of $\rho(z_i)$ can treated as negligible under Assumption 3.4.

In Section 7, I investigate the power properties of the thresholding test via simulation study. I find that combining the $JK(\beta_0)$ statistic with the sup-score statistic based on C improves power against distant alternatives and helps alleviate a power decline suffered by the $JK(\beta_0)$ statistic against a particular set of alternatives.

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 $\tilde{\Pi}_i := (\tilde{\Pi}_{1i},\dots,\tilde{\Pi}_{d_x i})' \in \mathbb{R}^{d_x}$, $\tilde{\Pi}_{\epsilon i} := (\mathbb{E}[\epsilon_i^2(\beta_0)])^{1/2}\tilde{\Pi}_i$, and

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

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 $\chi^2_{d_x}$.

In addition to characterizing the local behavior of $JK(\beta_0)$ with multiple endogenous variables, I show that the thresholding test of Section 4.2 can be applied with multiple endogenous variables with a generalized conditioning statistic.

5.1 MODIFIED INTERPOLATION APPROACH

As with a single endogenous variable, error taken on from the estimation of $\rho(z_i)$ prevents immediate comparison of $JK(\beta_0)$ to $JK_G(\beta_0)$. Instead as an intermediate step consider showing that the quantiles of $JK_I(\beta_0)$ can be approximated by corresponding quantiles of $JK_G(\beta_0)$ where $JK_I(\beta_0)$ is an infeasible statistic:

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

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 $(r_1, \ldots, r_n)'$ in place of their estimates $(\hat{r}_1, \ldots, \hat{r}_n)'$.

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 $\varphi_\gamma(\cdot) \in C^3_b(\mathbb{R})$ 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

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

The classical interpolation argument of Lindeberg (1922) would attempt to show (5.1) by one-by-one replacement of each pair, $(\epsilon_i(\beta_0), r_i)'$, in the expression of $\varphi_{\gamma_n}(JK_I(\beta_0))$ with its Gaussian analog, $(\tilde{\epsilon}_i(\beta_0), \tilde{r}_i)'$, and bounding of the size of each of these deviations. As mentioned in Section 3, the problem arises as the derivative of the test statistic, $JK_I(\beta_0)$, 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. Since the overall derivatives of $\varphi_{\gamma_n}(JK_I(\beta_0))$ with respect to $(\epsilon_i(\beta_0), r_i)'$ depend on the product of derivatives with respect to the test statistic and derivatives of $\varphi_{\gamma_n}(\cdot)$, which scale inversely with the approximation parameter, this adjustment of the approximation parameter allows control of the overall derivative. Details of this approach can be found in Appendix D.

This approach relies on stronger moment conditions, which I detail below. These strengthened

[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.

moment conditions are needed mainly needed to bound moments of the determinant of the denominator matrix. For all $\ell = 1, \ldots, d_x$ let $\zeta_{\ell i} := v_i - \rho_\ell(z_i)\eta_i$, noting that $\zeta_{\ell i} = r_{\ell i} - \mathbb{E}[r_{\ell i}]$. Recall also the definition of $\eta_i = \epsilon_i - v_i'(\beta - \beta_0)$, which is equal to $\epsilon_i(\beta_0) - \mathbb{E}[\epsilon_i(\beta_0)]$.

Assumption 5.1 (Moment Conditions). *Assume (i) there are constants $c > 1$ and $v \in (0, 1] \cup \{2\}$ such that $\|\epsilon_i\|_{\psi_a} \leq c$ and $\|\zeta_{\ell i}\|_{\psi_v} \leq c$, 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) For any $\ell = 1, \ldots, d_x$ let $s_{\ell,n}^{-2} = \max_{1\leq i\leq n} \mathbb{E}[(\widehat{\Pi}_{\ell i}^I)^2]$; 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 v-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 there be than two effective instruments in the hat matrix. Assumption 5.2 thus reduces to Assumption 3.2 when $d_x = 1$.

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 $\ell = 1, \ldots, d_x$.

Lemma 5.1 (Infeasible Uniform Approximation). *Suppose that Assumptions 5.1–5.3 hold. Then*

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

5.2 LIMITING BEHAVIOR OF TEST STATISTIC

Having derived the limiting behavior of the infeasible statistic, I next present a high-level condition under which estimation error taken on from estimation of $\rho(z_i)$ can be treated as negligible. I then verify this high-level condition for the ℓ_1-regularized estimators proposed in (2.2). For any $\ell = 1, \ldots, d_x$ define the scaled differences

$$\Delta_{N,\ell} := \frac{s_{\ell,n}}{\sqrt{n}}\sum_{i=1}^n \epsilon_i(\beta_0)(\widehat{\Pi}_{\ell,i} - \widehat{\Pi}_{\ell,i}^I)$$

$$\Delta_{D,\ell} := \frac{s_{\ell,n}^2}{n}\sum_{i=1}^n \epsilon_i^2(\beta_0)(\widehat{\Pi}_{\ell,i}^2 - (\widehat{\Pi}_{\ell,i}^I)^2)$$

As long as these scaled differences tend to zero, Lemma 5.2 shows that the difference between the feasible and infeasible test statistics converges to zero:

Lemma 5.2. *Suppose that Assumptions 5.1–5.3 hold and that* $(\Delta_{N,\ell}, \Delta_{D,\ell}) \to_p 0$ *for all* $\ell = 1, \ldots, d_x$. *Then* $|JK(\beta_0) - JK_I(\beta_0)| \to_p 0$.

As with Lemma 3.2, while Lemma 5.2 is a simple statement, it is not immediate. In particular, establishing Lemma 5.2 requires showing that $\lambda_{\max}(D^{-1})$ is bounded in probability, where D represents a scaled version of the denominator matrix. This requires some work as the scaled denominator matrix is not required to converge in distribution to a stable limit. Instead I directly show that $\lambda_{\max}(D^{-1})$ is bounded in probability by showing that $\Pr(\lambda_{\min}(D) \leq \delta_n) \to 0$ for any sequence $\delta_n \to 0$.

To do this, I first demonstrate that it is sufficient to show that $\Pr(a'Da \leq \delta_n) \to 0$ for any $\delta_n \to 0$ and fixed $a \in \mathcal{S}^{d_x - 1} = \{v \in \mathbb{R}^{d_x} : \|v\| = 1\}$. I then establish the claim for an arbitrary choice of a. As in Lemma 3.2 I do this by comparing the scaled quadratic form of the denominator matrix to a Gaussian analog and then establishing the corresponding result for the Gaussian analog. This corresponding result is again also useful for establishing the validity of the interpolation approach with a dynamic choice of approximation parameter.

I state conditions under which $(\Delta_{N,\ell}, \Delta_{D,\ell}) \to_p 0$ holds for the ℓ_1-regularized estimation procedure proposed in (2.2). These conditions are equivalent to those in Assumption 3.4 but hold for each the d_x estimation procedures.

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

Under Assumption 5.4 the conditions of Lemma 5.2 are satisfied. If these conditions are satisfied, Lemmas 5.1 and 5.2 can be combined to analyze the behavior of $JK(\beta_0)$ statistics in local neighborhoods of the null.

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 in Lemma 3.1, the result in Theorem 5.1 does not require $JK_G(\beta_0)$ to have a stable limiting distribution under H_1.

5.3 IMPROVING POWER AGAINST CERTAIN ALTERNATIVES

As discussed in Section 4.1, tests based on the jackknife K-statistic may suffer from suboptimal power properties. These properties are particularly bad 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* $\ell \in [d_x]$, $\mathbb{E}[\widehat{\Pi}^I_{\ell i}] = 0$ for all $i \in [n]$. Under the assumption that $\mathbb{E}[\widehat{\Pi}^I_{\ell i}] = 0, \forall i \in [n], \ell \in [d_x]$, 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:

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

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 $S(\beta_0)$.

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

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 $1 - \delta_n$,*

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

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$^{\text{th}}$ 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 seek to determine the effect of social spillovers in movie consumption. The sample consists of all 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 all movies wide released in theaters in the United States.[2] The data are obtained through Box Office Mojo, a subsidiary of the Internet Movie Database (IMDb). To focus on movies in theaters long enough for social spillovers to be a relevant factor, the authors consider only movies that remain in theaters for at least six weeks.

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

[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.

of and demand for movies, a vector of date controls are included. Formally, Gilchrist and Sands (2016) are interested in the parameters β_w, $w = 2, \ldots, 7$ from the linear IV model(s):

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

where, for $i = 1, \ldots, 6$, $\text{Sales}_{wi}^{\perp}$ represents gross national ticket sales, after the partialing out of date controls and a constant, $7w$ days after day i, of movies that opened on the opening weekend of i. The variable $\text{Sales}_{7i}^{\perp} = \sum_{w=1}^{6} \text{Sales}_{wi}^{\perp}$ denotes the cumulative national ticket sales from 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; more people seeing a movie on its opening weekend will mean more people telling their friends about the movie potentially leading to larger sales later on.

Because movies with high first-week sales may have high sales in succeeding weeks for reasons other than word of mouth spillover effects (e.g the movie may receive positive critical reviews prerelease or be part of a previously successful franchise), the parameter β_w cannot be plausibly recovered from ordinary least squares regression of $\text{Sales}_{wi}^{\perp}$ on $\text{Sales}_{1i}^{\perp}$. To identify the structural parameter, Gilchrist and Sands (2016) employ a vector of nationally aggregated weather measures. These weather measures reflect the proportion of movie theaters experiencing a particular type of weather on a particular weekend. The 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.

The nationally aggregated weather conditions on opening weekend days serve as plausibly exogenous instrumental variables, affecting ticket sales in later weeks only through their effect on opening-weekend-day sales. Same-day weather conditions may also have an effect on movie ticket sales: when the weather is particularly nice, people may be more inclined to engage in outdoor activities while in poorer, weather people may choose to stay indoors and see a new movie. Putting together the nationally aggregated weather measures leaves Gilchrist and Sands (2016) with a vector of 52 instrumental variables. After the partialing out a constant and the date controls, four of these are linearly dependent. I discard these and work with the remaining 48 partialed-out instruments in my analysis.

To handle the large number of instruments, the authors follow Belloni et al. (2012) and employ a post-LASSO estimate of the first stage. In their main specifications, 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.[3] 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.

Figure 6.1 provides evidence from a simple simulation experiment to demonstrate this. For the simulation experiment I generate an iid sample of 10 instrumental variables, $\{Z_{1i}, \ldots, Z_{10i}\}_{i=1}^{n}$ from a normal distribution with a Toeplitz covariance structure, $\text{Cov}(Z_{\ell i}, Z_{ki}) = 2^{-|j-k|}$, $1 \leq j, k \leq 10$. The endogenous variable is generated to only have a weak relationship with the instruments $X_i = \frac{1}{\sqrt{n}} \sum_{\ell=1}^{10} 0.7 \cdot Z_{\ell i} + v_i$, where the first-stage errors v_i are independent standard normals. From this initial set of 10 instrumental variables I generate an additional 55 technical

[3] Typical empirical practice is to use the Wald test when the first stage F-statistic is larger than 10.

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 over one thousand simulations. As seen in Figure 6.1, these first-stage F-statistics increase significantly as the number of selected instruments decreases. While the "true" F-statistic, computed with only the 10 initial instruments directly relevant for the first stage, is only 5.234, the average F-statistic on the selected variables can be larger than 40. The persistence of this pattern between sample sizes $n = 500$ and $n = 1000$ suggests that this is not a small-sample issue and that pretesting for weak identifications based on post-LASSO F-statistics may be problematic generally. Figure 6.2 shows how the first stage F-statistic changes with the number of LASSO-selected variables in the Gilchrist and Sands (2016) data. The pattern is similar to that seen in the Figure 6.1 simulation experiment.

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.

Table 2 compares the 95% confidence intervals for $\beta_1, \ldots, \beta_7$ generated by the jackknife K test to the confidence intervals generated by the sup-score test of Belloni et al. (2012) and the jackknife LM test (JLM) test of Matsushita and Otsu (2022). I form these confidence intervals by running the tests for each β_0 on a 300 point grid between zero and two and inverting the results; a point β_0 is included in the 95% confidence interval if the test fails to reject the null that $\beta_w = \beta_0$ at level $\alpha = 0.05$. For the $JK(\beta_0)$ statistic I use the choice of hat matrix in (2.3) and estimate the auxiliary parameter $\rho(z_i)$ as in (2.2). The penalty parameter λ is chosen with leave-one-out cross-validation using the `cv.glmnet` command from the `glmnet` package in R (R Core Team, 2021; Friedman et al., 2010). The critical value for the sup-score statistic $S(\beta_0)$ is simulated using 2,500 bootstrap draws. Confidence intervals based on the combination test, $T(\beta_0; \tau)$, are not directly reported as the pretesting procedure based on simulating the 75^{th} quantile of C as in (4.4) always suggests using the $JK(\beta_0)$ statistic.

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. 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 significantly 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 the parameters $\beta_2, \beta_4, \beta_5$, and β_6, the confidence intervals generated by the sup-score statistic are empty while the sup-score confidence interval for β_2 is nearly empty. This is also the case when using the jackknife AR-statistic of Crudu et al. (2021) and Mikusheva and Sun (2021), whose confidence intervals are not reported as they are always empty. With 48 instruments and a single parameter the linear IV model in (6.1) is overidentified and as such

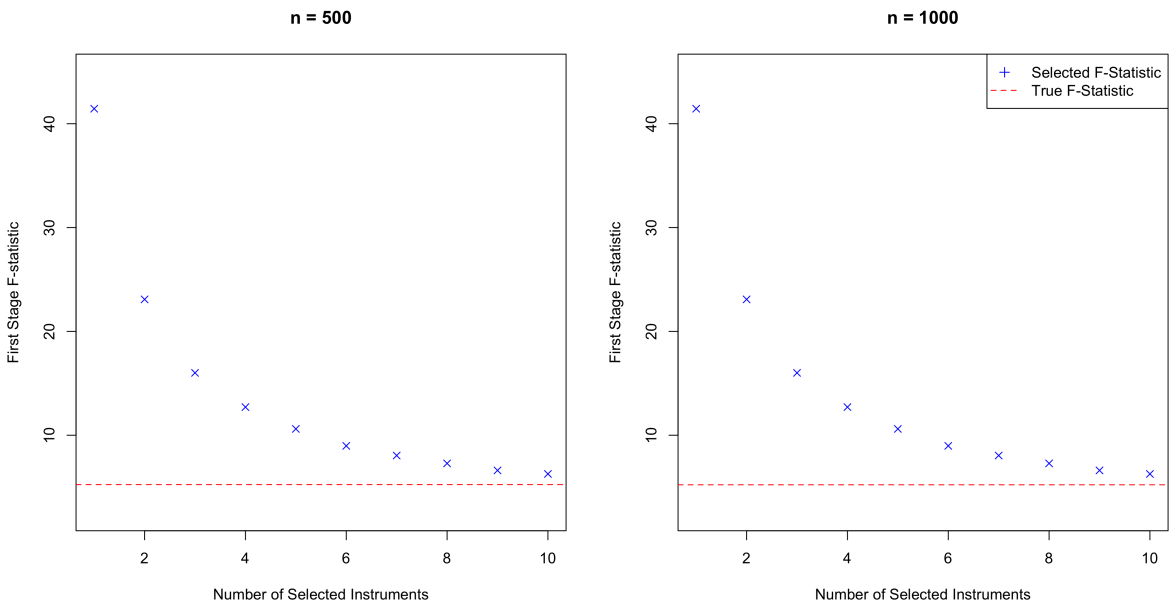

Figure 6.1: Results from Simulation Experiment. The endogenous variable is generated to be weakly related to a set of ten initial instruments. I take quadratic powers and interactions of these ten initial instruments to create an additional 55 technical instruments that do not directly enter the first stage. The LASSO penalty is then set to select a certain number of variables and I report the resulting average post-LASSO F-statistics over 1000 simulations. The average F-statistic using only the relevant ten initial instruments is 5.234 for both $n = 500$ and $n = 1000$.

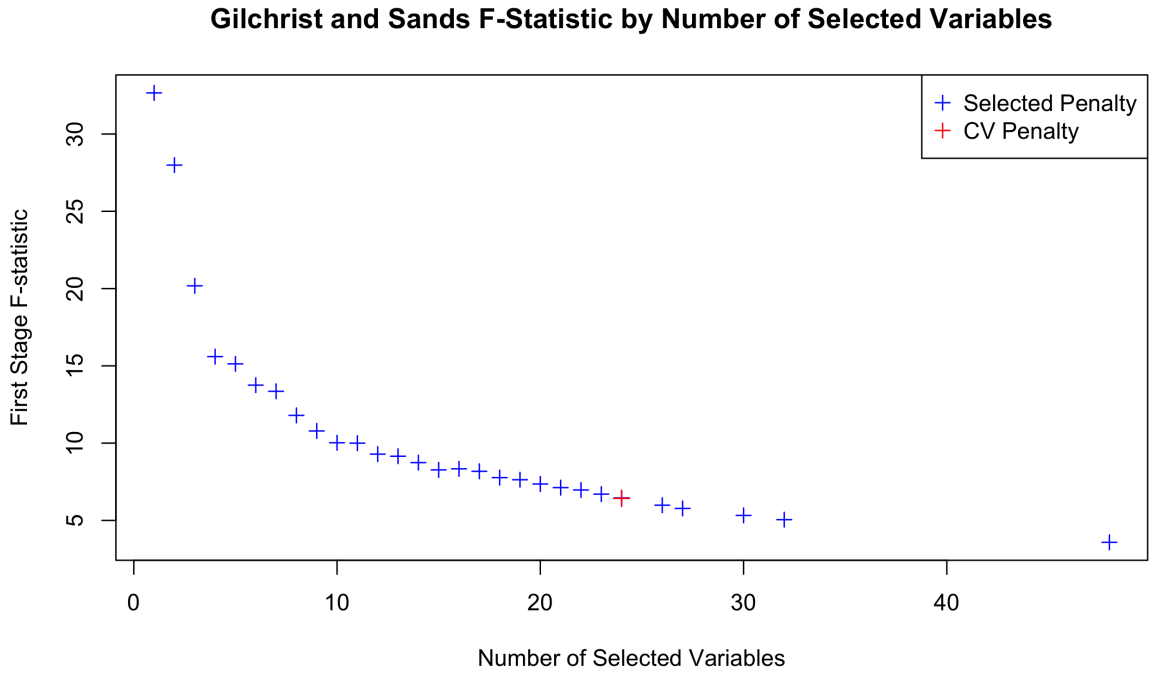

Figure 6.2: First-Stage F-statistic as Function of Number of LASSO-Selected Variables in the Data of Gilchrist and Sands (2016). When selecting variables using a cross-validated choice of LASSO penalty parameter, the first-stage F-statistic is 6.42.

the empty confidence intervals could be interpreted as evidence of model misspecification. For the parameter β_7 the confidence interval generated by inverting the sup-score statistic is not empty and is instead 36% larger than the $JK(\beta_0)$ confidence interval and 41% larger than the JLM confidence interval. This result suggests that the jackknife K tests and JLM tests may have better power properties than the sup-score test in this setting.

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)
$JK(\beta_0)$	[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]

Table 2: 95% Confidence Intervals based on inverting various test statistics. Instrument set used is the same as the Gilchrist and Sands (2016) instrument set less four collinear instruments; $d_z = 48$ with $n = 1,671$. Thresholding test confidence intervals are not reported as they coincide with confidence intervals for $JK(\beta_0)$.

Tables 3 and 4 repeat the analysis of Table 2 but with alternative instrument sets. The confidence intervals of Table 3 use only 5° Fahrenheit temperature bins ($d_z = 36$) while the confidence intervals of Table 4 include all the instruments used in Table 2 and all interactions between the 5° Fahrenheit temperature bins and the other weather measures for a total of 524 instruments.[4] For the most part, the confidence intervals generated by inverting the jackknife K-statistic are similar across Tables 2-4. The confidence intervals for the jackknife LM statistic however, become much narrower when using the largest set of instruments is used. This is interesting as the results from the $JK(\beta_0)$ test as well as the power analysis in Section 4 seem to suggest that use of the extra isntruments does not lead to better first-stage estimates. Interestingly, the JLM confidence intervals in for β_6, β_7 in Table 4 do not contain the point estimates for β_6 and β_7 from Gilchrist and Sands (2016). As with Table 2, Tables 3 and 4 do not report confidence intervals from $T(\beta_0; \tau)$ as these always agree with the $JK(\beta_0)$ confidence intervals and do not report jackknife AR confidence intervals as these are always empty.

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)
$JK(\beta_0)$	[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]

Table 3: 95% Confidence Intervals based on inverting various test statistics. Instrument set used includes only temperatures measures; $d_z = 36$, with $n = 1,671$. Thresholding test condidence intervals are not reported as they coincide with confidence intervals for $JK(\beta_0)$.

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)
$JK(\beta_0)$	[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 4: 95% Confidence Intervals based on inverting various test statistics. Instrument set used includes the original instrument set along with interactions of the temperature measures set with all other aggregated weather measures; $d_z = 524$, with $n = 1,671$. Thresholding test confidence intervals are not reported as they coincide with confidence intervals for $JK(\beta_0)$.

[4] The instrument set of Table 4 does not include interactions between temperature bins nor interactions between other weather measures.

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

$$\begin{aligned} y_i &= x_i + \epsilon_i \\ x_i &= \boldsymbol{z}_i'\pi + v_i \end{aligned} \tag{7.1}$$

where $\boldsymbol{z}_i = (\bar{z}_{1i}, \bar{z}_{1i}^2, \bar{z}_{1i}^3)'$ is a transformation of an initial set of instruments $\bar{z}_i \in \mathbb{R}^{10}$, which are generated as described below. The value of π varies depending on the strength of identification considered; for strong identification, $\pi = (1,1,1)' \in \mathbb{R}^3$, while under weak identification, $\pi = (1/\sqrt{n}, 1/\sqrt{n}, 1/\sqrt{n})' \in \mathbb{R}^3$. 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 $b = 1$.[1] Since the limiting χ^2 distribution of the jackknife K-statistic is exact 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 to the finite-sample behavior of the test. 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 full set instruments z_i is taken to be equal to $\bar{z}_i$ 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 final 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 performance 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 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 consider two different quantiles of the conditioning statistic under the assumption that $\mathbb{E}[\widehat{\Pi}_i^I] = 0$ for all $i \in [n]$; $\tau_{0.25}$ corresponding to the 25$^{\text{th}}$ quantile and $\tau_{0.75}$ corresponding to the 75$^{\text{th}}$

[1] The Laplace distribution is often referred to as a "double exponential" distribution. If X_1 and X_2 are independently distributed according Exponential(1), then $Y = X_1 - X_2$ has a Laplace distribution with parameters $\mu = 0$ and $b = 1$. If X has a Laplace distribution with parameters $\mu = 0$ and $b = 1$, then $|X| \sim$ Exponential(1).

quantile.

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	0.0468	0.0464	0.0472	0.0448	0.0276	0.0756	0.0380
		0.2	0.6	0.0428	0.0280	0.0404	0.0384	0.0232	0.0780	0.0380
		0.5	0.3	0.0440	0.0260	0.0392	0.0364	0.0292	0.0836	0.0328
		0.5	0.6	0.0456	0.0312	0.0432	0.0412	0.0256	0.0708	0.0396
	30	0.2	0.3	0.0448	0.0100	0.0396	0.0216	0.0100	0.0996	0.0400
		0.2	0.6	0.0592	0.0104	0.0544	0.0348	0.0124	0.0984	0.0428
		0.5	0.3	0.0480	0.0132	0.0400	0.0288	0.0096	0.1116	0.0300
		0.5	0.6	0.0460	0.0136	0.0412	0.0276	0.0100	0.1144	0.0236
	65	0.2	0.3	0.0512	0.0392	0.0464	0.0452	0.0232	0.0788	0.0412
		0.2	0.6	0.0544	0.0340	0.0528	0.0476	0.0252	0.0724	0.0380
		0.5	0.3	0.0528	0.0336	0.0476	0.0476	0.0256	0.0628	0.0400
		0.5	0.6	0.0424	0.0260	0.0428	0.0400	0.0268	0.0780	0.0472
500	10	0.2	0.3	0.0524	0.0488	0.0508	0.0524	0.0372	0.0636	0.0420
		0.2	0.6	0.0512	0.0444	0.0516	0.0476	0.0404	0.0724	0.0508
		0.5	0.3	0.0564	0.0372	0.0544	0.0524	0.0332	0.0740	0.0520
		0.5	0.6	0.0556	0.0420	0.0544	0.0528	0.0312	0.0748	0.0388
	30	0.2	0.3	0.0576	0.0204	0.0520	0.0428	0.0200	0.0864	0.0324
		0.2	0.6	0.0424	0.0136	0.0404	0.0328	0.0244	0.0800	0.0356
		0.5	0.3	0.0504	0.0220	0.0436	0.0376	0.0136	0.0912	0.0284
		0.5	0.6	0.0448	0.0252	0.0408	0.0328	0.0224	0.0848	0.0304
	65	0.2	0.3	0.0460	0.0328	0.0432	0.0408	0.0360	0.0748	0.0460
		0.2	0.6	0.0544	0.0404	0.0528	0.0488	0.0376	0.0732	0.0392
		0.5	0.3	0.0440	0.0348	0.0388	0.0376	0.0348	0.0768	0.0408
		0.5	0.6	0.0528	0.0380	0.0516	0.0468	0.0308	0.0688	0.0404

Table 5: Simulated Size of Identification and Heteroskedasticity Robust Tests under Weak Identification. Each DGP is simulated 2500 times. Critical values of the sup-score statistic and quantiles of the conditioning statistic are calculated using 1000 multiplier bootstrap simulations.

Tables 5 and 6 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 standard Anderson-Rubin test is significantly undersized in almost all setups considered.

The sup-score, jackknife AR, and jackknife LM test all seem to have particularly poor performance when identification is weak and $d_z = 30$, which is the setup with the most correlation between the instruments. While tests based on the jackknife AR statistic seems to overreject in this setting, both the sup-score and and jackknife LM tests can be quite conservative. The size of the sup-score test is always under 0.02 while the size of the JLM test can be under half of the nominal size. Notably, the size properties of the sup-score test do seem to improve in the weak identification and $d_z = 30$ case when the sample size increases from $n = 200$ to $n = 500$.

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	0.2	0.0448	0.0348	0.0448	0.0448	0.0312	0.0704	0.0448
		0.2	0.6	0.0452	0.0308	0.0452	0.0452	0.0308	0.0708	0.0496
		0.5	0.2	0.0440	0.0308	0.0440	0.0440	0.0172	0.0768	0.0472
		0.5	0.6	0.0456	0.0360	0.0456	0.0456	0.0252	0.0728	0.0468
	30	0.2	0.2	0.0500	0.0132	0.0500	0.0500	0.0088	0.0988	0.0300
		0.2	0.6	0.0368	0.0100	0.0368	0.0368	0.0128	0.0976	0.0344
		0.5	0.2	0.0344	0.0124	0.0344	0.0344	0.0096	0.1184	0.0332
		0.5	0.6	0.0424	0.0156	0.0424	0.0416	0.0128	0.1120	0.0304
	65	0.2	0.2	0.0524	0.0304	0.0524	0.0524	0.0284	0.0792	0.0404
		0.2	0.6	0.0456	0.0312	0.0456	0.0456	0.0312	0.0824	0.0500
		0.5	0.2	0.0444	0.0316	0.0444	0.0444	0.0268	0.0764	0.0396
		0.5	0.6	0.0496	0.0368	0.0496	0.0496	0.0260	0.0816	0.0492
500	10	0.2	0.2	0.0480	0.0424	0.0480	0.0480	0.0420	0.0716	0.0500
		0.2	0.6	0.0452	0.0440	0.0452	0.0452	0.0348	0.0664	0.0552
		0.5	0.2	0.0496	0.0436	0.0496	0.0496	0.0300	0.0712	0.0428
		0.5	0.6	0.0476	0.0372	0.0476	0.0476	0.0344	0.0616	0.0436
	30	0.2	0.2	0.0496	0.0224	0.0496	0.0496	0.0264	0.0784	0.0424
		0.2	0.6	0.0464	0.0276	0.0464	0.0464	0.0208	0.0828	0.0440
		0.5	0.2	0.0408	0.0192	0.0408	0.0408	0.0180	0.0940	0.0352
		0.5	0.6	0.0400	0.0224	0.0400	0.0400	0.0220	0.0924	0.0452
	65	0.2	0.2	0.0436	0.0424	0.0436	0.0436	0.0408	0.0676	0.0500
		0.2	0.6	0.0500	0.0468	0.0500	0.0500	0.0364	0.0716	0.0472
		0.5	0.2	0.0484	0.0372	0.0484	0.0484	0.0348	0.0680	0.0472
		0.5	0.6	0.0476	0.0384	0.0476	0.0476	0.0368	0.0704	0.0464

Table 6: Simulated Size of Identification and Heteroskedasticity Robust Tests under Strong Identification. Each DGP is simulated 2500 times. Critical values of the sup-score statistic and quantiles of the conditioning statistic are calculated using 1000 multiplier bootstrap simulations.

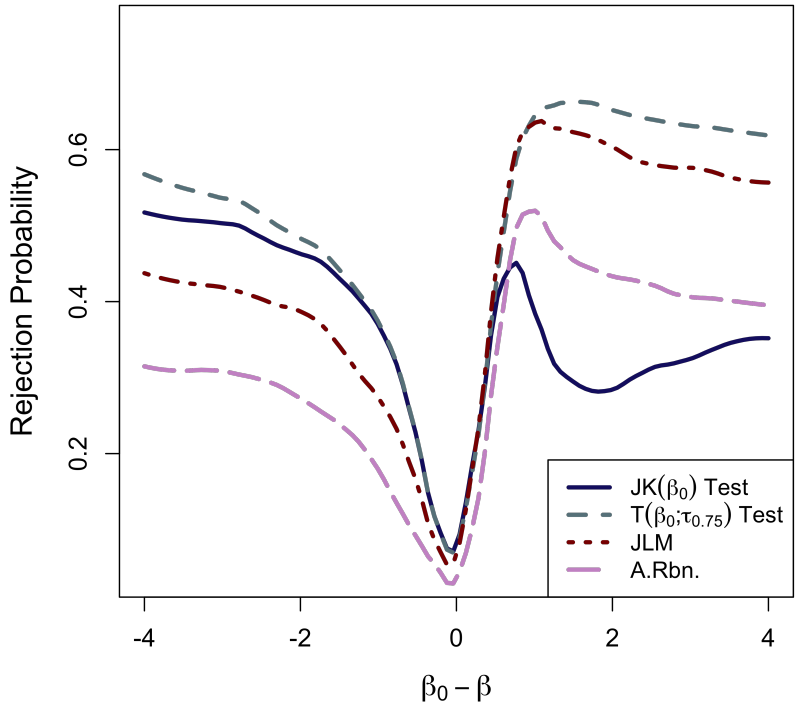

Figure 7.1: Simulated Local Power Curves under Intermediate identification Strength. Sample size is 500, $d_z = 65$, $\varrho_1 = 0.2$ and $\varrho_2 = 0.6$. Local power is calculated on a grid of 100 points between -4 and 4. At each point the DGP is simulated 2000 times.

This is in line with theoretical results showing that the sup-score test has exact asymptotic size under standard conditions. 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. This suggests that the requirement of $d_z \to \infty$ may be important for the quality of finite-sample approximation by its limiting distribution.

Figures 7.1–7.3 plot simulated local power curves under an intermediate-strength identification where the first stage is in a $n^{-1/3}$ neighborhood of zero, $d_z = 65$, $\varrho_1 \in \{0.2, 0.5\}$ and $\varrho_2 \in \{0.3, 0.6\}$. I compare the local power curves of the $JK(\beta_0)$ test, the combination test with cutoff $\tau_{0.75}$, the jackknife LM test, and the standard Anderson-Rubin test. The power curves for the jackknife AR and sup-score statistic are not plotted as the realized size of these tests consistently differers from the nominal size in all of the setups considered. Under alternate DGPs Matsushita and Otsu (2022) and Lim et al. (2022) show that the jackknife AR test seems to have similar overall power to the jackknife LM test, performing somewhat better against distant alternatives and somewhat worse against local alternatives.

In all three cases considered, the test based on the $JK(\beta_0)$ statistic faces a decline in power for values of $(\beta_0 - \beta)$ between one and three. This power decline appears to be avoided by the test based on the combination test. In the DGPs considered, the power curve of the combination test seems to agree with that of the jackknife K test under alternatives local to the null, but weighs the sup-score test more under distant alternatives. In the DGPs considered, the combination test appears to have weakly larger power than the jackknife LM and Anderon-Rubin tests at almost all points and strictly larger power for negative values of $(\beta_0 - \beta)$.

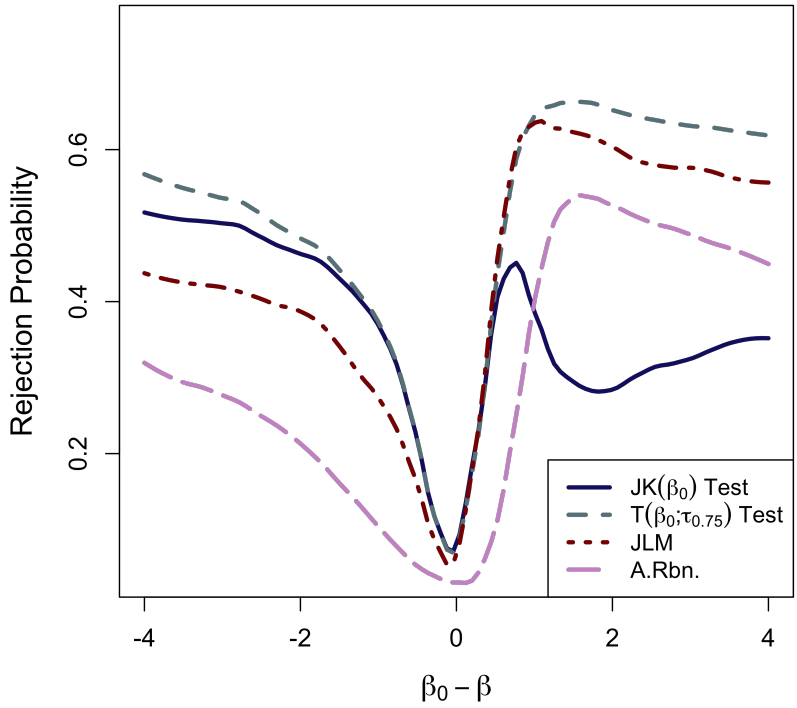

Figure 7.2: Simulated Local Power Curves under Intermediate Identification Strength. Sample size is 500, $d_z = 65$, $\varrho_1 = 0.5$ and $\varrho_2 = 0.3$. Local power is calculated on a grid of 100 points between -4 and 4. At each point the DGP is simulated 2000 times.

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.

8 Conclusion

I propose a new test for the structural parameter in a linear instrumental variables model. This test is based on a jackknife version of the K-statistic and the limiting behavior of the test is analyzed via a novel direct Gaussian approximation argument. I show that, as long as an auxiliary parameter can be consistently estimated, the test is robust to both the strength of identification and the number of instruments; the limiting distribution of the test statistic does not depend on either of these factors. Consistency of the auxiliary parameter can be achieved under approximate sparsity using simple-to-implement ℓ_1-penalized methods.

I characterize the behavior of the jackknife K-statistic in local neighborhoods of the null. To address a power deficiency that tests based on jackknife K-statistic inherit from their non-jackknife namesakes, I propose a testing procedure that decides whether the researcher should run a test via the jackknife K-statistic or one via the sup-score statistic based on the value of a conditioning statistic. While this combination does not fully address the power decline, I show that it works well in a simulation study and leave further refinements to future work.

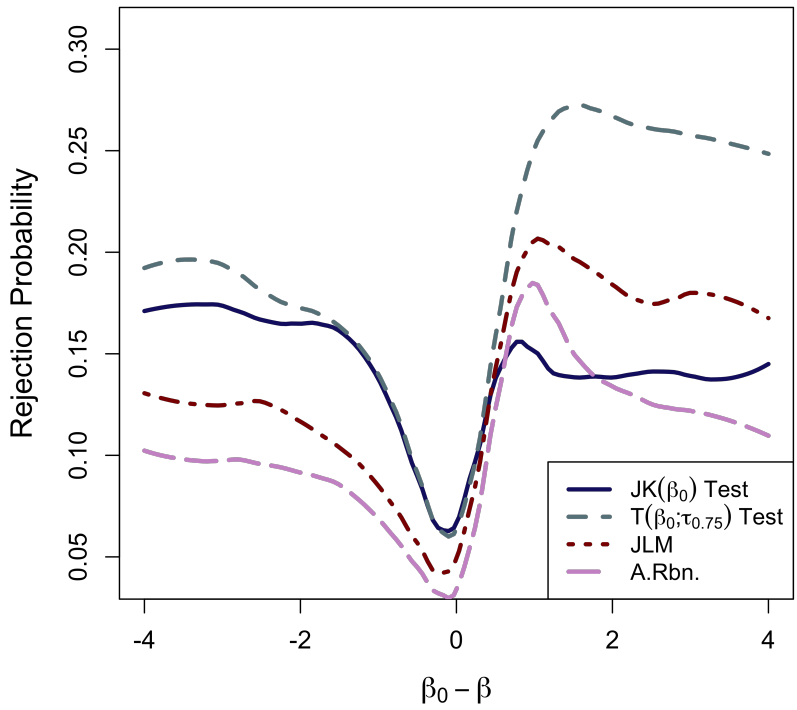

Figure 7.3: Simulated Local Power curves under Intermediate Identification Strength. Sample size is 500, $d_z = 65$, $\varrho_1 = 0.5$ and $\varrho_2 = 0.6$. Local power is calculated on a grid of 100 points between -4 and 4. At each point the DGP is simulated 2000 times.

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., M. Moreira, and J. H. Stock (2004, August). Optimal invariant similar tests for instrumental variables regression. (299).

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.

Bekker, P. A. (1994). Alternative approximations to the distributions of instrumental variable estimators. *Econometrica 62*(3), 657–681.

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.

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

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.

Chao, J. C., N. R. Swanson, J. A. Hausman, W. K. Newey, and T. Woutersen (2012). Asymptotic distribution of jive in a heteroskedastic iv regression with many instruments. *Econometric Theory 28*(1), 42–86.

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

Chatterjee, S. (2010). A new approach to strong embeddings.

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.

Chernozhukov, V., D. Chetverikov, and K. Kato (2017). Central limit theorems and bootstrap in high dimensions. *The Annals of Probability 45*(4), 2309–2352.

Chernozhukov, V., W. K. Newey, and R. Singh (2022). Automatic debiased machine learning of causal and structural effects. *Econometrica 90*(3), 967–1027.

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.

Dobbie, W., J. Goldin, and C. S. Yang (2018, February). The effects of pretrial detention on conviction, future crime, and employment: Evidence from randomly assigned judges. *American Economic Review 108*(2), 201–40.

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.

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.

Han, C. and P. C. B. Phillips (2006). Gmm with many moment conditions. *Econometrica 74*(1), 147–192.

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

Horn, R. and C. Johnson (2012). *Matrix Analysis*. Cambridge University Press.

Jou, A. and T. Morgan (2023). Do relief programs compensate affected populations? evidence from the great depression and the new deal. *Working Paper*.

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.

Kline, P., R. Saggio, and M. Sølvsten (2020). Leave-out estimation of variance components. *Econometrica 88*(5), 1859–1898.

Lee, D. S., J. McCrary, M. J. Moreira, and J. Porter (2022, October). Valid t-ratio inference for iv. *American Economic Review 112*(10), 3260–90.

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.

Maestas, N., K. J. Mullen, and A. Strand (2013, August). Does disability insurance receipt discourage work? using examiner assignment to estimate causal effects of ssdi receipt. *American Economic Review 103*(5), 1797–1829.

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.

Nazarov, F. (2003). *On the Maximal Perimeter of a Convex Set in R^n with Respect to a Gaussian Measure*. Berlin, Heidelberg: Springer Berlin Heidelberg.

Nelson, C. R. and R. Startz (1990). Some further results on the exact small sample properties of the instrumental variable estimator. *Econometrica 58*(4), 967–976.

Newey, W. K. and F. Windmeijer (2009). Generalized method of moments with many weak moment conditions. *Econometrica 77*(3), 687–719.

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.

Petersen, K. B. and M. S. Pedersen (2012, nov). The matrix cookbook. Version 20121115.

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

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Sambale, H. (2022). Some notes on concentration for α-subexponential random variables.

Sampat, B. and H. L. Williams (2019, January). How do patents affect follow-on innovation? evidence from the human genome. *American Economic Review 109*(1), 203–36.

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

Stock, J. and M. Yogo (2005). *Testing for Weak Instruments in Linear IV Regression*. New York: Cambridge University Press.

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.

van Wieringen, W. N. (2023). Lecture notes on ridge regression.

A Proofs of Results in Section 3

A.1 Proof of Lemma 3.1

The statement $\sup_{a<0}\left|\Pr(JK_I(\beta_0) \le a) - \Pr(JK_G(\beta_0) \le a)\right| = 0$ is immediate since both $JK_I(\beta_0)$ and $JK_G(\beta_0)$ are always weakly positive. It thus suffices to show

$$\sup_{a\ge 0}\left|\Pr(JK_I(\beta_0) \le a) - \Pr(JK_G(\beta_0) \le a)\right| \to 0$$

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. Recall that $\tilde{h}_{ii} = 0$ 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 $\kappa_i^2(\beta_0) = \mathbb{E}[\epsilon_i^2(\beta_0)]$. Under this notation we can write $JK_I(\beta_0) = \frac{N^2}{D}\mathbf{1}_{\{D>0\}}$ and $JK_G(\beta_0) = \frac{\tilde{N}^2}{\tilde{D}}$. Dealing with these forms of the statistics is difficult for the interpolation argument, since the denominator is random. Instead, we will notice that since $D = 0 \implies N = 0$ and $\Pr(\tilde{D} > 0) = 1$, for any $a \ge 0$ we can rewrite the events

$$\{JK_I(\beta_0) \le a\} = \{N^2 - aD \le 0\} \text{ and } \{JK_G(\beta_0) \le a\} \overset{\text{a.s}}{=} \{\tilde{N}^2 - a\tilde{D} \le 0\} \tag{A.1}$$

With this in mind define

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

Showing Lemma 3.1 is then equivalent to showing that $\sup_a |\Pr(JK^a \le 0) - \Pr(\tilde{JK}^a \le 0)| \to 0$. We do so in a few lemmas, the final result being shown in Lemma A.3 at the bottom of this subsection.

Lemma A.1 (Lindeberg Interpolation)**.** *Suppose that Assumptions 3.1–3.3 hold. Let $\varphi(\cdot) : \mathbb{R} \to \mathbb{R}$ be such that $\varphi(\cdot) \in C_b^3(\mathbb{R})$ with $L_2(\varphi) = \sup_x |\varphi''(x)|$ and $L_3(\varphi) = \sup_x |\varphi'''(x)|$. Then, there is a constant M that depends only on the constant c such that:*

$$|\mathbb{E}[\varphi(JK^a) - \varphi(\tilde{JK}^a)]| \le \frac{M(a^3 \vee 1)}{\sqrt{n}}(L_2(\varphi) + L_3(\varphi))$$

Proof of Lemma A.1. Begin by defining the leave-one-out numerator, denominator, and decomposed statistics

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

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

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 the dependence of these variables on i to simplify notation.

Next, define the one-step deviations

$$
\begin{aligned}
\Delta_{1i} &:= \epsilon_i(\beta_0)\sum_{j=1}^{n}\tilde{h}_{ij}\dot{r}_j + r_i\sum_{j=1}^{n}\tilde{h}_{ji}\dot{\epsilon}_j(\beta_0) \\
\tilde{\Delta}_{1i} &:= \tilde{\epsilon}_i(\beta_0)\sum_{j=1}^{n}\tilde{h}_{ij}\dot{r}_j + \tilde{r}_i\sum_{j=1}^{n}\tilde{h}_{ji}\dot{\epsilon}_j(\beta_0) \\
\Delta_{2i} &:= \underbrace{a\epsilon_i^2(\beta_0)(\sum_{j=1}^{n}\tilde{h}_{ij}\dot{r}_j)^2 + ar_i^2\sum_{j=1}^{n}\tilde{h}_{ji}^2\ddot{\epsilon}_j^2(\beta_0)}_{\Delta_{2i}^a} + \underbrace{2ar_i\sum_{j=1}^{n}\ddot{\epsilon}_j^2(\beta_0)\sum_{\ell\neq i}\tilde{h}_{j\ell}\tilde{h}_{ji}\dot{r}_\ell}_{\Delta_{2i}^b} \\
\tilde{\Delta}_{2i} &:= \underbrace{a\kappa_i^2(\beta_0)(\sum_{j=1}^{n}\tilde{h}_{ij}\dot{r}_j)^2 + a\tilde{r}_i^2\sum_{j=1}^{n}\tilde{h}_{ji}^2\ddot{\epsilon}_j^2(\beta_0)}_{\Delta_{2i}^a} + \underbrace{2a\tilde{r}_i\sum_{j=1}^{n}\ddot{\epsilon}_j^2(\beta_0)\sum_{\ell\neq i}\tilde{h}_{j\ell}\tilde{h}_{ji}\dot{r}_\ell}_{\tilde{\Delta}_{2i}^b}
\end{aligned}
\tag{A.2}
$$

These one-step deviations contain all the terms associated with observation i in the expression of the numerator and denominator of the test statistics. To demonstrate, note that these one-step deviations satisfy $N_{-1} + n^{-1/2}\Delta_{11} = N$ and $aD_{-1} + n^{-1}\Delta_{21} = aD$ as

$$
\begin{aligned}
N &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\epsilon_i(\beta_0)\sum_{j=1}^{n}\tilde{h}_{ij}r_j \\
&= \frac{1}{\sqrt{n}}\sum_{j>1}\epsilon_j(\beta_0)\sum_{\ell=1}^{n}\tilde{h}_{j\ell}r_j + \epsilon_1(\beta_0)\frac{1}{\sqrt{n}}\sum_{j>1}\tilde{h}_{1j}r_j \\
&= \frac{1}{\sqrt{n}}\sum_{j>1}\epsilon_j(\beta_0)\Big\{\tilde{h}_{j1}r_1 + \sum_{\ell>1}h_{j\ell}r_\ell\Big\} + \epsilon_1(\beta_0)\frac{1}{\sqrt{n}}\sum_{j>1}\tilde{h}_{1j}r_j \\
&= \underbrace{\frac{1}{\sqrt{n}}\sum_{j>1}\epsilon_j(\beta_0)\sum_{\ell>1}h_{j\ell}r_\ell}_{N_{-1}} + \underbrace{\epsilon_1(\beta_0)\frac{1}{\sqrt{n}}\sum_{j>1}\tilde{h}_{1j}r_j + r_1\frac{1}{\sqrt{n}}\sum_{j>1}\tilde{h}_{j1}\epsilon_j(\beta_0)}_{n^{-1/2}\Delta_{11}}
\end{aligned}
$$

and

$$
\begin{aligned}
D &= \frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2(\beta_0)(\sum_{j=1}^{n}\tilde{h}_{ij}r_j)^2 \\
&= \frac{1}{n}\sum_{j>1}\epsilon_j^2(\beta_0)(\sum_{\ell=1}^{n}\tilde{h}_{j\ell}r_\ell)^2 + \epsilon_1^2(\beta_0)\frac{1}{n}(\sum_{j>1}\tilde{h}_{1j}r_j)^2 \\
&= \frac{1}{n}\sum_{j>1}\epsilon_j^2(\beta_0)(\tilde{h}_{j1}r_1 + \sum_{\ell\neq 1}\tilde{h}_{\ell j}r_\ell)^2 + \epsilon_1^2(\beta_0)\frac{1}{n}(\sum_{j>1}\tilde{h}_{1j}r_j)^2 \\
&= \underbrace{\frac{1}{n}\sum_{j>1}\epsilon_j^2(\beta_0)(\sum_{\ell>1}\tilde{h}_{\ell,j}r_\ell)^2}_{D_{-1}}
\end{aligned}
$$

$$\underbrace{+\epsilon_1^2(\beta_0)\frac{1}{n}\Big(\sum_{j>1}\tilde{h}_{1j}r_j\Big)^2 + r_1^2\frac{1}{n}\sum_{j>1}\tilde{h}_{j1}^2\epsilon_j^2(\beta_0) + 2r_1\frac{1}{n}\sum_{j>1}\epsilon_j^2(\beta_0)\sum_{\ell>1}\tilde{h}_{\ell j}r_\ell}_{(an)^{-1}\Delta_{21}}$$

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

$$\begin{aligned}\mathbb{E}[\varphi(JK^a) - \varphi(J\tilde{K}^a)] = \sum_{i=1}^{n} &\mathbb{E}[\varphi(JK_{-i} + n^{-1/2}N_{-i}\Delta_{1i} + n^{-1}\Delta_{1i}^2 - n^{-1}\Delta_{2i})] \\ &- \mathbb{E}[\varphi(JK_{-i} + n^{-1/2}N_{-i}\tilde{\Delta}_{1i} + n^{-1}\tilde{\Delta}_{1i}^2 - n^{-1}\tilde{\Delta}_{2i})]\end{aligned} \tag{A.3}$$

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

$$\begin{aligned}\mathbb{E}[\Delta_{1i} - \tilde{\Delta}_{1i}|\mathcal{F}_{-i}] = \mathbb{E}[\Delta_{1i}^2 - \tilde{\Delta}_{1i}^2|\mathcal{F}_{-i}] &= \mathbb{E}[\Delta_{2i} - \tilde{\Delta}_{2i}|\mathcal{F}_{-i}] \\ = \mathbb{E}[\Delta_{2i}^b - \tilde{\Delta}_{2i}^b|\mathcal{F}_{-i}] &= \mathbb{E}[\Delta_{1i}\Delta_{2i}^b - \tilde{\Delta}_{1i}\tilde{\Delta}_{2i}^b|\mathcal{F}_{-i}] = 0\end{aligned} \tag{A.4}$$

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 write

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

Applications of Lemmas F.1 and F.2, Cauchy-Schwarz, and thte generalized Hölder inequality,[1]

[1] $\mathbb{E}[|fgk|]^3 \le \mathbb{E}[|f|^3]\mathbb{E}[|g|^3]\mathbb{E}[|k|^3]$

will allow us to bound for a fixed constant M that depends only on c,

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

and

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

Combining these bounds and summing over n gives the result. □

Lemma A.2 (Gaussian Denominator Anti-Concentration)**.** *Suppose that Assumptions 3.1 and 3.2 hold. Then for any sequence $\delta_n \searrow 0$,*

$$\Pr(\tilde{D} \leq \delta_n) \to 0$$

Proof of Lemma A.2. 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

$$\begin{aligned} \Pr(\tilde{D} \leq \delta_n) &\leq \Pr\Big(\frac{1}{cn} \sum_{i=1}^n \Big(\sum_{j=1}^n \tilde{h}_{ij} \tilde{r}_j \Big)^2 \leq \tilde{\delta}_n \Big) \\ &= \Pr\big(\|\tilde{r}' \bar{H}^{1/2}\|^2 \leq \delta_n \big) \end{aligned} \tag{A.5}$$

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.5), 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 $c^{-1} < \mathbb{E}[r' \bar{H} r] < c$:

$$\begin{aligned} \Pr(\tilde{r}' \bar{H} \tilde{r} \leq \delta_n) &= \Pr(\tilde{r}' \bar{H} \tilde{r} - \mathbb{E}[\tilde{r}' \bar{H} \tilde{r}] \leq \delta_n - \mathbb{E}[\tilde{r}' \bar{H} \tilde{r}]) \\ &\leq \Pr(\tilde{r}' \bar{H} \tilde{r} - \mathbb{E}[r' \bar{H} r] \geq \mathbb{E}[\tilde{r}' \bar{H} \tilde{r}] - \delta_n) \\ &\leq \Pr(|\tilde{r}' \bar{H} \tilde{r} - \mathbb{E}[r' \bar{H} r]| \geq \frac{1}{2c}) \\ &\leq 2c \operatorname{Var}(r' \bar{H} r) \end{aligned} \tag{A.6}$$

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

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 $\boldsymbol{R} = \operatorname{diag}(\operatorname{Var}(r_1), \ldots, \operatorname{Var}(r_n))$. 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 G.5, Λ_2 / Λ_1 is bounded away from zero. Using Theorem H.4 we can then bound for some constant $C > 0$

$$\Pr(\|r' H\|^{1/2} \leq \delta_n) \leq C \delta_n \Lambda_1^{-1} \tag{A.7}$$

Combining Bounds. To combine the bounds in (A.6) and (A.7), first write

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

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:

$$\begin{aligned}
\mu_r'\bar{H}\boldsymbol{R}\bar{H}\mu_r &= (\mu_r'\bar{H}^{1/2})'(\bar{H}^{1/2}\boldsymbol{R}\bar{H}^{1/2})(\bar{H}^{1/2}\mu_r) \\
&\leq \lambda_1(\bar{H}^{1/2}\boldsymbol{R}\bar{H}^{1/2})\|\mu_r'\bar{H}^{1/2}\|^2 \\
&= \sqrt{\lambda_1^2(\bar{H}^{1/2}\boldsymbol{R}\bar{H}^{1/2})}\|\mu_r'\bar{H}^{1/2}\|^2 \\
&= \sqrt{\lambda_1(\bar{H}^{1/2}\boldsymbol{R}\bar{H}\boldsymbol{R}\bar{H}^{1/2})}\|\mu_r'\bar{H}^{1/2}\|^2 \\
&\leq \sqrt{\text{trace}(\bar{H}^{1/2}\boldsymbol{R}\bar{H}\boldsymbol{R}\bar{H}^{1/2})}\|\mu_r'\bar{H}^{1/2}\|^2 \\
&= \sqrt{\text{trace}(\boldsymbol{R}\bar{H}\boldsymbol{R}\bar{H})}\|\mu_r'\bar{H}\|^2 \leq c^2\Lambda_1^{1/2}
\end{aligned} \tag{A.8}$$

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.6), (A.7), and (A.8) gives us

$$\Pr(\tilde{D} \leq \delta_n) \leq C\min\left\{\Lambda_1 + \Lambda_1^{1/2}, \delta_n\Lambda_1^{-1}\right\} \tag{A.9}$$

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

Remark A.1 (Final Anticoncentration Bound)**.** To give an explicit bound on (A.9) in terms of δ_n we note that, if $x^\star$ solves

$$x^\star + \sqrt{x^\star} = \frac{c}{x^\star}$$

then for any $x \geq 0$, $\min\{x + \sqrt{x}, c/x\} \leq x^\star + \sqrt{x^\star}$. Using this, notice that $(x^\star)^2 + (x^\star)^{3/2} = c$ so that $x^\star \leq \sqrt{c}$. This allows us to bound (A.9)

$$\Pr(\tilde{D} \leq \delta_n) \leq C\min\{\Lambda_1 + \Lambda_1^{1/2}, \delta_n\Lambda_1^{-1}\} \leq C(\delta_n^{1/2} + \delta_n^{1/4})$$

Lemma A.3 (Approximate Distribution)**.** *Under Assumptions 3.1–3.3*

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

Proof of Lemma A.3. First, fix a $\Delta \geq 0$ and consider any $a \leq \Delta$. As in Lemma A.2, 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 Lemma A.1 we can write for some constant M that depends only on Δ:

$$\Pr(JK_I(\beta_0) \leq a) = \Pr(JK^a \leq 0) \leq \mathbb{E}[\varphi_n(JK^a)]$$

$$\leq \mathbb{E}[\varphi_n(J\tilde{K}^a)] + \frac{M}{\sqrt{n}}(\gamma_n^2 + \gamma_n^{-3})$$
$$\leq \Pr(J\tilde{K}^a \leq 0) + \Pr(0 \leq \tilde{N}^2 - a\tilde{D} \leq \gamma_n) + \frac{M}{\sqrt{n}}(\gamma_n^2 + \gamma_n^{-3})$$

Applying Lemma A.4 and $\{J\tilde{K}^a \leq 0\} = \{JK_G(\beta_0) \leq a\}$ gives:

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

By Lemma F.3, we can bound $\mathbf{A} \leq M\gamma_n^{1/2}$ while by Lemma A.2 and Remark A.1, $\mathbf{B} \leq M\gamma_n^{1/4}$. 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

$$\Pr(JK_G(\beta_0) \leq a) - \boldsymbol{e} \leq \Pr(JK_I(\beta_0) \leq a) \leq \Pr(JK_G(\beta_0) \leq a) + \boldsymbol{e}$$

where

$$\boldsymbol{e} = M\Big(\frac{\gamma_n^{-2} + \gamma_n^{-3}}{\sqrt{n}} + \gamma_n^{1/2} + \gamma_n^{1/4}\Big) = o(1)$$

Since the constant M depends only on Δ, this gives us that for any fixed $\Delta > 0$

$$\sup_{a \leq \Delta} \big|\Pr(JK_I(\beta_0) \leq a) - \Pr(JK_G(\beta_0) \leq a)\big| \leq C\Big(\frac{\gamma_n^{-2} + \gamma_n^{-3}}{\sqrt{n}} + \gamma_n^{1/2} + \gamma_n^{1/4}\Big) = o(1) \tag{A.10}$$

where C is a constant that depends only on Δ. 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 Lemma A.2, we can apply Lemma A.6. This step shows that the result in (A.10) implies that the approximation error tends to zero uniformly over the real line, which is the desired result. Optimizing over γ_n in the expression of (A.10) yields the rate of decay in Remark 3.3. □

Lemma A.4. *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

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

Proof. Define the event $\Omega = \{Y_n \geq \delta^{1/2}\}$. We can bound

$$\begin{aligned}\Pr(0 \leq X_n - cY_n \leq \delta) &= \Pr(cY_n \leq X_n \leq \delta + cY_n) \\ &\leq \Pr(\{cY_n \leq X_n \leq \delta + cY_n\} \cap \Omega) + \Pr(\Omega^c) \\ &= \Pr(\{c \leq W_n \leq \delta/Y_n + c\} \cap \Omega) + \Pr(\Omega^c) \\ &\leq \Pr(c \leq W_n \leq \delta^{1/2} + c) + \Pr(\Omega^c)\end{aligned}$$

The second statement of the lemma follows symmetrically. □

Lemma A.5. *Suppose that X_n and Y_n are sequences of (real-valued) random variables such that $Y_n = O_p(1)$ and for any $x \in \mathbb{R}$*

$$|\Pr(X_n \leq x) - \Pr(Y_n \leq x)| \to 0$$

Then $X_n = O_p(1)$.

Proof. Pick any $\epsilon > 0$, and let $M_{\epsilon/2}$ be such that $\Pr(Y_n > M_{\epsilon/2}) \leq \epsilon/2$ for all $n \geq N_\epsilon$. In addition, let $\tilde{N}_\epsilon$ be such that $|\Pr(X_n \leq M_{\epsilon/2}) - \Pr(Y_n \leq M_{\epsilon/2})| \leq \epsilon/2$ for all $n \geq \tilde{N}_\epsilon$. Then for all $n \geq N_\epsilon \vee \tilde{N}_{\epsilon/2}$,

$$\begin{aligned}\Pr(X_n > M_{\epsilon/2}) &\leq \Pr(Y_n > M_{\epsilon/2}) + |\Pr(X_n > M_{\epsilon/2}) - \Pr(Y_n > M_{\epsilon/2})| \\ &\leq \epsilon/2 + |\Pr(Y_n \leq M_{\epsilon/2}) - \Pr(X_n \leq M_{\epsilon/2})| \\ &\leq \epsilon/2 + \epsilon/2 = \epsilon\end{aligned}$$

□

Lemma A.6. *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 $\sup_{x \in \mathbb{R}} |\Pr(X_n \leq x) - \Pr(Y_n \leq x)| \to 0$.

Proof. Pick an $\epsilon > 0$. By Lemma A.5, $X_n = O_p(1)$. Pick a constant $M_{\epsilon/3}$ such that $\Pr(X_n > M_{\epsilon/3}) \leq \epsilon/3$ and $\Pr(Y_n > M_{\epsilon/3}) \leq \epsilon/3$. Then for any $x \in \mathbb{R}$ we can bound $|\Pr(X_n \leq x) - \Pr(Y_n \leq x)|$ by considering two cases:

Case 1. If $x \leq M_{\epsilon/3}$, then,

$$|\Pr(X_n \leq x) - \Pr(Y_n \leq x)| \leq \sup_{x \leq M_{\epsilon/3}} |\Pr(X_n \leq x) - \Pr(Y_n \leq x)| \tag{A.11}$$

by hypothesis, there is an N_ϵ such that for $n \geq N_\epsilon$ the RHS of (A.11) is less than ϵ.

Case 2. If $x > M_{\epsilon/3}$ we can bound

$$\begin{aligned}|\Pr(X_n \leq x) - \Pr(Y_n \leq x)| &\leq |\Pr(X_n \leq M_{\epsilon/3}) - \Pr(Y_n \leq M_{\epsilon/3})| \\ &\quad + |\Pr(M_{\epsilon/3} < X_n \leq x) - \Pr(M_{\epsilon/3} < Y_n \leq x)| \\ &\leq |\Pr(X_n \leq M_{\epsilon/3}) - \Pr(Y_n \leq M_{\epsilon/3})| + \epsilon/3 + \epsilon/3 \end{aligned} \tag{A.12}$$

By hypothesis, there is an $N_{\epsilon/3}$ such that $|\Pr(X_n \leq M_{\epsilon/3}) - \Pr(Y_n \leq N_{\epsilon/3})| \leq \epsilon/3$.

WLOG $N_{\epsilon/3} \geq N_\epsilon$. Combining the bounds in (A.11) and (A.12), for any $n \geq N_{\epsilon/3}$ and any $x \in \mathbb{R}$,

$$|\Pr(X_n \leq x) - \Pr(Y_n \leq x)| \leq \epsilon$$

Since this holds for all x, this gives the result. □

A.2 PROOF OF PROPOSITION 3.1

Proof of Proposition 3.1. As at the top of Appendix A.1, recall that $\tilde{h}_{ii} = 0$, and define

$$N = \frac{1}{\sqrt{n}} \sum_{i=1}^n \epsilon_i(\beta_0) \sum_{j=1}^n \tilde{h}_{ij} r_j \qquad D = \frac{1}{n} \sum_{i=1}^n \epsilon_i^2(\beta_0) (\sum_{j=1}^n \tilde{h}_{ij} r_j)^2$$

where $\tilde{h}_{ij} = s_n h_{ij}$. The goal is to show that $\Pr(JK_I(\beta_0) \leq a) \to 0$ for any fixed $a \in \mathbb{R}_+$. The event $\{JK_I(\beta_0) \leq a\}$ is equivalently expressed $\{N^2 - aD \leq 0\}$ so that $\Pr(JK(\beta_0) \leq a) = \Pr(N^2 - aD \leq 0)$. Under Assumptions 3.1 and 3.2, $aD = O_p(1)$ so by Lemma A.8 it suffices to show that $\Pr(|N| \leq M) \to 0$ for any fixed $M \geq 0$. By assumption $P = \mathbb{E}[N^2] \to \infty$ so we move to show that $\mathrm{Var}(N) = O(1)$ and then apply Lemma A.7 to conclude. To this end, recall the definition of $\eta_i = \epsilon_i(\beta_0) - \mathbb{E}[\epsilon_i(\beta_0)]$, define $\mu_i = \mathbb{E}[\epsilon_i(\beta_0)] = \Pi_i(\beta - \beta_0)$, and let

$$N_1 := \frac{1}{\sqrt{n}} \sum_{i=1}^n \eta_i \sum_{j=1}^n \tilde{h}_{ij} r_j \qquad\qquad N_2 := \frac{1}{\sqrt{n}} \sum_{i=1}^n \mu_i \sum_{j=1}^n \tilde{h}_{ij} r_j$$

Notice that $N = N_1 + N_2$. To show that $\mathrm{Var}(N_1) = O(1)$, define $\boldsymbol{a}_i = \eta_i \sum_{j=1}^n \tilde{h}_{ij} r_j$. Since $\mathbb{E}[\eta_i r_i] = 0$, we have that $\mathrm{Cov}(\boldsymbol{a}_i, \boldsymbol{a}_j) = 0$ for $i \neq j$. Thus,

$$\mathrm{Var}(N_1) = \mathrm{Var}(\sum_{i=1}^n \boldsymbol{a}_i/\sqrt{n}) = n^{-1} \sum_{i=1}^n \mathrm{Var}(\boldsymbol{a}_i) = n^{-1} \sum_{i=1}^n \mathrm{Var}(\eta_i) \mathbb{E}[(\sum_{j=1}^n \tilde{h}_{ij} r_j)^2] \leq c^2$$

where the final inequality follows from an upper bound on $\mathrm{Var}(\eta_i)$ from Assumption 3.1 and by definition of $\tilde{h}_{ij} = s_n h_{ij}$ from Assumption 3.2.

To show that $\mathrm{Var}(N_2) = O(1)$ let $\boldsymbol{b}_i = \sum_{j=1}^n \tilde{h}_{ji} \tilde{\Pi}_j (\beta - \beta_0)$ and rewrite $N_2 = \frac{1}{\sqrt{n}} \sum_{i=1}^n r_i \boldsymbol{b}_i$. Under Assumption 3.3(ii), $|\boldsymbol{b}_i| = |\mathbb{E}[\sum_{j=1}^n \tilde{h}_{ji} \epsilon_j(\beta_0)]| \leq c^{1/2}$, so we can bound

$$\mathrm{Var}(N_2) = \mathrm{Var}(\sum_{i=1}^n r_i \boldsymbol{b}_i/\sqrt{n}) = n^{-1} \sum_{i=1}^n \boldsymbol{b}_i^2 \mathrm{Var}(r_i) \leq c^2$$

Since $\mathrm{Var}(N) \leq 2\,\mathrm{Var}(N_1) + 2\,\mathrm{Var}(N_2)$, we can conclude. □

Lemma A.7. *Suppose that X_n is a sequence of random variables such that $\mathbb{E}[X_n^2] \to \infty$ while $\mathrm{Var}(X_n) = O(1)$. Then, for any $M \geq 0$, $\Pr(|X_n| \leq M) \to 0$.*

Proof. First, note that $\mathrm{Var}(|X_n|) \leq \mathrm{Var}(X_n)$ so $\mathrm{Var}(|X_n|) = O(1)$. Moreover $\mathrm{Var}(|X_n|) = \mathbb{E}[X_n^2] - (\mathbb{E}[|X_n|])^2$, so $\mathbb{E}[X_n^2] \to \infty$ and $\mathrm{Var}(|X_n|) = O(1)$ implies that $\mathbb{E}[|X_n|] \to \infty$. Then,

$$\begin{aligned}
\Pr(|X_n| \leq M) &= \Pr(|X_n| - \mathbb{E}[|X_n|] \leq M - \mathbb{E}[|X_n|]) \\
&= \Pr(\mathbb{E}[|X_n|] - |X_n| \geq \mathbb{E}[|X_n| - M) \\
&\leq \Pr(|\mathbb{E}[|X_n|] - |X_n|| \geq \mathbb{E}[|X_n|] - M) \\
&\leq \frac{\mathrm{Var}(|X_n|)}{\mathbb{E}[|X_n|] - M}
\end{aligned}$$

Since $\mathrm{Var}(|X_n|) = O(1)$ but $\mathbb{E}[|X_n|] \to \infty$, this tends to zero. □

Lemma A.8. *Suppose that X_n and Y_n are random variables such that $Y_n = O_p(1)$ and, for any $M \geq 0$, $\Pr(|X_n| \leq M) \to 0$. Then, for any $M_1 \geq 0$, $\Pr(X_n^2 - Y_n \leq M_1) \to 0$.*

Proof. Pick any $\epsilon > 0$. We want to show that, eventually, $\Pr(X_n^2 - Y_n > M_1) \geq 1 - \epsilon$. Since $Y_n = O_p(1)$, there is a fixed constant M_Y such that $\Pr(|Y_n| \leq M_Y) \geq 1 - \epsilon/2$. Since $\Pr(|X_n| \leq M) \to 0$ for any $M \geq 0$, there exists an N_X such that, for $n \geq N_X$, $\Pr(X_n^2 \leq M_1 + M_Y) \leq \epsilon/2$. A union bound completes the argument (on the eventuality $n \geq N_X$):

$$\Pr(X_n^2 - Y_n > M) \geq \Pr(X_n^2 > M_1 + M_Y, |Y_n| \leq M_Y)$$

$$= 1 - \Pr(\{X_n^2 < M_1 + M_Y\} \cup \{|Y_n| > M_Y\})$$
$$\geq 1 - \epsilon/2 - \epsilon/2 = 1 - \epsilon$$

□

A.3 PROOF OF LEMMA 3.2

Proof of Lemma 3.2. For N and D defined at the top of Appendix A.1 define $\widehat{N} = N + \Delta_N$ and $\widehat{D} = D + \Delta_D$. We can then write $JK(\beta_0) = \widehat{N}^2/\widehat{D}$ and rewrite

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

Apply Lemma F.2 to see that $N^2 = O_p(1)$ while under Assumption 3.2, $D = O_p(1)$. Thus, $2ND\Delta_n + D\Delta_n - N^2\Delta_D = o_p(1)$. Meanwhile, by Lemma A.11, $\Pr(D^2 \leq \delta_n) \to 0$ for any sequence $\delta_n \to 0$. Apply Lemma A.9 to conclude. □

Lemma A.9. *Let A_n, B_n and Y_n be sequences of random variables such that $A_n = o_p(1)$ and $B_n = o_p(1)$. If Y_n is such that for any sequence $\delta_n \to 0$, $\Pr(|Y_n| \leq \delta_n) \to 0$, then,*

$$\left|\frac{A_n}{Y_n + B_n}\right| = o_p(1)$$

Proof. Fix any $\epsilon > 0$. We show that

$$\left|\frac{A_n}{Y_n + B_n}\right| \leq \epsilon$$

on an intersection of events whose probability tends to one. By Lemma G.1 there is a sequence $\epsilon_n \searrow 0$ such that

$$\Pr(|A_n| \leq \epsilon_n) \to 1 \text{ and } \Pr(\epsilon|B_n| \leq \epsilon_n) \to 1$$

Consider the intersection of events $\Omega_1 \cap \Omega_2 \cap \Omega_3$ where

$$\Omega_1 := \{\epsilon|Y_n| \geq 2\epsilon_n\}, \quad \Omega_2 := \{\epsilon|B_n| \leq \epsilon_n\}, \quad \Omega_3 := \{|A_n| \leq \epsilon_n\}$$

By assumption, $\Pr(\Omega_1 \cap \Omega_2 \cap \Omega_3) \to 1$. On this event $|Y_n + B_n| \geq \epsilon_n/\epsilon > 0$ and $|A_n| \leq \epsilon_n$ so that $|A_n/(Y_n + B_n)| \leq |\epsilon_n/(\epsilon_n/\epsilon)| \leq \epsilon$. □

Lemma A.10 (Denominator Interpolation)**.** *Suppose that Assumptions 3.1 and 3.2 hold. Let $\varphi(\cdot) : \mathbb{R} \to \mathbb{R}$ be such that $\varphi(\cdot) \in C_b^3(\mathbb{R})$ with $L_2(\varphi) = \sup_x |\varphi''(x)|$ and $L_3(\varphi) = \sup_x |\varphi'''(x)|$. Then there is a constant M that depends only on the constant c such that:*

$$|\mathbb{E}[\varphi(D) - \varphi(\tilde{D})]| \leq \frac{M}{\sqrt{n}}(L_2(\varphi) + L_3(\varphi))$$

Proof of Lemma A.10. We inherit the definitions of D_{-i}, Δ_{2i}^a, Δ_{2i}^b, $\tilde{\Delta}_{2i}^a$, and $\tilde{\Delta}_{2i}^b$ from the proof of Lemma A.1 with $a = 1$. Then, as before we can write

$$\mathbb{E}[\varphi(D) - \varphi(\tilde{D})] = \sum_{i=1}^{n} \mathbb{E}[\varphi(D_{-i} + n^{-1}\Delta_{2i}^a + n^{-1}\Delta_{2i}^b)]$$
$$- \mathbb{E}[\varphi(D_{-i} + n^{-1}\tilde{\Delta}_{2i}^a + n^{-1}\tilde{\Delta}_{2i}^b)]$$

We examine each term via a second-order Taylor expansion around D_{-i}

$$\begin{aligned}\mathbb{E}[\text{Term}_i] = &\frac{1}{n}\mathbb{E}[\varphi'(D_{-i})\{(\Delta_{2i}^a - \tilde{\Delta}_{2i}^a) + (\Delta_{2i}^b - \tilde{\Delta}_{2i}^b)\}] \\ &+ \frac{1}{2n^2}\mathbb{E}[\varphi''(D_{-i})\{((\Delta_{2i}^a)^2 - (\tilde{\Delta}_{2i}^a)^2) + 2(\Delta_{2i}^a\Delta_{2i}^b - \tilde{\Delta}_{2i}^a\tilde{\Delta}_{2i}^b) + ((\Delta_{2i}^b)^2 - (\Delta_{2i}^b)^2)\}] \\ &+ R_i + \tilde{R}_i\end{aligned}$$

where R_i and $\tilde{R}_i$ are remainder terms to be analyzed later. Using the restrictions in (A.4) we can simplify the above display:

$$\begin{aligned}\mathbb{E}[\text{Term}_i] = &\underbrace{0.5n^{-2}\mathbb{E}[\varphi''(D_{-i})((\Delta_{2i}^a)^2 - (\tilde{\Delta}_{2i}^a)^2)]}_{\dot{\mathbf{A}}_i} + \underbrace{n^{-2}\mathbb{E}[\varphi''(K_{-i})(\Delta_{2i}^a\Delta_{2i}^b - \tilde{\Delta}_{2i}^a\tilde{\Delta}_{2i}^b)}_{\dot{\mathbf{B}}_i} \\ &+ R_i + \tilde{R}_i\end{aligned}$$

Using Lemma F.1 we can bound

$$|\mathbf{A}_i| \leq \frac{M}{n^2}L_2(\varphi) \qquad |\mathbf{B}_i| \leq \frac{M}{n^{3/2}}L_2(\varphi)$$

For some $\bar{D}_{1i}$ and $\bar{D}_{2i}$ we can express

$$\begin{aligned}R_i &= \mathbb{E}[\varphi'''(\bar{D}_{1i})\{n^{-1}\Delta_{2i}^a + \Delta_{2i}^b\}^3] \leq \frac{M}{n^{3/2}}L_3(\varphi) + \frac{M}{n^3}L_3(\varphi) \\ R_i &= \mathbb{E}[\varphi'''(\bar{D}_{2i})\{n^{-1}\tilde{\Delta}_{2i}^a + \tilde{\Delta}_{2i}^b\}^3] \leq \frac{M}{n^{3/2}}L_3(\varphi) + \frac{M}{n^3}L_3(\varphi)\end{aligned}$$

where the inequalities again come from applications of Lemma F.1. Combining these bounds and summing over the n terms gives the result. □

Lemma A.11 (Denominator anti-concentration)**.** *Suppose that Assumptions 3.1 and 3.2 hold. Then, for any sequence $\delta_n \searrow 0$,*

$$\Pr(D \leq \delta_n) \to 0$$

Proof of Lemma A.11. 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 second sequence $\gamma_n \searrow 0$ slowly enough such that $(\gamma_n^{-2} + \gamma_n^{-3})/\sqrt{n} \to 0$. Take $\varphi_n(x) = \tilde{\varphi}(\frac{x-\delta_n}{\gamma_n})$. By Lemma A.10 and since $\tilde{\varphi}(\cdot)$ has bounded derivatives up to the third order, there is a fixed constant $M_1 > 0$ that depends only on c such that

$$\Pr(D \leq \delta_n) \leq \Pr(\tilde{D} \leq \delta_n + \gamma_n) + \frac{M_1}{\sqrt{n}}(\gamma_n^{-2} + \gamma_n^{-3})$$

Let γ_n be a sequence tending to zero such that $(\gamma_n^{-2} + \gamma_n^{-3})/\sqrt{n} \to 0$ and conclude by applying Lemma A.2. □

A.4 PROOF OF LEMMA 3.3

For any $j = 1, \ldots, d_b$ define the matrix $B_j = \text{diag}(b_j(z_1), \ldots, b_j(z_n))$ and collect observations $\epsilon(\beta_0) = (\epsilon_1(\beta_0), \ldots, \epsilon_n(\beta_0))' \in \mathbb{R}^n$, $r = (r_1, \ldots, r_n)' \in \mathbb{R}^n$, $\hat{r} = (\hat{r}_1, \ldots, \hat{r}_n)' \in \mathbb{R}^n$, and $\xi = (\xi_1, \ldots, \xi_n)' \in \mathbb{R}^n$. In addition, collect $b_\epsilon = (b_{\epsilon 1}, \ldots, b_{\epsilon n}) \in \mathbb{R}^{d_b \times n}$ where $b_{\epsilon i} = \epsilon_i(\beta_0)b(z_i) \in \mathbb{R}^{d_b}$. Finally, let $\boldsymbol{H} = \frac{s_n}{\sqrt{n}}H$, $\tilde{H} = s_n H$ and $\tilde{h}_{ij} = s_n h_{ij}$.

Step 1: $\Delta_N \to_p 0$. To show that $\Delta_N \to_p 0$ write

$$\begin{aligned}\Delta_N &= |\epsilon(\beta_0)'\boldsymbol{H}(\hat{r} - r)| \\ &= |\epsilon(\beta_0)'\boldsymbol{H}(b_\epsilon'\hat{\gamma} - b_\epsilon'\gamma) - \epsilon(\beta_0)'\boldsymbol{H}\xi| \\ &\leq \underbrace{\max_{1\leq j\leq d_b} |\epsilon(\beta_0)'\boldsymbol{H}B_j\epsilon(\beta_0)|\|\hat{\gamma} - \gamma\|_1}_{\mathbf{A}} + \underbrace{\|\epsilon(\beta_0)'\boldsymbol{H}\|_2\|\xi_2\|_2}_{\mathbf{B}}\end{aligned}$$

To bound **A** we move to apply Theorem H.1 to the quadratic form $\epsilon(\beta_0)'(\boldsymbol{H}B_j)\epsilon(\beta_0)$. First notice that, under Assumption 3.4(v), we have

$$\|\mathbb{E}[\boldsymbol{H}b_j\epsilon(\beta_0)]\|_2 = \frac{1}{n}\sum_{i=1}^{n}(\mathbb{E}[s_n\sum_{j\neq i} h_{ij}b(z_j)\epsilon_j(\beta_0)])^2 \leq c^2$$

In the notation of Theorem H.1 this give us an upper bound on $\|\mathbb{E}f^{(1)}(X)\|_{\text{HS}}$. Next, Assumption 3.2 gives us that the Frobenius norm of $\boldsymbol{H} = \frac{s_n}{\sqrt{n}}\boldsymbol{H}$ is bounded, since the rows of s_nH are square summable, $\sum_{j\neq i}(s_nh_{ij})^2 \leq c$ for all $i = 1,\ldots,n$. In the notation of Theorem H.1 this gives us an upper bound on $\|\mathbb{E}f^{(2)}(X)\|_{\text{HS}}$. Applying Theorem H.1 and a union bound then gives us that

$$\max_{1\leq j\leq d_b} |\epsilon(\beta_0)'\boldsymbol{H}B_j\epsilon(\beta_0) - \mathbb{E}[\epsilon(\beta_0)'\boldsymbol{H}B_j\epsilon(\beta_0)]| = O_p(\log^{2/a}(d_b)) \tag{A.13}$$

Since $\max_{1\leq j\leq d_b}|\mathbb{E}[\epsilon(\beta_0)'\boldsymbol{H}B_j\epsilon(\beta_0)]| \leq c$ under Assumption 3.4(v), (A.13) gives that

$$\max_{1\leq j\leq d_b} |\epsilon(\beta_0)'\boldsymbol{H}B_j\epsilon(\beta_0)| = O_p(\log^{2/a}(d_b))$$

Since $\log^{2/a}(d_b)\|\widehat{\gamma} - \gamma\|_1 \to_p 0$ by assumption, this yields that $\mathbf{A} \to_p 0$.

To bound **B** see that $\|\epsilon(\beta_0)'\boldsymbol{H}\|_2 = \frac{s_n^2}{n}\sum_{i=1}^{n}(\sum_{j\neq i} h_{ij}\epsilon_i(\beta_0))^2 = O_p(1)$ under Assumption 3.3(ii) while under Assumption 3.4 $\|\xi\|_2 = o(1)$.

Step 2: $\Delta_D \to_p 0$. Notice that $a^2 - b^2 = 2b(a-b) + (a-b)^2$ and bound:

$$\begin{aligned}|\Delta_D| \leq \underbrace{\frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2(\beta_0)\Big|\sum_{j\neq i}\tilde{h}_{ij}r_j\Big|}_{\mathbf{E}} \times \max_i\Big|\sum_{j\neq i}\tilde{h}_{ij}(\hat{r}_j - r_j)\Big| \\ + \underbrace{\frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2(\beta_0)}_{\mathbf{F}} \times \max_i\Big|\sum_{j\neq i}\tilde{h}_{ij}(\hat{r}_j - r_j)\Big|^2\end{aligned}$$

Since both $\mathbf{E} = O_p(1)$ and $\mathbf{F} = O_p(1)$ under Assumptions 3.1 and 3.2, it suffices to show that

$$\max_i\Big|\sum_{j\neq i}\tilde{h}_{ij}(\widehat{r}_j - r_j)\Big| \to_p 0$$

To do so write

$$\max_i \Big|\sum_{j\neq i} \tilde{h}_{ij}\{\hat{r}_j - r_j\}\Big| \leq \underbrace{\max_{\substack{1\leq i\leq n\\ 1\leq j\leq d_b}} \Big|\sum_{j\neq i} \tilde{h}_{ij} b(z_j)\epsilon_j(\beta_0)\Big| \|\hat{\gamma}-\gamma\|_1}_{\mathbf{A}} + \underbrace{\max_{\substack{1\leq i\leq n\\ 1\leq j\leq d_b}} \Big|\sum_{j\neq i} \tilde{h}_{ij} b(z_j)\xi_j\Big|}_{\mathbf{B}}$$

To bound **A**, note that by Assumption 3.4(v) $\max_{i,j} |\mathbb{E}[\sum_{j\neq i} \tilde{h}_{ij} b(z_j)\epsilon_j(\beta_0)| \leq c$. Under Assumptions 3.2 and 3.4(ii), $\max_{i,j} \sum_{j\neq i} \tilde{h}_{ij}^2 b^2(z_j) \leq c^2$ so we can apply Theorem H.1 and a union bound to obtain that

$$\max_{\substack{1\leq i\leq n\\ 1\leq j\leq d_b}} \Big|\sum_{j\neq i} \tilde{h}_{ij} b(z_j)\epsilon_j(\beta_0)\Big| = O_p(\log^{1/a}(d_b n))$$

Along with the implied rate on $\|\hat{\gamma}-\gamma\|_1$ from Assumption 3.4(iv) this shows that $\mathbf{A} \to_p 0$.

To show that $\mathbf{B} \to 0$, use Cauchy-Schwarz, $\sum_{j\neq i} \tilde{h}_{ij}^2 b^2(z_j) \leq c$ for any i, j by Assumptions 3.2 and 3.4(ii), and $\sum_{i=1}^n \xi_i^2 = o(1)$ by Assumption 3.4(iii).

A.5 PROOF OF THEOREM 3.1

Apply Lemma A.12 with $X_n = JK(\beta_0)$, $Y_n = JK_I(\beta_0)$ and $Z_n = JK_G(\beta_0)$. The density of Z_n is uniformly bounded by Lemma F.3.

Lemma A.12. *Let X_n, Y_n, and Z_n be sequences of random variables such that $|X_n - Y_n| \to_p 0$, the distribution of Z_n is absolutely continuous with respect to Lebesgue measure and the density functions of Z_n are uniformly bounded and $\sup_{a\in\mathbb{R}} |\Pr(Y_n \leq a) - \Pr(Z_n \leq a)| \to 0$. Then $\sup_{a\in\mathbb{R}} |\Pr(X_n \leq a) - \Pr(Z_n \leq a)| \to 0$.*

Proof. For any $a \in \mathbb{R}$ and $\epsilon > 0$ we have that $\{X_n \leq a\} \subseteq \{Y_n \leq a+\epsilon\} \cup \{|X_n - Y_n| > \epsilon\}$; thus, by applying union bound and rearranging we obtain:

$$\begin{aligned}\Pr(X_n \leq a) &\leq \Pr(Y_n \leq a+\epsilon) + \Pr(|Y_n - X_n| > \epsilon)\\ &\leq \Pr(Z_n \leq a+\epsilon) + |\Pr(Y_n \leq a+\epsilon) - \Pr(Z_n \leq a+\epsilon)|\\ &\quad + \Pr(|Y_n - X_n| > \epsilon)\end{aligned}$$

so that

$$\begin{aligned}\Pr(X_n \leq a) - \Pr(Z_n \leq a) \leq \Pr(a < Z_n \leq a+\epsilon) &+ |\Pr(Y_n \leq a+\epsilon) - \Pr(Z_n \leq a+\epsilon)|\\ &+ \Pr(|Y_n - X_n| > \epsilon)\end{aligned}$$

Let $\epsilon_n \to 0$ be a sequence tending to zero such that $\Pr(|X_n - Y_n| > \epsilon_n) \to 0$ (Lemma G.1). Applying a supremum to the above display yields

$$\begin{aligned}\sup_{a\in\mathbb{R}} \Pr(X_n \leq a) - \Pr(Z_n \leq a) \leq &\sup_{a\in\mathbb{R}} \Pr(a < Z_n \leq a+\epsilon_n)\\ &+ \sup_{a\in\mathbb{R}} |\Pr(Y_n \leq a+\epsilon_n) - \Pr(Z_n \leq a+\epsilon_n)|\\ &+ \Pr(|Y_n - X_n| > \epsilon_n)\end{aligned}$$

The first term goes to zero as $\epsilon_n \to 0$ since Z_n has a uniformly bounded density; the second term goes to zero by $\sup_{a\in\mathbb{R}} |\Pr(Y_n \leq a) - \Pr(Z_n \leq a)| \to 0$ and the third term goes to zero by definition of ϵ_n and $|Y_n - X_n| \to_p 0$.

We can apply a symmetric argument to show that $\sup_{a\in\mathbb{R}} \Pr(Z_n \leq a) - \Pr(X_n \leq a) \leq o(1)$ which

completes the claim of the lemma. □

B PROOFS OF RESULTS IN SECTION 4

The statement of Theorem 4.1 relies on showing

$$\sup_{(a_1,a_2)\in\mathbb{R}^2} \left|\Pr(JK(\beta_0) \le a_1, C \le a_2) - \Pr(JK_G(\beta_0) \le a_1, C_G \le a_2)\right| \to 0$$
$$\text{and} \quad \sup_{(a_1,a_2)\in\mathbb{R}^2} \left|\Pr(S(\beta_0) \le a_1, C \le a_2) - \Pr(S_G(\beta_0) \le a_1, C_G \le a_2)\right| \to 0$$

In particular, since $(JK_G(\beta_0) \perp C_G)$ and $(S_G(\beta_0) \perp C_G)$ under H_0, showing the above will imply the test based on $T(\beta_0;\tau)$ has asymptotic size α for any choice of cutoff τ. The second line in the above display follows imediately from Theorem H.5 after verifying Assumption H.2, below.

The first line in the top display relies on a joint interpolation of the infeasible $JK_I(\beta_0)$ test statistic and the infeasible conditioning statistic C_I, which could be constructed if $\rho(z_i)$ was known to the researcher.

$$C_I := \max_{1\le i\le n} \left|\frac{1}{\sqrt{n}} \sum_{i=1}^n h_{ij} r_j \big/ (n^{-1} \sum_{i=1}^n h_{ij}^2)^{1/2}\right| \tag{B.1}$$

This joint interpolation argument is rather involved however, and deferred to Appendix D. The interpolation argument for the conditioning statistic very closely follows the results in Chernozhukov et al. (2013). The results of Section 4 rely on showing that the difference between C and C_I can be treated as negligible. This in turn reduces to verifying Assumption H.2, which is done in Lemma B.1, below.

Lemma B.1. *Suppose that Assumption 3.4 holds. Then there are sequences $\delta_n \searrow 0$, $\beta_n \searrow 0$ such that*

$$\Pr\Big(\max_{i\in[n]} n^{-1} \sum_{j=1}^n \dot{h}_{ij}^2 (\widehat{r}_j - r_j)^2 > \delta_n^2 / \log^2(n)\Big) \le \beta_n$$

where $\dot{h}_{ij} = h_{ij}/(n^{-1}\sum_{j=1}^n h_{ij}^2)^{1/2}$.

Proof. In view of Lemma G.1 it suffices to show

$$\max_{1\le i\le n} \frac{1}{n} \sum_{j=1}^n \dot{h}_{ij}^2 (\hat{r}_i - r_i)^2 = o_p(1/\log^2(n)) \tag{B.2}$$

Notice that we can bound

$$\begin{aligned}
\max_{1\le i\le n} \frac{1}{n} \sum_{j=1}^n (\hat{r}_i - r_i)^2 &= \max_{1\le i\le n} \Big|(\widehat{\gamma}-\gamma)' n^{-1} \sum_{j=1}^n \epsilon_j^2(\beta_0) b(z_i) b(z_j)' (\widehat{\gamma}-\gamma)\Big| \\
&\quad + \max_{1\le i\le n} |n^{-1} \sum_{j=1}^n \dot{h}_{ij}^2 \xi_j^2| \\
&\le \max_{\substack{1\le i\le n \\ 1\le j,k\le d_b}} \underbrace{\Big|n^{-1} \sum_{j=1}^n \epsilon_j^2(\beta_0) b_j(z_j) b_k(z_j)\Big|}_{\mathbf{A}_{ijk}} \|\widehat{\gamma}-\gamma\|_1^2
\end{aligned}$$

$$+ n^{-1/2} \max_{1\le i\le n}(n^{-1}\sum_{j=1}^{n} \dot{h}_{ij}^4)^{1/2}(\sum_{j=1}^{n} \xi_j^4)^{1/2}$$

Under Assumption 3.4(i,ii) each $\mathbf{A}_{ijk}$ is v-sub-exponential by Theorem H.1 (that is $\|\mathbf{A}_{ijk}\|_{\psi_v}$ is bounded). An application of Lemma G.2 then yields that $\max_{i,j,k}|\mathbf{A}_{ijk}| = O_p(\log^{1/v}(d_b n))$. Along with Assumption 3.4(iv) this gives that $\max_{i,j,k}|\mathbf{A}_{ijk}|\|\widehat{\gamma}-\gamma\|_1 = O_p(\log^{-3/(v\wedge 1)}(d_b n)) = o_p(\log^{-2}(n))$. Meanwhile by definition of $\dot{h}_{ij}$, $\max_i(n^{-1}\sum_{j=1}^n \dot{h}_{ij}^4)^{1/2} = O(1)$ while by Assumption 3.4(iii) $(\sum_{j=1}^n \xi_j^4)^{1/2} = o(1)$. Since $\log^2(n)/\sqrt{n} \to 0$ this shows (B.2). □

B.1 PROOF OF THEOREM 4.1

The first result in Theorem 4.1 with $JK(\beta_0)$ and C replaced with their infeasible analogs $JK_I(\beta_0)$ and C_I follows from the argument in Appendix D. After verifying that $|JK(\beta_0) - JK(\beta_0)| \to_p 0$ via Lemma 3.3 and that Assumption H.2 is satisfied via Lemma B.1 follow the same steps as in the proof of Belloni et al. (2018), Theorem 2.1 to see that approximation result holds for the feasible $JK(\beta_0)$ and C.

For the second statement, I show that the conditions of Theorem H.6 are satisfied. To see that Assumption H.1(i,ii) is satisfied under Assumption 3.1 use (i) the definition of $\dot{h}_{ij} = h_{ij}/(n^{-1}\sum_{j=1}^n h_{ij}^2)^{1/2}$; (ii) that the variance of each r_j is bounded away from zero and (iii) that the fourth moments of r_j are bounded from above. Assumption H.1(iii) is satisfied with $B_n = \log^{1/v}(n)$ by Assumption 4.1(i,iii) and Lemma G.2. Finally Assumption H.2 is satisfied by applying Lemma B.1. Apply Theorem H.6 to conclude.

C PROOFS OF RESULTS IN SECTION 5

Throughout this section, define the scaled elements of the infeasible and gaussian numerators and denominators

$$N_\ell = \frac{s_{n,\ell}}{\sqrt{n}}\sum_{i=1}^n \epsilon_i(\beta_0)\sum_{j=1}^n h_{ij}r_j \qquad \tilde{N}_\ell = \frac{s_{n,\ell}}{\sqrt{n}}\sum_{i=1}^n \tilde{\epsilon}_i(\beta_0)\sum_{j=1}^n h_{ij}\tilde{r}_j$$

$$D_{\ell k} = \frac{s_{\ell,n}s_{m,k}}{n}\sum_{i=1}^n \epsilon_i^2(\beta_0)(\sum_{j=1}^n h_{ij}r_{\ell j})(\sum_{j=1}^n h_{ij}r_{kj}) \qquad \tilde{D}_{\ell k} = \frac{s_{\ell,n}s_{m,k}}{n}\sum_{i=1}^n \epsilon_i^2(\beta_0)(\sum_{j=1}^n h_{ij}\tilde{r}_{\ell j})(\sum_{j=1}^n h_{ij}\tilde{r}_{kj})$$

Collect these in $N = (N_1,\dots N_{d_x})' \in \mathbb{R}^{d_x}$, $\tilde{N} = (\tilde{N}_1,\dots,\tilde{N}_{d_x})' \in \mathbb{R}^{d_x}$, $D = [D_{\ell k}]_{\ell,k\in[d_x]} \in \mathbb{R}^{d_x\times d_x}$, and $\tilde{D} = [\tilde{D}_{\ell k}]_{\ell,k\in[d_x]} \in \mathbb{R}^{d_x\times d_x}$. After multiplying by scaling matrix $\text{diag}(s_{1,n},\dots,s_{d_x,n})$ and the inverse of the scaling matrix we rewrite the infeasible and gaussian test statistics

$$JK_I(\beta_0) = N'D^{-1}N\mathbf{1}_{\{\lambda_{\min}(D)>0\}} \qquad JK_G(\beta_0) = \tilde{N}'\tilde{D}^{-1}\tilde{N}$$

These are the representations of the test statsitics we will largely work through in this section.

C.1 PROOF OF LEMMA 5.1

Lemma 5.1 follows immediately from the joint gaussian approximation argument established in Appendix D.

C.2 PROOF OF LEMMA 5.2

Define the matrix $\Delta_D = [(\Delta_D)_{\ell k}]_{\ell,k\in[d_x]}$ and the vector $\Delta_N = [(\Delta_N)_\ell]_{\ell\in[d_x]}$ where

$$(\Delta_D)_{\ell k} := \frac{s_{\ell,n}s_{k,n}}{n}\sum_{i=1}^{n}\epsilon_i^2(\beta_0)\big(\widehat{\Pi}_{\ell,i}\widehat{\Pi}_{k,i} - \widehat{\Pi}^I_{\ell,i}\widehat{\Pi}^I_{k,i}\big)$$

$$(\Delta_N)_\ell := \frac{s_{\ell,n}}{\sqrt{n}}\sum_{i=1}^{n}\epsilon_i(\beta_0)(\widehat{\Pi}_{\ell,i} - \widehat{\Pi}^I_{\ell,i})$$

Under the conditions of Lemma 5.2 we have that $\|\Delta_D\| \to_p 0$ and $\|\Delta_N\| \to_p 0$. Using this notation, we can write the infeasible version of the test statistic as $JK^I(\beta_0) = N'D^{-1}N$ while the feasible version is written $JK(\beta_0) = (N+\Delta_N)'(D+\Delta_D)^{-1}(N+\Delta_N)$. Add and subtract D^{-1} to get

$$\begin{aligned}
JK(\beta_0) &= \big(N+\Delta_N\big)'\big((D+\Delta_D)^{-1} \pm D^{-1}\big)\big(N+\Delta_N\big)\\
&= JK^I(\beta_0) + N'\big((D+\Delta_D)^{-1} - D^{-1}\big)N + \Delta_N\big((D+\Delta_D)^{-1} - D^{-1}\big)N\\
&\quad + \Delta_N'\big((D+\Delta_D)^{-1} - D^{-1}\big)\Delta_N + N'D^{-1}\Delta_N + \Delta_N D^{-1}N + \Delta_N D^{-1}\Delta_N
\end{aligned}$$

Via Lemma D.2 we have that $\|D^{-1}\| = (\lambda_{\min}(D))^{-1} = O_p(1)$ and by assumption we have that $\Delta_N \to_p 0$. It therefore suffices to show that

$$\|(D+\Delta_D)^{-1} - D^{-1}\| \to_p 0 \tag{C.1}$$

To do so, we can use the following equality from Horn and Johnson (2012), p. 381.

$$\|(D+\Delta_D)^{-1} - D^{-1}\| \leq \frac{\|D^{-1}\|^2\|\Delta_D\|}{1 - \|D^{-1}\Delta_D\|}$$

Since $\|D^{-1}\| = O_p(1)$ and $\Delta_D \to_p 0$, this gives (C.1).

C.3 PROOF OF THEOREM 5.1

Under Assumption 5.4, the conditions of Lemma 5.2 can be verified following the same steps as the proof of Lemma 3.3. Combine Lemma 5.2 and Lemma 5.1 as in the proof of Theorem 3.1 to conclude.

C.4 PROOF OF THEOREM 5.2

D JOINT GAUSSIAN APPROXIMATION OF $JK(\beta_0)$ AND C

The main results of Sections 4 and 5 rely on a joint interpolation of the conditioning and testing statistics as well as a joint interpolation of the conditioning and testing statistics. The joint interpolation of $JK(\beta_0)$ and the conditioning statistic C is given in Appendix D.2 after introducing some notation in Appendix D.1. The joint gaussian approximation of $S(\beta_0)$ and C follows immediately from results in Belloni et al. (2018), Chernozhukov et al. (2017). The result is presented below for the general form of the $JK(\beta_0)$ sttatistic under H_0 however the proof strategy is very similar when using the decomposed form of $JK(\beta_0)$ when $d_x = 1$. This proof is available on request.

D.1 NOTATION

JACKKNIFE STATISTIC DEFINITIONS. Define $\tilde{h}_{\ell,ij} = s_{n,\ell} h_{ij}$ for each $\ell = 1, \ldots, d_x$ and the scaled leave-one-out quasi-numerator and denominators

$$U_{-i} = \left[\frac{1}{\sqrt{n}} \sum_{j=1}^{n} \dot{\epsilon}_j(\beta_0) \sum_{k \neq i} \tilde{h}_{\ell,jk} \dot{r}_{\ell k} \right]_{1 \leq \ell \leq d_x} \in \mathbb{R}^{d_x}$$

$$D_{-i} = \left[\frac{1}{n} \sum_{j=1}^{n} \ddot{\epsilon}_i^2(\beta_0) \Big(\sum_{k \neq i} \tilde{h}_{\ell,ij} \dot{r}_{\ell j} \Big) \Big(\sum_{k \neq i} \tilde{h}_{\ell,ij} \dot{r}_{mj} \Big) \right]_{\substack{1 \leq \ell \leq d \\ 1 \leq m \leq d_x}} \in \mathbb{R}^{d_x \times d_x}$$

where $\dot{\epsilon}_j(\beta_0)$ is equal to $\tilde{\epsilon}_j(\beta_0)$ if $j < i$ and equal to $\epsilon_j(\beta_0)$ if $j > i$, $\dot{r}_{\ell j}$ is equal to $\tilde{r}_{\ell j}$ if $j < i$ and equal to r_j if $j > i$, and $\ddot{\epsilon}_j(\beta_0)$ is equal to $\mathbb{E}[\epsilon_j^2(\beta_0)]$ if $j < i$ and equal to $\epsilon_j(\beta_0)$ if $j > i$. As in the proof of Lemma 3.1 while the definitions of $\dot{\epsilon}_j(\beta_0)$, $\dot{r}_{\ell j}$, and $\ddot{\epsilon}_j(\beta_0)$ depend on i this dependence is suppressed to conslidate notation and since we only consider one step deviations at a time.

Also define the one step deviations

$$\Delta_{Ui} = \Big[\epsilon_i(\beta_0) \sum_{j=1}^{n} \tilde{h}_{\ell,ij} \dot{r}_{\ell j} + r_{\ell i} \sum_{j=1}^{n} \tilde{h}_{\ell,ji} \dot{\epsilon}_j(\beta_0) \Big]_{1 \leq \ell \leq d} \in \mathbb{R}^d$$

$$\tilde{\Delta}_{Ui} = \Big[\tilde{\epsilon}_i(\beta_0) \sum_{j=1}^{n} \tilde{h}_{\ell,ij} \dot{r}_{\ell j} + \tilde{r}_{\ell i} \sum_{j=1}^{n} \tilde{h}_{\ell,ji} \dot{\epsilon}_j(\beta_0) \Big]_{1 \leq \ell \leq d} \in \mathbb{R}^d$$

$$\Delta_{Di} = \underbrace{\big[(\Delta_{Di}^a)_{\ell m} \big]_{\substack{1 \leq \ell \leq d \\ 1 \leq m \leq d}}}_{\Delta_{Di}^a} + \underbrace{\big[(\Delta_{Di}^b)_{\ell m} \big]_{\substack{1 \leq \ell \leq d \\ 1 \leq m \leq d}}}_{\Delta_{Di}^b}$$

$$\tilde{\Delta}_{Di} = \underbrace{\big[(\tilde{\Delta}_{Di}^a)_{\ell m} \big]_{\substack{1 \leq \ell \leq d \\ 1 \leq m \leq d}}}_{\tilde{\Delta}_{Di}^a} + \underbrace{\big[(\tilde{\Delta}_{Di}^b)_{\ell m} \big]_{\substack{1 \leq \ell \leq d \\ 1 \leq m \leq d}}}_{\tilde{\Delta}_{Di}^b}$$

where

$$(\Delta_{Di}^a)_{\ell m} = \epsilon_i^2(\beta_0) \Big(\sum_{j=1}^{n} \tilde{h}_{\ell,ij} r_{\ell j} \Big) \Big(\sum_{j=1}^{n} \tilde{h}_{\ell,ij} \dot{r}_{\ell j} \Big) \Big(\sum_{j=1}^{n} h_{m,ij} r_{m,ij} \Big)^2 + r_{\ell i} r_{ki} \sum_{j=1}^{n} \tilde{h}_{\ell,ij} \tilde{h}_{m,ij} \ddot{\epsilon}_j^2(\beta_0)$$

$$(\tilde{\Delta}_{Di}^a)_{\ell m} = \tilde{\epsilon}_i^2(\beta_0) \Big(\sum_{j=1}^{n} \tilde{h}_{\ell,ij} r_{\ell j} \Big) \Big(\sum_{j=1}^{n} \tilde{h}_{\ell,ij} \dot{r}_{\ell j} \Big) \Big(\sum_{j=1}^{n} h_{m,ij} r_{m,ij} \Big)^2 + \tilde{r}_{\ell i} \tilde{r}_{ki} \sum_{j=1}^{n} \tilde{h}_{\ell,ij} \tilde{h}_{m,ij} \ddot{\epsilon}_j^2(\beta_0)$$

$$(\Delta_{Di}^b)_{\ell m} = r_{\ell i} \sum_{j=1}^{n} \ddot{\epsilon}_j^2(\beta_0) \sum_{k \neq i} \tilde{h}_{\ell,ji} \tilde{h}_{m,jk} \dot{r}_{mk} + r_{ki} \sum_{j=1}^{n} \ddot{\epsilon}_j^2(\beta_0) \sum_{k \neq i} \tilde{h}_{\ell,ji} \tilde{h}_{m,jk} \dot{r}_{\ell k}$$

$$(\tilde{\Delta}_{Di}^b)_{\ell m} = \tilde{r}_{\ell i} \sum_{j=1}^{n} \ddot{\epsilon}_j^2(\beta_0) \sum_{k \neq i} \tilde{h}_{\ell,ji} \tilde{h}_{m,jk} \dot{r}_{mk} + \tilde{r}_{ki} \sum_{j=1}^{n} \ddot{\epsilon}_j^2(\beta_0) \sum_{k \neq i} \tilde{h}_{\ell,ji} \tilde{h}_{m,jk} \dot{r}_{\ell k}$$

Notice that in this notation we can write the test statistic and gaussian test statistics, after scaling by $\mathrm{diag}(s_{n,1}, \ldots, s_{n,d_x})$, as

$$C(\beta_0) = (U_{-1} + \Delta_{U1}/\sqrt{n})'(D_{-1} + \Delta_{D1}/n)^{-1}(U_{-1} + \Delta_{U1}/\sqrt{n}) \mathbf{1}\{\lambda_{\min}(D_{-1} + \Delta_{D1})^{-1}) > 0\}$$

$$\tilde{C}(\beta_0) = (U_{-n} + \tilde{\Delta}_{Un}/\sqrt{n})'(\tilde{D}_{-1} + \tilde{\Delta}_{D1}/n)^{-1}(U_{-n} + \tilde{\Delta}_{U1}/\sqrt{n})$$

In this proof we will use these representations for the test statistics. Finally define

$$U = U_{-1} + \Delta_{U1}/\sqrt{n} \qquad \tilde{U} = U_{-n} + \tilde{\Delta}_{Un}/\sqrt{n}$$
$$D = D_{-1} + \Delta_{D1}/n \qquad \tilde{D} = D_{-n} + \Delta_{Dn}/n$$

CONDITIONING STATISTIC DEFINITIONS. Let $h_{\ell,ii} = 0$ for any $\ell = 1, \ldots, d_x$ and $i = 1, \ldots, n$. Define $\tilde{h}_{\ell,ij} = h_{\ell,ij}/\omega_{\ell i}$ for $\omega_{\ell i} = n^{-1}\sum_{j=1}^{n} |h_{\ell,ij}|$. Also define the one-step deviations:

$$\Delta_{Ci} := (\tilde{h}_{1,ji} r_{1i}, -\tilde{h}_{1,ji} r_{1i}, \ldots, \tilde{h}_{d_x,ji} r_{d_x i}, -\tilde{h}_{d_x,ji} r_{d_x i})'_{1\le j\le n} \in \mathbb{R}^{2nd_x}$$
$$\Delta_{Ci} := (\tilde{h}_{1,ji} \tilde{r}_{1i}, -\tilde{h}_{1,ji} \tilde{r}_{1i}, \ldots, \tilde{h}_{d_x,ji} \tilde{r}_{d_x i}, -\tilde{h}_{d_x,ji} \tilde{r}_{d_x i})'_{1\le j\le n} \in \mathbb{R}^{2nd_x}$$

And the leave-one-out vector

$$C_{-i} := \frac{1}{\sqrt{n}}\sum_{j<i} \tilde{\Delta}_{Cj} + \frac{1}{\sqrt{n}}\sum_{j>i} \Delta_{Cj} \in \mathbb{R}^{2nd_x}$$

Notice that $C = \max_{1\le\iota\le 2nd_x}(C_{-1} + \frac{1}{\sqrt{n}}\Delta_{C1})_\iota$ while $\tilde{C} = \max_{1\le\iota\le 2nd_x}(C_{-n} + \Delta_{Cn})_\iota$.

FUNCTION DEFINITIONS. As in Chernozhukov et al. (2013) consider the "smooth max" function, $F_\beta : \mathbb{R}^p \to \mathbb{R}$ defined

$$F_\beta(z) = \beta^{-1}\log\left(\sum_{i=1}^{n}\exp(\beta z_i)\right)$$

which satisfies

$$0 \le F_\beta(z) - \max_{1\le i\le n} z_i \le \beta^{-1}\log p.$$

Appendix F.2 notes some useful properties of the smooth max function which we will use in the joint interpolation argument. In addition let $\varphi(\cdot) \in C_b^3(\mathbb{R})$ be such that $\varphi(x) = 1$ if $x \le 0$, $\varphi'(x) < 0$ for $x \in (0,1)$, and $\varphi(x) = 0$ for $x \ge 1$. For any $\gamma > 0$ and $a = (a_1, a_2)' \in \mathbb{R}^2$ define the function $\tilde{\varphi}(\cdot,\cdot,\cdot) : \mathbb{R}^{d_x} \times \mathrm{vec}(\mathbb{R}^{d_x \times d_x}) \times \mathbb{R}^{2nd_x} \to \mathbb{R}$ via

$$\tilde{\varphi}_{\gamma,a}(u, \mathrm{vec}(d), c) := \phi_{\gamma,a_1}(u, \mathrm{vec}(d))\tau_{\gamma,a_2}(c) \tag{D.1}$$

where

$$\phi_{\gamma,a_1}(u, \mathrm{vec}(d)) := \varphi\left(\frac{u'd^{-1}u - a_1}{\gamma\lambda_{\min}^5(d)}\right)$$
$$\tau_{\gamma,a}(c) := \varphi\left(\frac{F_{1/\gamma}(c) - a_2}{\gamma}\right)$$

The function $\tilde{\varphi}_{\gamma,a}(\cdot,\cdot,\cdot)$ is meant to approximate the indicator function $\mathbf{1}\{K(\beta_0) \le a_1\}\mathbf{1}\{C \le a_2\}$ with γ governing the quality of approximation. Where it is obvious, we will supress the subscripts γ, a from our notation.

D.2 MAIN ARGUMENT

Lemma D.1 (Joint Lindeberg Interpolation). *Suppose that Assumptions 5.1–5.3 hold. Then there is a fixed constant M*

$$\left|\mathbb{E}[\tilde{\varphi}_{\gamma,a}(U, vec(D), C) - \tilde{\varphi}_{\gamma,a}(\tilde{U}, vec(\tilde{D}), \tilde{C})]\right| \leq \frac{M_1 \log^{M_2}(n)}{\sqrt{n}}(\gamma^{-1} + \gamma^{-2} + \gamma^{-3}) \tag{D.2}$$

Proof of Lemma D.1. We can bound the difference on the left hand side of (D.2) using the telescoping sum

$$\begin{aligned}\sum_{i=1}^{n} \Big|&\mathbb{E}[\tilde{\varphi}_{\gamma,a}(U_{-i} + \Delta_{Ui}/\sqrt{n}, \text{vec}(D_{-i} + \Delta_{Di}/n), C_{-i} + \Delta_{Ci}/\sqrt{n})] \\ &- \mathbb{E}[\tilde{\varphi}_{\gamma,a}(U_{-i} + \Delta_{Ui}/\sqrt{n}, \text{vec}(D_{-i} + \Delta_{Di}/n), C_{-i} + \Delta_{Ci}/\sqrt{n})]\Big|\end{aligned} \tag{D.3}$$

By second degree Taylor expansion, we break each of the summands in (D.3) into first order, second order, and remainder terms; each of which are bounded below. We make use of the following moment conditions implied by (i) indpendence of observations across $i = 1, \ldots, n$ and (ii) the mean and covariance matrix of $(\epsilon_i(\beta_0), r_i)$ being equal to the mean and covariance matrix of $(\tilde{\epsilon}_i(\beta_0), r_i)$

$$\begin{aligned}0 &= \mathbb{E}[\Delta_{Ui} - \tilde{\Delta}_{Ui}|\mathcal{F}_{-i}] = \mathbb{E}[\Delta_{Ui}\Delta_{Ui}' - \tilde{\Delta}_{Ui}\tilde{\Delta}_{Ui}'|\mathcal{F}_{-i}] = \mathbb{E}[\text{vec}(\Delta_{Di}) - \text{vec}(\tilde{\Delta}_{Di})|\mathcal{F}_{-i}] \\ &= \mathbb{E}[\Delta_{Ci} - \tilde{\Delta}_{Ci}|\mathcal{F}_{-i}] = \mathbb{E}[\Delta_{Ui} \otimes \text{vec}(\Delta_{Di}^b)' - \tilde{\Delta}_{Ui} \otimes \text{vec}(\tilde{\Delta}_{Di}^b)'|\mathcal{F}_{-i}] \\ &= \mathbb{E}[\Delta_{Ci} \otimes \Delta_{Ui} - \tilde{\Delta}_{Ci} \otimes \tilde{\Delta}_{Ui}|\mathcal{F}_{-i}] = \mathbb{E}[\Delta_{Ci} \otimes \text{vec}(\tilde{\Delta}_{Di}^b) - \tilde{\Delta}_{Ci} \otimes \text{vec}(\tilde{\Delta}_{Di}^b)|\mathcal{F}_{-i}] \\ &= \mathbb{E}[\text{vec}(\Delta_{Di}^b)\text{vec}(\Delta_{Di}^b)' - \text{vec}(\tilde{\Delta}_{Di}^b)\text{vec}(\tilde{\Delta}_{Di}^b)'|\mathcal{F}_{-i}]\end{aligned} \tag{D.4}$$

where $\mathcal{F}_{-i}$ denotes the sub-sigma algebra generated by all observations not equal to i, $\otimes$ denotes the Kronecker product, and I apologize for the abuse of the equal sign in the above display.

First Order Terms. First order terms can be expressed

$$\begin{aligned}\text{First Order}_i = &\sum_{\ell=1}^{d_x} \mathbb{E}\left[\frac{\partial}{\partial U_\ell}\tilde{\varphi}(U_{-i}, \text{vec}(D_{-i}), C_{-i})((\Delta_{Ui})_\ell - (\tilde{\Delta}_{Ui})_\ell)\right]/\sqrt{n} \\ &+ \sum_{\ell=1}^{d_x}\sum_{m=1}^{d_x} \mathbb{E}\left[\frac{\partial}{\partial D_{\ell m}}\tilde{\varphi}(U_{-i}, \text{vec}(D_{-i}), C_{-i})((\Delta_{Di})_{\ell m} - (\tilde{\Delta}_{Di})_{\ell m})\right]/n \\ &+ \sum_{\ell=1}^{2nd_x} \mathbb{E}\left[\frac{\partial}{\partial C_\ell}\tilde{\varphi}(U_{-i}, \text{vec}(D_{-i}), C_{-i})((\Delta_{Ci})_\ell - (\tilde{\Delta}_{Ci})_\ell)\right]/\sqrt{n}\end{aligned}$$

These terms are all equal to zero after applying the matched moments in (D.4).

Second Order Terms. After canceling out terms using the matched moments in (D.4) the second order terms that remain can be expressed

$$\begin{aligned}\text{2nd Order}_i &= \frac{1}{n^{3/2}}\sum_{\ell=1}^{d_x}\sum_{m=1}^{d_x}\sum_{n=1}^{d_x} \underbrace{\mathbb{E}\left[\frac{\partial^2}{\partial U_\ell \partial D_{mn}}\tilde{\varphi}(U_{-i}, \text{vec}(D_{-i}), C_{-i})((\Delta_{Ui})_\ell(\Delta_{Di}^a)_{mn} - (\tilde{\Delta}_{Ui})_\ell(\tilde{\Delta}_{Di}^a)_{mn})\right]}_{\mathbf{A}_{\ell mn}} \\ &= \frac{1}{n^2}\sum_{\ell=1}^{d_x}\sum_{m=1}^{d_x}\sum_{n=1}^{d_x}\sum_{o=1}^{d_x} \underbrace{\mathbb{E}\left[\frac{\partial^2}{\partial U_\ell \partial D_{mn}}\tilde{\varphi}(U_{-i}, \text{vec}(D_{-i}), C_{-i})((\Delta_{Di}^a)_{\ell m}(\Delta_{Di}^a)_{no} - (\tilde{\Delta}_{Di}^a)_{\ell m}(\tilde{\Delta}_{Di}^a)_{no})\right]}_{\mathbf{B}_{\ell mno}}\end{aligned}$$

$$= \frac{2}{n^2} \sum_{\ell=1}^{d_x} \sum_{m=1}^{d_x} \sum_{n=1}^{d_x} \sum_{o=1}^{d_x} \underbrace{\mathbb{E}\left[\frac{\partial^2}{\partial U_\ell \partial D_{mn}} \tilde{\varphi}(U_{-i}, \text{vec}(D_{-i}), C_{-i})((\Delta^b_{Di})_{\ell m}(\Delta^a_{Di})_{no} - (\tilde{\Delta}^a_{Di})_{\ell m}(\tilde{\Delta}^b_{Di})_{no}) \right]}_{\mathbf{C}_{\ell mno}}$$

$$= \frac{1}{n^{3/2}} \sum_{\ell=1}^{2nd_x} \sum_{m=1}^{d_x} \sum_{n=1}^{d_x} \underbrace{\mathbb{E}\left[\frac{\partial^2}{\partial C_\ell \partial D_{mn}} \tilde{\varphi}(U_{-i}, \text{vec}(D_{-i}), C_{-i})((\Delta_{Ci})_\ell(\Delta^a_{Di})_{mn} - (\tilde{\Delta}_{Ci})_\ell(\tilde{\Delta}^a_{Di})_{mn}) \right]}_{\mathbf{D}_{\ell mn}}$$

To bound each $\mathbf{A}_{\ell mn}$, $\mathbf{B}_{\ell mno}$, and $\mathbf{C}_{\ell mno}$ we use the fact that the second order derivatives of $\tilde{\varphi}$ are bounded up to a log power of n via repeated application of Lemmas F.12 and F.15. Under Assumption 5.1 the absolute value of terms $(\Delta_{Ui})_\ell$,$|\Delta^a_{Di}|_{mn}$, and $(\Delta^b_{Di}/\sqrt{n})_{no}$ can also be shown to have bounded third moments via the exact same steps as in the proof of Lemma F.1. Putting these together with generalized Holder's inequality will yield a finite constants M_1 and M_2 such that $|\mathbf{A}_{lmn}| \leq M_1 \log^{M_2}(n)(\gamma^{-1} + \gamma^{-2})$, $\mathbf{B}_{\ell mno} \leq M_1 \log^{M_2}(n)(\gamma^{-1} + \gamma^{-2})$, and $|\mathbf{C}_{\ell mno}| \leq M_1 \log^{M_2}(n) n^{1/2}(\gamma^{-1} + \gamma^{-2})$. To bound $\mathbf{D}_{\ell mn}$ terms notice that

$$\sum_{\ell=1}^{2nd_x} \mathbf{D}_{\ell mn} = \sum_{\ell=1}^{2nd_x} \mathbb{E}\left[\frac{\partial}{\partial D_{mn}} \phi(U_{-i}, \text{vec}(D_{-i})) \frac{\partial}{\partial C_\ell} \tau(C_{-i})((\Delta_{C-i})_\ell(\Delta^a_{Di})_{mn} - (\tilde{\Delta}_{Ci})_\ell(\tilde{\Delta}^a_{Di})_{mn})) \right]$$

Apply Lemma F.1 to bound Δ^a_{Di}, and Lemmas F.12 and F.15 to bound the derivative of $\phi(\cdot)$ and Cauchy-Schwarz to split up the Δ_{Ci} and Δ_{Di} terms

$$\leq \sqrt{M_1 \log^{M_2}(n)\gamma^{-2}} \mathbb{E}\left[\sum_{\ell=1}^{2nd_x} (\partial_\ell \tau(C_{-i}))^2((\Delta_{Ci})_\ell + (\tilde{\Delta}_{Ci})_\ell)^2 \right]^{1/2}$$

$$\leq \sqrt{M_1 \log^{M_2}(n)\gamma^{-2}} \mathbb{E}\left[\max_{1 \leq \ell \leq n} ((\Delta_{Ci})_{2\ell} + (\tilde{\Delta}_{Ci})_{2\ell})^2 \sum_{\ell=1}^{2nd_x} (\partial_\ell \tau(C_{-i}))^2 \right]^{1/2}$$

By Lemma F.8 and chain rule we have that $\sum_{\ell=1}^{2nd_x} (\partial_\ell \tau(C_{-i}))^2 \leq \gamma^{-2}$. Moreover $(\Delta_{Ci})_\ell^{a/2}$ is sub-exponential so via Lemma G.2 the second moment of the maximum is bounded by a power of $\log(n)$. After updating the constant M_1 and M_2 this yields

$$\leq M_1 \log^{M_2}(n)\gamma^{-2}$$

Putting these all together and summing over the remaining indices gives

$$|\text{Second Order}_i| \leq \frac{M_1 \log^{M_2}(n)}{n^{3/2}}(\gamma^{-1} + \gamma^{-2}) \tag{D.5}$$

Remainder Terms. The first remainder term can be expressed

$$\begin{aligned}
\text{Remainder}_i = &\frac{1}{n^{3/2}} \sum_{\ell=1}^{d_x} \sum_{m=1}^{d_x} \sum_{n=1}^{d_x} \mathbb{E}\left[\frac{\partial^3}{\partial U_\ell \partial U_m \partial U_n} \tilde{\varphi}(\bar{U}, \text{vec}(\bar{D}), \bar{C})(\Delta_{Ui})_\ell(\Delta_{Ui})_m(\Delta_{Ui})_n \right] \\
&+ \frac{1}{n^3} \sum_{(\ell,m)} \sum_{(n,o)} \sum_{(q,p)} \mathbb{E}\left[\frac{\partial^3}{\partial D_{\ell m} \partial D_{no} \partial D_{pq}} \tilde{\varphi}(\bar{U}, \text{vec}(\bar{D}), \bar{C})(\Delta_{Di})_{\ell m}(\Delta_{Di})_{no}(\Delta_{Di})_{qp} \right] \\
&+ \frac{1}{n^{3/2}} \sum_{\ell=1}^{2nd_x} \sum_{m=1}^{2nd_x} \sum_{n=1}^{2nd_x} \mathbb{E}\left[\frac{\partial^3}{\partial C_\ell \partial C_m \partial C_n} \tilde{\varphi}(\bar{U}, \text{vec}(\bar{D}), \bar{C})(\Delta_{Ci})_\ell(\Delta_{Ci})_m(\Delta_{Ci})_n \right] \\
&+ \frac{1}{n^2} \sum_{\ell=1}^{d_x} \sum_{m=1}^{d_x} \sum_{(n,o)} \mathbb{E}\left[\frac{\partial^3}{\partial U_\ell \partial U_m \partial D_{no}} \tilde{\varphi}(\bar{U}, \text{vec}(\bar{D}), \bar{C})(\Delta_{Ui})_\ell(\Delta_{Ui})_m(\Delta_{Di})_{no} \right]
\end{aligned}$$

$$+\frac{1}{n^{5/2}}\sum_{\ell=1}^{d_x}\sum_{(m,n)}\sum_{(o,p)}\mathbb{E}\left[\frac{\partial^3}{\partial U_\ell \partial D_{mn}\partial D_{op}}\tilde{\varphi}(\bar{U},\mathrm{vec}(\bar{D}),\bar{C})(\Delta_{Ui})_\ell(\Delta_{Di})_{mn}(\Delta_{Di})_{op}\right]$$

$$+\frac{1}{n^{5/2}}\sum_{\ell=1}^{2nd_x}\sum_{(m,n)}\sum_{(o,p)}\mathbb{E}\left[\frac{\partial^3}{\partial C_\ell \partial D_{mn}\partial D_{op}}\tilde{\varphi}(\bar{U},\mathrm{vec}(\bar{D}),\bar{C})(\Delta_{Ci})_\ell(\Delta_{Di})_{mn}(\Delta_{Di})_{op}\right]$$

$$+\frac{1}{n^{2}}\sum_{\ell=1}^{2nd_x}\sum_{m=1}^{2nd_x}\sum_{(n,o)}\mathbb{E}\left[\frac{\partial^3}{\partial C_\ell \partial C_m\partial D_{no}}\tilde{\varphi}(\bar{U},\mathrm{vec}(\bar{D}),\bar{C})(\Delta_{Ci})_\ell(\Delta_{Ci})_m(\Delta_{Di})_{no}\right]$$

$$+\frac{1}{n^{3/2}}\sum_{\ell=1}^{2nd_x}\sum_{m=1}^{2nd_x}\sum_{n=1}^{d_x}\mathbb{E}\left[\frac{\partial^3}{\partial C_\ell \partial C_m\partial U_n}\tilde{\varphi}(\bar{U},\mathrm{vec}(\bar{D}),\bar{C})(\Delta_{Ci})_\ell(\Delta_{Ci})_m(\Delta_{Ui})_n\right]$$

$$+\frac{1}{n^{2}}\sum_{\ell=1}^{2nd_x}\sum_{m=1}^{2nd_x}\sum_{n=1}^{d_x}\mathbb{E}\left[\frac{\partial^3}{\partial C_\ell \partial C_m\partial U_n}\tilde{\varphi}(\bar{U},\mathrm{vec}(\bar{D}),\bar{C})(\Delta_{Ci})_\ell(\Delta_{Ci})_m(\Delta_{Ui})_n\right]$$

where $\bar{U},\mathrm{vec}(\bar{D})$, and $\bar{C}$ vary term by term but are always in the hyper-rectangles $[U_{-i}, U+\Delta_{Ui}]$, $[\mathrm{vec}(D_{-i}),\mathrm{vec}(D_{-i}+\Delta_{Di})]$, and $[C_{-i}, C_{-i}+\Delta_{Ci}]$, respectively. As such, any moment conditions that apply to U, D, C also apply to $(\bar{U},\bar{D},\bar{C})$. Repeated application of generalized Hölder inequality, Lemma F.1 to bound moments of Δ_{Ui} and $(\Delta_{Di}/\sqrt{n})$, Lemma F.15 to bound moments of the second and third derivatives of $\phi(\tilde{U},\mathrm{vec}(\tilde{D}))$, Lemma F.11 to bound the sums of derivatives of $\tau(\tilde{C})$, and Lemma G.2 to bound moments of $\max_{1\le\ell\le n}(\Delta_{Ci})_\ell$ will yield that

$$|\text{Remainder}_i| \le \frac{M_1\log^{M_2}(n)}{n^{3/2}}(\gamma^{-1}+\gamma^{-2}+\gamma^{-3}) \tag{D.6}$$

Symmetric logic will bound the other remainder term. Summing (D.5) and (D.6) over indices gives the result. □

Lemma D.2 (Denominator Anticoncentration)**.** *Suppose that Assumptions 5.1–5.3 hold. Then for any sequence $\delta_n \to 0$ we have that $\Pr(\lambda_{\min}(\tilde{D}) \le \delta_n) \to 0$.*

Proof. By Lemma D.4 it suffices to show that for any fixed $a \in \mathcal{S}^{d_x-1}$ and any $\delta_n \to 0$, $\Pr(a'Da \le \delta_n) \to 0$. For any such a write:

$$a'\tilde{D}a = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\epsilon_i^2(\beta_0)]\Big(\sum_{\ell=1}^{d_x}\sum_{j=1}^{n}a_\ell\tilde{h}_{\ell,ij}r_{\ell,j}\Big)^2$$

$$\ge \frac{1}{cn}\sum_{i=1}^{n}\Big(\sum_{\ell=1}^{d_x}\sum_{j=1}^{n}a_\ell\tilde{h}_{\ell,ij}r_{\ell,j}\Big)^2$$

Define $\dot{s}_{n,j} = \max_{\{\ell:a_\ell\neq 0\}} s_{n,\ell}$ and $\dot{h}_{ij} = s_n h_{ij}$

$$= \frac{1}{cn}\sum_{i=1}^{n}\Big(\sum_{j=1}^{n}\dot{h}_{ij}\sum_{\ell=1}^{d_x}\frac{a_\ell s_{n,\ell}}{s_n}r_{\ell,j}\Big)^2$$

By Assumption 5.1 we have that $\lambda_{\min}(\mathbb{E}[D]) \ge \underline{c}$ so that $\mathbb{E}[\frac{1}{n}\sum_{i=1}^{n}\big(\sum_{\ell=1}^{d_x}\sum_{j=1}^{n}a_\ell\tilde{h}_{\ell,ij}r_{\ell,j}\big)^2] \ge c^{-1}$. Moreover, by Assumption 5.1, $\mathrm{Var}(\sum_{\ell=1}^{d_x}\frac{a_\ell s_{n,\ell}}{s_n})$ is bounded from above and below. Define the matrix $\tilde{H} = [\dot{h}_{ij}]_{ij}$ and follow the same steps as Lemma D.2 to conclude. □

Lemma D.3 (Gaussian Approximation). *Suppose that Assumptions 5.1–5.3 hold. Then*

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

Proof. Let $a = (a_1, a_2)$ and $\tilde{\phi}_{\gamma,a}$ be as in (D.1):

$$\begin{aligned}
\Pr(N'D^{-1}N \leq a_1, C \leq a_2) &\leq \mathbb{E}[\tilde{\phi}_{\gamma,a}(U, \text{vec}(D), C)] \\
&\leq \mathbb{E}[\tilde{\phi}_{\gamma,a}(\tilde{U}, \text{vec}(\tilde{D}), \tilde{C})] + \frac{M_1 \log_2^M(n)}{\sqrt{n}}(\gamma^{-1} + \gamma^{-2}) \\
&\leq \Pr(\tilde{N}'\tilde{D}^{-1}\tilde{N} \leq a_1, \tilde{C} \leq a_2) + \Pr(a_1 \leq \tilde{N}'\tilde{D}^{-1}N \leq a_1 + \gamma\lambda_{\min}^5(D)) \\
&\quad + \Pr(a_2 \leq C \leq a_2 + \gamma) + \frac{M_1 \log_2^{M_2}(n)}{\sqrt{n}}(\gamma^{-1} + \gamma^{-2} + \gamma^{-3}) \\
&\leq \Pr(\tilde{N}'\tilde{D}^{-1}\tilde{N} \leq a_1, \tilde{C} \leq a_2) + \Pr(a_1 \leq \tilde{N}'\tilde{D}^{-1}N \leq a_1 + \gamma\lambda_{\min}^5(D)) \\
&\quad + \Pr(a_2 \leq C \leq a_2 + \gamma) + \frac{M_1 \log_2^{M_2}(n)}{\sqrt{n}}(\gamma^{-1} + \gamma^{-2} + \gamma^{-3})
\end{aligned}$$

Let $\gamma \to 0$ at a rate such that $\frac{\log^{M_2}(n)}{\sqrt{n}}\gamma^{-3} \to 0$ and apply Lemmas D.1 and D.2 to conclude as in the proof of Lemma A.3. A symmetric argument shows that the lower bound tends to zero.

□

Lemma D.4. *Let $\Sigma_n \in \mathbb{R}^{d\times d}$ be a sequence of random positive-semidefinite matrices. Suppose that for any fixed $a \in \mathcal{S}^{d-1}$ and any $\delta_n \to 0$ we have that $\Pr(a'\Sigma_n a \leq \delta_n) \to 0$ and $\Pr(\lambda_{\max}^2(\Sigma_n) \geq \delta_n^{-1}) \to 0$. Then for any $\delta_n \to 0$, $\Pr(\lambda_{\min}^2(\Sigma_n) \leq \delta_n) \to 0$.*

Proof. Take any preliminary sequence $\delta_n \to 0$. It suffices to show that there is another sequence $\tilde{\delta}_n$ weakly larger than $\delta_n/2$ such that $\Pr(\lambda_{\min}^2(\Sigma_n) \leq \tilde{\delta}_n) \to 0$. For any $m \in \mathbb{N}$ let $\mathcal{A}_m$ be a set of points in $\mathcal{S}^{d-1}$ such that

$$\max_{a\in\mathcal{S}^{d-1}} \min_{\tilde{a}\in\mathcal{A}_m} \|a - \tilde{a}\| \leq \delta_m^2$$

From here let $\tilde{n}_j$ be defined

$$\tilde{n}_j = \inf\{n \geq j : \min_{\tilde{a}\in\mathcal{A}_{n,j}} \Pr(\tilde{a}'\Sigma_n a \leq 2\delta_{n_j}) < \delta_{n_j}\}$$

Define a new sequence $\tilde{\delta}_n \to 0$, weakly larger than δ_n, via

$$\tilde{\delta}_n = \begin{cases} 1 & \text{if } 0 \leq 0 \leq n < \tilde{n}_1 \\ \delta_i & \text{if } \tilde{n}_i \leq n < \tilde{n}_{i+1} \end{cases}$$

and notice that, by definition $\Pr(\min_{a\in\mathcal{A}_{\tilde{n}_j}} a'\Sigma_n a \leq 2\tilde{\delta}_n) < \delta_{\tilde{n}_j}$. We wish to show that $\lambda_{\min}^2(\Sigma_n) > \tilde{\delta}_n$ on an intersection of events whose probability tends to one. Since Σ_n is positive semi-definite, $\|x\|_{\Sigma_n}^2 = x'\Sigma_n x$ defines a seminorm. By triangle inequality

$$\lambda_{\min}^2(\Sigma_{n_j}) \geq \min_{\mathcal{A}_{n_j}} a'\Sigma_{n_j}a - \lambda_{\max}^2(\Sigma_n)\tilde{\delta}_{n_j}^2$$

Define the events

$$\Omega_1 = \{\min_{\mathcal{A}_{\tilde{n}_j}} a'\Sigma_n a \geq 2\tilde{\delta}_n\} \text{ and } \Omega_2 = \{\lambda_{\max}(\Sigma_n) \leq \tilde{\delta}_n^{-1/2}\}$$

On the intersection of these events, whose probabilities tend to one, we have $\lambda^2_{\min}(\Sigma_n) \geq \tilde{\delta}_n$. □

E Incorporating Exogenous Controls

In this section, I analyze the model with exogeneous controls. To this end, define the vector $z_2 = (z_{21}', \ldots, z_{2n}')' \in \mathbb{R}^{n \times d_c}$. Let $P_2 = z_2(z_2'z_2)^{-1}z_2' \in \mathbb{R}^{n \times n}$ denote the projection onto the column space of z_2 and $M_2 = I_n - P_2$ denote the projection onto to orthocomplement of the column space. Focus will be on the case where $d_x = 1$ to simplify notation, but the basic concepts apply generally to $d_x > 1$.

For $y \coloneqq (y_1, \ldots, y_n)' \in \mathbb{R}^n$ and $x \coloneqq (x_1', \ldots, x_n')' \in \mathbb{R}^{n\times}$ define $y^\perp \coloneqq M_2 y$ and $x^\perp \coloneqq M_2 x$ as the "partialled out" versions of y and x, respectively. Let $y_i^\perp$ be the i^{th} element of $y^\perp$ and $x_i^\perp$ be the i^{th} element of $x^\perp$. From here we can define $\epsilon(\beta_0) \coloneqq y - x\beta_0$, $\epsilon^\perp(\beta_0) = M_2\epsilon(\beta_0)$ and $r^\perp \coloneqq M_2 r$ where as in the main text $r = (r_1, \ldots, r_n)'$ is constructed $r_i = x_i - \rho(z_i)\epsilon_i(\beta_0)$. The definition of $\rho(z_i)$ does not change after partialling out z_2 since all expectations are understood to be conditional on the instruments z. Notice that $\epsilon^\perp(\beta_0)$ is mean zero. Finally I assume that the controls have been partialled out of hat matrix so that the effective hat matrix is M_2H and the vector $\widehat{\Pi} \in \mathbb{R}^n$ is defined $\widehat{\Pi} = (M_2H)(M_2r)$. This does not make a difference for the numerator of the $JK(\beta_0)$ statistic but does affect the denominator slightly. When this is not done, inference may be conservative.

Using matrix notation in the numerator to make things clear, we can write the version of the $JK(\beta_0)$ statistic with the partialled out vectors, $\epsilon^\perp(\beta_0)$ and $r^\perp$, in terms of the original vectors, $\epsilon(\beta_0)$ and r,

$$\begin{aligned} JK_I(\beta_0) &= \frac{\left(\frac{1}{\sqrt{n}}\epsilon(\beta_0)'M_2\tilde{H}M_2r\right)^2}{\frac{1}{n}\sum_{i=1}^n(\epsilon_i^\perp(\beta_0))^2\left(\sum_{j=1}^n \boldsymbol{h}_{ij}r_j\right)^2} \\ &= \frac{\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i(\beta_0)\sum_{j=1}^n \boldsymbol{h}_{ij}r_j\right)^2}{\frac{1}{n}\sum_{i=1}^n(\epsilon_i^\perp(\beta_0))^2\left(\sum_{j=1}^n \boldsymbol{h}_{ij}r_j\right)^2} \end{aligned}$$

where $\boldsymbol{h}_{ij} = [M_2\tilde{H}M_2]_{ij}$, $\tilde{H} = s_nH$, and $m_{ij} = [M_2]_{ij}$. I seek to characterize the limiting distribution of $JK(\beta_0)$ under H_0. To do so, we show that quantiles $JK(\beta_0)$ can be approximated by quantiles of the gaussian analog statistic

$$JK_G(\beta_0) = \frac{\left(\frac{1}{\sqrt{n}}\tilde{\epsilon}(\beta_0)'M_2\tilde{H}M_2\tilde{r}\right)^2}{\frac{1}{n}\sum_{i=1}^n \text{Var}(\epsilon_i)\left(\sum_{j=1}^n \boldsymbol{h}_{ij}\tilde{r}_j\right)^2}$$

where $(\tilde{\epsilon}_i, \tilde{\epsilon}_i(\beta_0), \tilde{r}_i)$ are generated gaussian independent of the data and with the same mean and covariance as $(\epsilon_i, \epsilon_i(\beta_0), r_i)$. Since $\text{Var}(\tilde{\epsilon}(\beta_0)) = \text{Var}(\epsilon_i)$ under H_0, $\mathbb{E}[\tilde{\epsilon}(\beta_0)'M_2] = 0$, and $\tilde{r} \perp \tilde{\epsilon}(\beta_0)$, this gaussian analog statistic has a χ_1^2 distribution conditional on any realization of $\tilde{r}$ and thus its unconditional distribution is also χ_1^2.

Showing that quantiles of $JK(\beta_0)$ can be approximated by quantiles of $\tilde{JK}(\beta_0)$ proceeds in two steps. In the first step, we show that $JK(\beta_0)$ converges in probability to an intermediate statistic.

$$JK^{\text{int}}(\beta_0) = \frac{\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i(\beta_0)\sum_{j=1}^n \boldsymbol{h}_{ij}r_j\right)^2}{\frac{1}{n}\sum_{i=1}^n \epsilon_i^2(\sum_{j\neq i} \boldsymbol{h}_{ij}r_j)^2}$$

We will then show that quantiles of this intermediate statistic can be approximated by quantiles

of $\tilde{JK}(\beta_0)$. In view of Lemma 3.2, it suffices to show for the first step that $\Delta_D \to_p 0$, where

$$\Delta_D = \frac{1}{n}\sum_{i=1}^{n}((\epsilon_i^{\perp}(\beta_0))^2 - \epsilon_i^2)\widehat{\Pi}_i^2$$

To do this, notice that under H_0 we can write $\epsilon_i^{\perp}(\beta_0) = \epsilon_i + z_{2i}'(\widehat{\Gamma} - \Gamma)$ where $\widehat{\Gamma} = (z_2'z_2)^{-1}z_2\epsilon(\beta_0)$ is a $\sqrt{n}$-consistent estimate of Γ. Exploiting this fact we get

$$\Delta_D = (\widehat{\Gamma} - \Gamma)'\frac{1}{n}\sum_{i=1}^{n}(\widehat{\Pi}_i)^2 z_{2i}z_{2i}'(\widehat{\Gamma} - \Gamma) + 2(\widehat{\Gamma} - \Gamma)'\frac{1}{n}\sum_{i=1}^{n}\epsilon_i z_{2i}\widehat{\Pi}_i$$

Both of these terms will tend to zero by the consistency $\widehat{\Gamma}$ to Γ, giving that $\Delta_D \to_p 0$.

In our second step, we argue that quantiles of $JK^{\text{int}}(\beta_0)$ can be approximated by quantiles of $JK_G(\beta_0)$. To make this comparasion, we can follow almost exactly the same steps as in Appendix A. The only difference between analysis in this case and analysis in the original case is that the partialling out of controls leads the test statistic to not strictly have a jackknife form; the effective hat matrix M_2HM_2 no longer has a deleted diagonal. However, as I will argue below, this will not make a difference in the interpolation argument since the diagonal terms of $[P_2]_{ii}$ are small in the sense that they sum to d_c.

The (A.2) analog one step deviations for the numerator are given

$$\Delta_{1i} = \epsilon_i(\beta_0)\sum_{j\neq i}\boldsymbol{h}_{ij}\dot{r}_j + r_i\sum_{j\neq i}\boldsymbol{h}_{ji}\dot{\epsilon}_j(\beta_0) + \boldsymbol{h}_{ii}\epsilon_i(\beta_0)r_i$$
$$\tilde{\Delta}_{1i} = \tilde{\epsilon}_i(\beta_0)\sum_{j\neq i}\boldsymbol{h}_{ij}\dot{r}_j + \tilde{r}_j\sum_{j\neq i}\boldsymbol{h}_{ji}\dot{\epsilon}_j(\beta_0) + \boldsymbol{h}_{ii}\tilde{\epsilon}_i(\beta_0)\tilde{r}_i$$

where as Appendix A, a dotted variable is equal to the gaussian analog if $j > i$ but equal to the standard version otherwise. The first and second moments of the first two terms in Δ_{1i} can be matched with their gaussian analog terms as in the proof of Lemma A.1. While we cannot match seconds moments of the third term in the one step deviation, this sum of all these third terms can be treated as negligible after scaling by $1/\sqrt{n}$ as $\sum_{i=1}^{n}|\boldsymbol{h}_{ii}| \lesssim d_c$. This is because $M_2\tilde{H}M_2 = \tilde{H} - P_2\tilde{H} - \tilde{H}P_2 - P_2\tilde{H}P_2$. The matrix $\tilde{H}$ has zeros on it's diagonal. Meanwhile

$$|[P_2\tilde{H}]_{ii}|^2 = \Big|\sum_{j=1}^{n}[P_2]_{ij}\tilde{H}_{ji}\Big|^2 \leq \Big(\sum_{j=1}^{n}[P_2]_{ij}^2\Big)\Big(\sum_{j\neq i}H_{ji}^2\Big) \lesssim [P_2]_{ii}$$

where the final inequality comes because the matrix P_2 is symmetric and idempotent and since $\left(\sum_{j\neq i}H_{ji}^2\right) \lesssim 1$ by Assumption 3.2(ii). A similar argument can be used to show that $[P_2\tilde{H}P_2]_{ii}^2 \lesssim [P_2]_{ii}$. Since P_2 is a projection matrix we must have that $\|P_2He_j\| \leq \|He_j\|$ for any basis vector $e_j \in \mathbb{R}^n$. Thus $\sum_{j=1}^{n}[P_2H]_{ji}^2 \leq \sum_{j=1}^{n}[H]_{ji}^2$. Finally, we can use the fact that the trace of P_2 is equal to its rank to show that $\sum_{i=1}^{n}|\boldsymbol{h}_{ii}| \lesssim d_c$

The one step deviations in the denominator can be bounded using the same logic. These one step deviations are given

$$\Delta_{2i} = \epsilon_i^2(\sum_{j\neq i}\boldsymbol{h}_{ij}\dot{r}_j)^2 + r_i^2\sum_{j\neq i}\boldsymbol{h}_{ji}^2\ddot{\epsilon}_j^2 + r_i\sum_{j\neq i}\ddot{\epsilon}_j(\sum_{k\neq j,i}\boldsymbol{h}_{ji}\boldsymbol{h}_{jk}r_k)$$

$$+ \epsilon_i^2 (\boldsymbol{h}_{ii}^2 r_i^2 + 2\boldsymbol{h}_{ii} r_j \sum_{j \neq i} \boldsymbol{h}_{ij} r_j)^2$$

$$\tilde{\Delta}_{2i} = \tilde{\epsilon}_i^2 (\sum_{j \neq i} \boldsymbol{h}_{ij} \dot{r}_j)^2 + \tilde{r}_i^2 \sum_{j \neq i} \boldsymbol{h}_{ji}^2 \ddot{\epsilon}_j^2 + \tilde{r}_i \sum_{j \neq i} \ddot{\epsilon}_j (\sum_{k \neq j,i} \boldsymbol{h}_{ji} \boldsymbol{h}_{jk} r_k)$$

$$+ \epsilon_i^2 (\boldsymbol{h}_{ii}^2 r_i^2 + 2\boldsymbol{h}_{ii} r_j \sum_{j \neq i} \boldsymbol{h}_{ij} r_j)^2$$

where $\ddot{\epsilon}_j$ is equal to $\text{Var}(\epsilon_j)$ if $j < i$ and equal to ϵ_j if $j > i$. The first three terms in this expansion are can be dealt with exactly as in the proof of Lemma A.1. The fourth term is new, however summing over the fourth terms and scaling by $1/n$ will be negligible as $\sum_{i=1}^n |\boldsymbol{h}_{ii}| \lesssim d_c$. After showing the lindeberg interpolation step, the rest of the proof follows exactly as in Appendix A.

F RELEVANT MOMENT BOUNDS

F.1 MOMENT BOUNDS FOR SECTION 3

Here I provide some lemmas that are useful in the proof of Lemmas A.1–A.3

Lemma F.1. *Let $\Delta_{1i}, \tilde{\Delta}_{1i}, \Delta_{2i}^a, \tilde{\Delta}_{2i}^a, \Delta_{2i}^b, \tilde{\Delta}_{2i}^b$ be as in (A.2). Then under Assumptions 3.1 and 3.2 there is a constant $M > 0$ such that for any $k = 1, \dots, 6$:*

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

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

$$\mathbb{E}[|\Delta_{2i}^a|^k] \leq M\alpha^k \qquad \mathbb{E}[|\tilde{\Delta}_{2i}^k|] \leq M\alpha^k$$
$$\mathbb{E}[|\Delta_{2i}^b/\sqrt{n}|^k] \leq M\alpha^k \qquad \mathbb{E}[|\tilde{\Delta}_{2i}^b/\sqrt{n}|^k] \leq M\alpha^k$$

Proof. First, since

$$\sum_{j=1}^n h_{ij}^2 \mathbb{E}[(r_j - \mathbb{E}[r_j])^2 \leq \mathbb{E}[(\sum_{i=1}^n \tilde{h}_{ij} r_j)^2] \leq 1$$

the constants are bounded, $\sum_{i=1}^n \tilde{h}_{ij}^2 \leq c$. Applying Lemma F.4 with $X_i = h_{ij} r_j$ and $X_i = h_{ij} \epsilon_j(\beta_0)$ we see that there is a constant A such that for any $k = 1, \dots, 6$

$$\mathbb{E}\Big[\Big|\sum_{i=1}^n \tilde{h}_{ij} r_j\Big|^k\Big] \leq A \text{ and } \mathbb{E}\Big[\Big|\sum_{i=1}^n \tilde{h}_{ij} \epsilon_j(\beta_0)\Big|^k\Big] \leq A \tag{F.1}$$

The bounds on $\mathbb{E}[|\Delta_{1i}^k|]$ and $\mathbb{E}[|\tilde{\Delta}_{1i}^k|]$ immediately follow from this result and the bounds on moments of r_i and $\epsilon_i(\beta_0)$ in Assumption 3.1. The bounds on $\mathbb{E}[|\Delta_{2i}^a|^k]$ and $\mathbb{E}[|\tilde{\Delta}_{2i}^a|^k]$ also follow from (F.1) after noting that there is a finite constant B such that:

$$\mathbb{E}[(\sum_{i=1}^n \tilde{h}_{ij}^2 \epsilon_i^2(\beta_0))^k] \leq B$$

Finally to bound $\mathbb{E}[|\Delta_{2i}^b/\sqrt{n}|^k]$ and $\mathbb{E}[|\tilde{\Delta}_{2i}^b/\sqrt{n}|^k]$ apply Lemma F.6 with $v_j = \epsilon_j^2(\beta_0) \sum_{k \neq i,j} \tilde{h}_{jk} r_k$, noting that $\mathbb{E}[|v_j|^3]$ is bounded by (F.1). □

Lemma F.2. *Let N and N_{-i} be defined as in Appendix A.1. Under Assumptions 3.1–3.3 there is a fixed*

constant M such that for all $i = 1, \ldots, n$ and any $k = 1, \ldots, 6$,

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

Proof. We show the bound for $\mathbb{E}[|N|^k]$ and note that the bound for N_{-i} follows from symmetric logic. Write $\epsilon_i(\beta_0) = \eta_i + \gamma_i$ where $\gamma_i = \Pi_i(\beta - \beta_0)$ and η_i is mean zero. Decompose $N = N_1 + N_2 + N_3$:

$$N_1 = \frac{1}{\sqrt{n}} \sum_{i=1}^n \eta_i \sum_{j=1}^n \tilde{h}_{ij} \dot{r}_j, \; N_2 = \frac{1}{\sqrt{n}} \sum_{i=1}^n r_i \sum_{j=1}^n \tilde{h}_{ji} \gamma_j, \text{ and } N_3 = \frac{1}{\sqrt{n}} \sum_{i=1}^n \eta_i \sum_{j=1}^n \tilde{h}_{ij} \mathbb{E}[r_j]$$

where $\dot{r}_j = r_j - \mathbb{E}[r_j]$.

Since via Assumption 3.2, $\sum_{i=1}^n h_{ji}^2 \leq c$ and via Assumption 3.1, $|\gamma_j| \leq c$, we can bound,

$$(\sum_{j=1}^n h_{ji}\gamma_j/\sqrt{n})^4 \leq (\frac{c}{\sqrt{n}} \sum_{i=1}^n |h_{ji}|)^4 \leq c^8 \implies (\sum_{j=1}^n h_{ji}\gamma_j/\sqrt{n})^6 \leq c^8 (\sum_{j=1}^n h_{ji}\gamma_j/\sqrt{n})^2$$

Under Assumption 3.3, $\mathbb{E}[N_2^2] \leq c$ while Assumption 3.2 implies that $(\sum_{i=1}^n h_{ij}\mathbb{E}[r_j])^2 \leq c$ so that $\mathbb{E}[N_3^2] \leq c^2$.

An absolute bound on the higher moments of N_2 then follows from an application of Lemma F.4 with $X_i = r_i \sum_{j=1}^n h_{ji}\gamma_j/\sqrt{n}$. An absolute bound on the higher moments of N_3 follows from symmetric logic.

To bound higher moments of N_1 define $v_i = \sum_{j<i}\{\eta_i h_{ij} r_j + \dot{r}_i h_{ji} \eta_j\}$ and write $N_1 = \frac{1}{\sqrt{n}} \sum_{i=2}^n v_i$. The sequence $v_2, \ldots, v_n$ is a martingale difference array. Via the same procedure as the bounds on $\mathbb{E}[|\Delta_{1i}|^k]$ as in Lemma F.1 one can verify that there is a fixed constant M such that $\mathbb{E}[|v_i|^k] \leq M$ for all $k = 1, \ldots, 6$. The bound on the higher moments of N then follows from Lemma F.7.

The bounds for moments of N_{-i} follow symmetric logic. □

Lemma F.3. *Let $\tilde{N}$ and $\tilde{D}$ be defined as in Appendix A.1. 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 $\tilde{r}$.*

Proof. Recall that

$$\tilde{N} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \tilde{\epsilon}_i(\beta_0) \sum_{j=1}^n \tilde{h}_{ij} \tilde{r}_j \text{ and } \tilde{D}^{1/2} = \sqrt{\frac{1}{n} \sum_{i=1}^n \kappa_i^2(\beta_0)(\sum_{j=1}^n \tilde{h}_{ij}\tilde{r}_j)^2}$$

The distribution of $\tilde{\epsilon}_i(\beta_0)|\tilde{r}_i$ is

$$\tilde{\epsilon}_i(\beta_0)|\tilde{r} \sim N\left(\mu_i(r_i), (1 - \rho_i^2)\operatorname{Var}(\epsilon_i(\beta_0))\right)$$

where $\mu_i(r_i) = \Pi_i(\beta - \beta_0) + \frac{\operatorname{Cov}(\epsilon_i(\beta_0), r_i)}{\operatorname{Var}(r_i)}(r_i - \mathbb{E}[r_i])$ and $\rho_i = \operatorname{corr}(\epsilon_i(\beta_0), r_i)$. Define $\bar{\Pi}_i := \sum_{j=1}^n \tilde{h}_{ij}\tilde{r}_j$. Then, conditional on $\tilde{r}$,

$$\frac{\tilde{N}}{\tilde{D}^{1/2}} \sim N\left(\frac{\frac{1}{\sqrt{n}}\sum_{i=1}^n \mu_i(r_i)\bar{\Pi}_i}{\sqrt{\frac{1}{n}\sum_{i=1}^n \kappa_i^2(\beta_0)\bar{\Pi}_i^2}}, \frac{\frac{1}{n}\sum_{i=1}^n (1-\rho_i^2)\operatorname{Var}(\epsilon_i(\beta_0))\bar{\Pi}_i^2}{\frac{1}{n}\sum_{i=1}^n \kappa_i^2(\beta_0)\bar{\Pi}_i^2}\right) \tag{F.2}$$

The maximum of the normal density is proportional to the inverse of the standard deviation so it suffices to show that the variance in (F.2) is bounded away from zero. To this end, notice that under Assumptions 3.1 and 3.3

$$(1-\delta^2)c^{-2} \leq (1-\rho_i^2)\frac{\text{Var}(\epsilon_i(\beta_0))}{\kappa_i^2(\beta_0)} \leq c^2$$

By Lemma G.8 to this gives that the conditional variance is also larger than $(1-\delta^2)c^{-2} > 0$.

□

Lemma F.4. *Let $X_1, \ldots, X_n$ be random variables such that $\mathbb{E}[X_i] = \mu_i$ and $\mathbb{E}[(\sum_{i=1}^n X_i)^2] \leq C$. Suppose that for any $i = 1, \ldots, n$ there is a constant U such that*

$$\mathbb{E}[(X_i-\mu_i)^3] \leq U\mathbb{E}[(X_i-\mu_i)^2] \text{ and } \mathbb{E}[(X_i-\mu_i)^6]^{1/3} \leq U\mathbb{E}[(X_i-\mu_i)^2]$$

Then $\mathbb{E}[(\sum_{i=1}^n X_i)^6] \leq 64U^3C^3 + 32C^3$.

Proof. First write

$$\mathbb{E}[(\sum_{i=1}^n X_i)^2] = \sum_{i=1}^n \mathbb{E}(X_i-\mu_i)^2 + (\sum_{i=1}^n \mu_i)^2 \leq C$$

To bound $\mathbb{E}[(\sum_{i=1}^n X_i)^6]$ expand out

$$\begin{aligned}
\mathbb{E}[(\sum_{i=1}^n X_i)^6] &= \mathbb{E}[(\sum_{i=1}^n (X_i-\mu_i) + \sum_{i=1}^n \mu_i)^6] \\
&\lesssim \mathbb{E}[(\sum_{i=1}^n (X_i-\mu_i))^6] + (\sum_{i=1}^n \mu_i)^6 \\
&= \sum_{i=1}^n \mathbb{E}[(X_i-\mu_i)^6] + \sum_{i=1}^n \sum_{j=1}^n \mathbb{E}[(X_i-\mu_i)^3(X_j-\mu_j)^3] \\
&\quad + \sum_{i=1}^n \sum_{j=1}^n \mathbb{E}[(X_i-\mu_i)^4(X_j-\mu_j)^2] \\
&+ \sum_{i=1}^n \sum_{j=1}^n \sum_{k\neq i,j} \mathbb{E}[(X_i-\mu_i)^2(X_j-\mu_i)^2(X_k-\mu_k)^2] + (\sum_{i=1}^n \mu_i)^6 \\
&\leq \sum_{i=1}^n \mathbb{E}[(X_i-\mu_i)^6] + \sum_{i=1}^n \sum_{j=1}^n \mathbb{E}[(X_i-\mu_i)^3]\mathbb{E}[(X_j-\mu_j)^3] \\
&+ \sum_{i=1}^n \sum_{j=1}^n \mathbb{E}[(X_i-\mu_i)^6]^{4/6}\mathbb{E}[(X_j-\mu_j)^6]^{2/6} \\
&+ \sum_{i=1}^n \sum_{j=1}^n \sum_{k\neq i,j} \mathbb{E}[(X_i-\mu_i)^6]^{1/3}\mathbb{E}[(X_j-\mu_i)^6]^{1/3}\mathbb{E}[(X_k-\mu_k)^6]^{1/3} \\
&+ C^3 \\
&= \left(\sum_{i=1}^n (\mathbb{E}[(X_i-\mu_i)^6])^{1/3}\right)^3 + \sum_{i=1}^n \sum_{j=1}^n \mathbb{E}[(X_i-\mu_i)^3]\mathbb{E}[(X_j-\mu_j)^3] + C^3
\end{aligned}$$

$$\begin{aligned}
&\leq \left(\sum_{i=1}^{n}(\mathbb{E}[(X_i-\mu_i)^6])^{1/3}\right)^3 + \left(\sum_{i=1}^{n}\mathbb{E}[(X_i-\mu_i)^3]\right)^2 + C^3 \\
&\leq 2U^3\left(\sum_{i=1}^{n}\mathbb{E}[(X_i-\mu_i)^2]\right)^3 + C^3 \\
&\leq 2U^3C^3 + C^3
\end{aligned}$$

where the implied constant in the second line is 32 by an application of Lemma G.8, the third line comes from expanding out the power, the first inequality by application of Hölder's inequality, and the penultimate inequality comes from applying bounds on the third and sixth central moments in terms of the second moments. □

Lemma F.5. *Let $h = (h_1, \ldots, h_n) \in \mathbb{R}^n$ be such that $\sum_{i=1}^n h_i^2 \leq b$. Suppose that $X_1, \ldots, X_n$ are such that $\mathbb{E}[|X_i|^k] \leq M$ for all $k = 1, 2, 3$. Then*

$$\mathbb{E}\Big[\Big|\sum_{i=1}^{n} h_i^2 X_i\Big|^3\Big] \leq b^3M^3$$

Proof. We can expand out

$$\begin{aligned}
\mathbb{E}\Big[\Big|\sum_{i=1}^{n} h_i^2 X_i\Big|^3\Big] &\leq \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} h_i^2h_j^2h_k^2\mathbb{E}[|X_i||X_j||X_k|] \\
&\leq M^3\sum_{i=1}^{n}h_i^2\sum_{j=1}^{n}h_j^2\sum_{k=1}^{n}h_k^2 \\
&\leq M^3\Big(\sum_{i=1}^{n}h_i^2\Big)^3 \leq c^3M^3
\end{aligned}$$

□

Lemma F.6. *Let $v_1, \ldots, v_n$ be random variables such that $\mathbb{E}[|v_i|^3] \leq M$ for all $i = 1, \ldots, n$. Let $h = (h_1, \ldots, h_n) \in \mathbb{R}^n$ be a vector of weights such that $\|h\|_2 \leq c$. Then*

$$\mathbb{E}\Big[\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} h_iv_i\Big|^3\Big] \leq c^3M$$

Proof. We can expand out

$$\begin{aligned}
\mathbb{E}\Big[\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} h_iv_i\Big|^3\Big] &\leq \frac{1}{n^{3/2}}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}|h_i||h_j||h_k|\mathbb{E}[|v_i||v_j||v_k|] \\
&\leq \frac{M}{n^{3/2}}\sum_{i=1}^{n}|h_i|\sum_{j=1}^{n}|h_j|\sum_{k=1}^{n}|h_k| \leq \frac{M}{n^{3/2}}\|h\|_1^3 \leq Mc^3
\end{aligned}$$

where the second inequality follows from generalized Hölder's inequality,

$$|\mathbb{E}[fgh]| \leq (\mathbb{E}[|f|^3]\mathbb{E}[|g|^3]\mathbb{E}[|h|^3])^{1/3}$$

and the fourth inequality from $\|h\|_1 \leq \sqrt{n}\|h\|_2$. □

Lemma F.7. *Let $v_1, \dots, v_n$ be a martingale difference array such that $\mathbb{E}[|v_i|^l] \le M$ for all $l = 1, \dots, k$. Then there is a fixed constant C_k that only depends on k such that*

$$\mathbb{E}[(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} v_i)^k] \le C_k M$$

Proof. We move to apply Theorem H.3 with $X_t = \sum_{i=1}^{t} v_i/\sqrt{n}$.

$$\begin{aligned}\mathbb{E}[(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} v_i)^k] &\le \mathbb{E}[(\max_{s \le n} \sum_{t=1}^{s} X_s)^k] \\ &\le C_k \mathbb{E}\big[\big(\sum_{i=1}^{n} v_i^2/n\big)^{k/2}\big] \le C_k \mathbb{E}\big[\frac{1}{n}\sum_{i=1}^{n} v_i^k\big] \le C_k M\end{aligned}$$

where the second inequality comes from Theorem H.3 and the third comes from an application of Jensen's inequality to the sample mean. □

F.2 Useful Properties of Smooth Max

Lemma F.8 (Chernozhukov et al. (2013), Lemma A.2)**.** *For every $1 \le j, k, l \le p$,*

$$\partial_j F_\beta(z) = \pi_j(z), \qquad \partial_j \partial_k F_\beta(z) = \beta w_{jk}(z), \qquad \partial_j \partial_k \partial_l F_\beta(z) = \beta^2 q_{jkl}(z)$$

where for $\delta_{jk} := \mathbf{1}\{j = k\}$,

$$\begin{aligned}\pi_j(z) &:= e^{\beta z_j} \Big/ \sum_{i=1}^{n} e^{\beta z_i}, \qquad w_{jk} := (\pi_j \delta_{jk} - \pi_j \pi_k)(z) \\ q_{jkl}(z) &:= (\pi_j \delta_{jl} \delta_{jk} - \pi_j \pi_l \delta_{jk} - \pi_j \pi_k (\delta_{jl} + \delta_{kl}) + 2\pi_j \pi_k \pi_l)(z)\end{aligned}$$

Moreover,

$$\pi_j(z) \ge 0, \qquad \sum_{j=1}^{p} \pi_i(z) = 1, \qquad \sum_{j,k=1}^{p} |w_{jk}(z)| \le 2, \qquad \sum_{j,k,l=1}^{p} |q_{jkl}| \le 6$$

Lemma F.9 (Chernozhukov et al. (2013), Lemma A.3)**.** *For every $x, z \in \mathbb{R}^p$,*

$$|F_\beta(x) - F_\beta(z)| \le \max_{1 \le j \le p} |x_j - z_j|.$$

Lemma F.10 (Chernozhukov et al. (2013), Lemma A.4)**.** *Let $\varphi(\cdot) : \mathbb{R} \to \mathbb{R}$ be such that $\varphi \in C_b^3(\mathbb{R})$ and define $m : \mathbb{R}^p \to \mathbb{R}$, $z \mapsto \varphi(F_\beta(z))$. The derivatives (up to the third order) of m are given*

$$\begin{aligned}\partial_j m(z) &= (\partial g(F(\beta))\pi_j)(z) \\ \partial_j \partial_k m(z) &= (\partial^2 g(F_\beta)\pi_j \pi_k + \partial g(F_\beta)\beta w_{jk})(z) \\ \partial_j \partial_k \partial_l m(z) &= (\partial^3 g(F_\beta)\pi_j \pi_k \pi_l + \partial^2 g(F_\beta)\beta(w_{jk}\pi_l + w_{jl}\pi_k + w_{kl}\pi_j) + \partial g(F_\beta)\beta^2 q_{jkl})(z)\end{aligned}$$

where π_j, w_{jk}, q_{jkl} are as described in Lemma F.8.

Lemma F.11 (Chernozhukov et al. (2013), Lemma A.5)**.** *Define $L_1(\varphi) = \sup_x |\varphi'(x)|, L_2(\varphi) =$*

$\sup_x |\varphi''(x)|$, *and* $L_3(\varphi) = \sup_x |\varphi'''(x)|$. *For every* $1 \le j, k, l \le p$,

$$|\partial_j \partial_k m(z)| \le U_{jk}(z) \;\; \textit{and} \;\; |\partial_j \partial_k \partial_l m(z)| \le U_{jkl}(z)$$

where for $W_{jk}(z) := (\pi_j \delta_{jk} + \pi_j \pi_k)(z)$,

$$\begin{aligned} U_{jk}(z) &:= (L_2 \pi_j \pi_k + L_1 \beta W_{jk}(z) \\ U_{jkl}(z) &:= (L_3 \pi_j \pi_k \pi_l + L_2 \beta (W_{jk}\pi_l + W_{jl}\pi_k + W_{kl}\pi_j) + L_1 \beta^2 Q_{jkl})(z) \\ Q_{jkl}(z) &:= (\pi_j \delta_{jl} \delta_{jk} + \pi_j \pi_k \delta_{jk} + \pi_j \pi_k (\delta_{jl} + \delta_{kl}) + 2\pi_j \pi_k \pi_l)(z). \end{aligned}$$

Moreover,

$$\sum_{j,k=1}^{p} U_{jk}(z) \le (L_2 + 2L_1\beta) \;\; \textit{and} \;\; \sum_{j,k,l=1}^{p} U_{jkl}(z) \le (L_3 + 6L_2\beta + 6L_1\beta^2).$$

F.3 MOMENT BOUNDS FOR SECTIONS 4 AND 5

Lemma F.12. *Suppose that Assumption 5.1 holds and let N and D be as defined at the top of Appendix D.2 Then under H_0, for any k there is a fixed constant C_k such that for any $\ell = 1, \dots, d_x$*

$$\mathbb{E}[|N_\ell|^k] \le C_k \;\; \textit{and} \;\; \mathbb{E}[|D_{\ell\ell}|^k] \le C_k \log^{2k/a}(n)$$

Proof. Let $\eta_{\ell i} = r_i - \mathbb{E}[r_i]$ and write

$$N_\ell = \underbrace{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \epsilon_i(\beta_0) \sum_{j=1}^{n} \tilde{h}_{ij} \eta_{\ell j}}_{N_\ell^1} + \underbrace{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \epsilon_i(\beta_0) \sum_{j=1}^{n} \tilde{h}_{ij} \mathbb{E}[r_{\ell j}]}_{N_\ell^2}$$

To bound moments of N_ℓ^1 use the fact that N_ℓ^1 is a quadratic form in mean-zero a-sub-exponential variables. By Theorem H.1, N_ℓ^1 is therefore also a-sub-exponential with parameter $a/2$; thus $(N_\ell^1)^{a/2}$ is sub-exponential and Lemma G.2 provides the moment bound for arbitrary moments. To bound moments of N_ℓ^2 we use the fact that $\max_i \left| \sum_{j=1}^n \tilde{h}_{ij} \mathbb{E}[r_{\ell j}] \right|$ is bounded by assumption and apply Burkholder-Davis-Gundy (Theorem H.3) after adding and subtracting $\mathbb{E}[\epsilon_i(\beta_0)]$.

To bound moments of $D_{\ell\ell}$ we decompose

$$|D| \le \frac{1}{n} \sum_{i=1}^{n} \epsilon_i^2(\beta_0) \max_{1 \le i \le n} \left| \sum_{j=1}^{n} h_{ij} r_j \right|^2$$

Apply Theorem H.1 to see that $\sum_{j=1}^n h_{ij} r_j$ is α-sub-exponential and Lemma G.2 to bound the RHS by a log-power of n. □

F.4 MATRIX DERIVATIVE LEMMAS

The purpose of this section is largely to establish some matrix derivative expressions that will be useful for the Lindeberg interpolation in

Lemma F.13. *Let $D \in \mathbb{R}^{d \times d}$ be a symmetric, real matrix such that* $\det(D) \neq 0$. *Let $N \in \mathbb{R}^d$ be a vector. The derivatives up to the derivatives of quadratic form $N' D^{-1} N$ are given.*

First Order:

$$\frac{\partial}{\partial N_l} = 2\sum_{j=1}^{d}(D^{-1})_{jl}N_j, \quad \frac{\partial}{\partial D_{lm}} = -2\sum_{j=1}^{d}\sum_{k=1}^{d}(D^{-1})_{jl}(D^{-1})_{km}N_jN_k,$$

Second Order:

$$\frac{\partial^2}{\partial N_l N_m} = 2(D^{-1})_{lm}, \quad \frac{\partial^2}{\partial N_l \partial D_{pq}} = -2\sum_{j=1}^{d}(D^{-1})_{jp}(D^{-1})_{ql}N_j,$$

$$\frac{\partial^2}{\partial D_{lm}\partial D_{qj}} = \sum_{j=1}^{d}\sum_{k=1}^{d}\left\{(D^{-1})_{lp}(D^{-1})_{qj})(D^{-1})_{km} + (D^{-1})_{kp}(D^{-1})_{mq}(D^{-1})_{lj}\right\}N_jN_k$$

Third Order:

$$\frac{\partial^3}{\partial N_l \partial N_m \partial N_p} = 0, \quad \frac{\partial^3}{\partial N_l \partial N_m \partial D_{pq}} = -2(D^{-1})_{lp}(D^{-1})_{qm}$$

$$\frac{\partial^3}{\partial D_{lm}\partial D_{pq}\partial N_r} = 2\sum_{j=1}^{d}\left\{(D^{-1})_{lp}(D^{-1})_{qj}(D^{-1})_{rm} + (D^{-1})_{rp}(D^{-1})_{mq}(D^{-1})lj\right\}N_j$$

$$\begin{aligned}\frac{\partial^3}{\partial D_{lm}D_{pq}D_{rs}} = 2\sum_{j=1}^{d}\sum_{j=1}^{d}\Big\{&(D^{-1})_{lr}(D^{-1})_{ps}(D^{-1})_{qj}(D^{-1})_{km} + (D^{-1})_{lp}(D^{-1})_{qr}(D^{-1})_{js}(D^{-1})_{km}\\ &+ (D^{-1})_{lp}(D^{-1})_{qj}(D^{-1})_{kr}(D^{-1})_{ms} + (D^{-1})_{kr}(D^{-1})_{ps}(D^{-1})_{mq}(D^{-1})_{lj}\\ &+ (D^{-1})_{kp}(D^{-1})_{mr}(D^{-1})_{qs}(D^{-1})_{lj} + (D^{-1})_{rp}(D^{-1})_{mq}(D^{-1})_{lr}(D^{-1})js\Big\}N_jN_k\end{aligned}$$

Proof. The derivative of an element of the the inverse of a matrix $\boldsymbol{X}$ can be expressed (Petersen and Pedersen, 2012)

$$\frac{\partial(\boldsymbol{X}^{-1})_{kl}}{\partial \boldsymbol{X}_{ij}} = -(\boldsymbol{X}^{-1})_{ki}(\boldsymbol{X}^{-1})_{jl} \tag{F.3}$$

repeated application of this identity as well as the expression of the quadratic form

$$N'D^{-1}N = \sum_{j=1}^{d}\sum_{k=1}^{d}(D^{-1})_{jk}N_jN_k$$

leads to the result, bearing in mind that the inverse of a symmetric matrix is symmetric. □

Lemma F.14. *Let D be a symmetric positive definite matrix. Then, for any $p > 3$, the derivatives of $(\det(D))^p$ are given up to the third order by*

$$\begin{aligned}\frac{\partial(\det(D))^p}{\partial D_{lm}} &= p(\det(D))^{p-1}(D^{-1})_{lm}\\ \frac{\partial^2(\det(D))^p}{\partial D_{lm}\partial D_{pq}} &= \frac{p!}{(p-2)!}(\det(D))^{p-2}(D^{-1})_{pq}(D^{-1})_{lm}\\ &\quad + p(\det(D))^{p-1}(D^{-1})_{lp}(D^{-1})_{mq}\\ \frac{\partial^3(\det(D))^p}{\partial D_{lm}\partial D_{pq}\partial D_{rs}} &= \frac{p!}{(p-3)!}(\det(D))^{p-3}(D^{-1})_{rs}(D^{-1})_{pq}(D^{-1})_{lm}\\ &\quad + \frac{p!}{(p-2)!}(\det(D))^{p-2}\Big\{(D^{-1})_{pq}(D^{-1})_{lr}(D^{-1})_{ps} + (D^{-1})_{pr}(D^{-1})_{qs}(D^{-1})_{lm}\end{aligned}$$

$$+ (D^{-1})_{rs}(D^{-1})_{lp}(D^{-1})_{mq}\Big\}$$
$$+ p(\det(D))^{p-1}\Big\{(D^{-1})_{lr}(D^{-1})_{qs}(D^{-1})_{mq} + (D^{-1})_{lp}(D^{-1})_{mr}(D^{-1})_{qs}\Big\}$$

Proof. We can express the derivative of the detrminant (Petersen and Pedersen, 2012),

$$\frac{\partial, \det(\boldsymbol{X})}{\partial \boldsymbol{X}_{ij}} = \det(\boldsymbol{X})(\boldsymbol{X}^{-1})_{ij} \tag{F.4}$$

Repeated application of this and (F.3) yields the result. □

Lemma F.15. *For any $p > 4$ define the function $\gamma(N, vec(D)) : \mathbb{R}^d \times \mathbb{R}^{d^2}$ by*

$$\gamma(N, vec(D)) := \begin{cases} (\det(D))^p (N' D^{-1} N - c) & \text{if } \det(D) \neq 0 \\ 0 & \text{if } \det(D) = 0 \end{cases}$$

This function is thrice continously differentiable. Futher the k^{th} moments of all partial derivatives of this function up to the third order are bounded

$$\mathbb{E}[(\partial^\alpha \gamma(N, vec(D))^k] \leq C_k(\max_{\iota \leq d} \mathbb{E}[|D_{\iota\iota}|^{2pdk}] \vee \max_{\iota \leq d} \mathbb{E}[|N_{\iota\iota}|^{6k})$$

where C_k is a positive constant that only depends on k and d.

Proof. The first statement is clear by examination of the derivatives in Lemmas F.13 and F.14 as well as the inequality (F.5) below. For the moment bounds, we may extensive use of following bounds on elements of D^{-1} for a positive-definite D^{-1}:

$$\begin{aligned} |\det(D)(D^{-1})_{jk}| \leq \det(D)\text{trace}(D^{-1}) &\leq d\lambda_{\max}(D^{-1})\Big(\prod_{m=1}^{d} \lambda_m(D)\Big) \\ &= d\prod_{m=2}^{d} \lambda_m(D) \\ &\leq d\Big(\sum_{m=2}^{d} \lambda_m(D)\Big)^{d-1} \\ &\leq d(\text{trace}(D))^{d-1} \end{aligned} \tag{F.5}$$

where the first inequality uses the fact that the largest element of a positive semidefinite matrix is on the diagonal and the fact that the diagonal elements of a positive semidefinite matrix are weakly positive, the second inequality uses the fact that the trace is the sum of the eigenvalues and the determinant is the product of the eigenvalues, the equality comes from $\frac{1}{\lambda_{\min}(D)} = \lambda_{\max}(D^{-1})$, the third inequality uses the AM-GM inequality and the fourth again uses that the trace is the sum of the (weakly positive) eigenvalues.

The moment bounds follow from (F.5) and the expressions in Lemmas F.13 and F.14. We give an example of how this is done for the first order derivatives, higher order derivatives follow from similar logic. For the following let A be an arbitrary random variable. *First Order.*

$$\mathbb{E}\left|A\frac{\partial \gamma}{\partial N_l}\right|^k \lesssim \sum_{j=1}^{d} \mathbb{E}|(\text{trace}(D))^{kdp} N_j^k A^k|$$

$$\lesssim \sum_{j=1}^{d}\sum_{\iota=1}^{d}\mathbb{E}[D_{\iota\iota}^{kdp}N_j^kA^k]$$
$$\leq \sum_{j=1}^{d}\sum_{\iota=1}^{d}\gamma^{2kdp}\mathbb{E}[N_j^{2k}A^{2k}]$$
$$\mathbb{E}\left|A\frac{\partial\gamma}{\partial D_{lm}}\right|^k = p\mathbb{E}\left|A\det(D)^{p-1}\sum_{j=1}^{d}\sum_{j'=1}^{d}(D^{-1})_{lm}(D^{-1})_{jj'}N_jN_{j'}\right|^k$$
$$\lesssim p\sum_{j=1}^{d}\sum_{j'=1}^{d}\mathbb{E}[|(\mathrm{trace}(D))^{2k(d-1)+(p-3)kd}A^kN_j^kN_{j'}^k|$$
$$\leq \sum_{j=1}^{d}\sum_{j'=1}^{d}\gamma^{2kd(p-1)}\mathbb{E}[A^{2k}N_j^{2k}N_{j'}^{2k}]$$

□

G Technical Lemmas

G.1 Probability Lemmas

Lemma G.1. *Let X_n be a sequence of random variables such that $X_n = o_p(1)$, that is for any $\delta > 0$, $\Pr(|X_n| \geq \delta) \to 0$. Then, there is a sequence $\delta_n \to 0$ such that $\Pr(|X_n| \geq \delta_n) \to 0$.*

Proof. Take a preliminary sequence $\tilde{\delta}_n \to 0$ and define

$$\tilde{n}_j = \inf\{n : \Pr(|X_n| > \tilde{\delta}_j) < \tilde{\delta}_j\}$$

Because $\Pr(|X_n| > \delta) \to 0$ for any fixed δ, we know that n_j is finite. Define a new sequence $\delta_n \to 0$ as below:

$$\delta_n = \begin{cases} 1 & \text{if } 0 \leq n < \tilde{n}_1 \\ \tilde{\delta}_i & \text{if } \tilde{n}_i \leq n < \tilde{n}_{i+1} \end{cases} \tag{G.1}$$

By construction, this sequence satisfies $\Pr(X_n \geq \delta_n) \leq \delta_n$ whenever $n \geq n_1$. □

Lemma G.2. *Suppose that $X_1, \ldots, X_n$ are α-subexponential such that $\Pr(|X_i| \geq t) \leq 2\exp(-t^\alpha/K)$ for all $t \geq 0$ and fixed constants K. For any $p \geq 1$ there is a constant C that depends only on p, K such that:*

$$\mathbb{E}\left[\max_{i\leq n}\frac{|X_j|^p}{(1+\log i)^{p/\alpha}}\right] \leq C$$

As a consequence

$$\mathbb{E}\left[\max_{i\leq n}|X_i|^p\right] \leq C(\log n)^{p/\alpha}$$

Proof. Argument below is provided for $\alpha = 1$. This can be extended to $\alpha \neq 1$ by noting that if $\Pr(|X_i| \geq t) \leq 2\exp(-t^\alpha/K)$ for some $\alpha > 0$ then $\Pr(|X_i|^\alpha \geq t) \leq 2\exp(-t/K)$.

$$\mathbb{E}\max_{i\leq n}\frac{|X_i|^p}{(1+\log i)^p} = \int_0^\infty \Pr\left(\max_i \frac{|X_i|^p}{(1+\log i)^p} > t\right)dt$$

$$
\begin{aligned}
&= \int_0^{2^{p/\alpha}} \Pr\left(\max_i \frac{|X_i|^p}{(1+\log i)^p} > t \right) dt + \int_{2^{p/\alpha}}^{\infty} \Pr\left(\max_i \frac{|X_i|^p}{(1+\log i)^p} > t \right) dt \\
&\leq 2^p + \int_{2^{p/\alpha}}^{\infty} \sum_{i=1}^{n} \Pr\left(\frac{|X_i|}{1+\log i} > t^{1/p} \right) dt \\
&\leq 2^p + \int_{2^p}^{\infty} \sum_{i=1}^{n} 2\exp\left(-\frac{t^{1/p}(1+\log i)}{K} \right) dt \\
&= 2^p + 2\sum_{i=1}^{n} \int_{2^p}^{\infty} \exp\left(-\frac{t^{1/p}}{K} \right) i^{-t^{1/p}}\, dt \\
&\leq 2^p + 2\sum_{i=1}^{n} \int_{2^p}^{\infty} \exp(-t^{-1/p}/K) i^{-2}\, dt \\
&\leq 2^p + 2\left(\sum_{i=1}^{n} i^{-2} \right)\left(\int_{2^p}^{\infty} \exp(-t^{-1/p}/K)\, dt \right)
\end{aligned}
$$

Both the integral and the summation are bounded, which gives the result. □

G.2 Matrix Lemmas

Lemma G.3. *Given a matrix M and a matrix P of full rank, the matrix M and the matrix $P^{-1}MP$ have the same eigenvalues.*

Proof. Suppose λ is a eigenvalue of $P^{-1}MP$ with eigenvector p. Then

$$P^{-1}MPv = \lambda v \implies M(Pv) = \lambda Pv$$

Hence Pv is an eigenvector of M with eigenvalue λ. Similarly, given an eigenvector v of M, it can be shown that $P^{-1}v$ is an eigenvector of $P^{-1}MP$;

$$P^{-1}MP(P^{-1}v) = P^{-1}Mv = \lambda P^{-1}v$$

□

Lemma G.4. *Let $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times n}$ be real symmetric positive semidefinite matrices. For an arbitary square matrix M let $\lambda_k(M)$ denote the k^{th} largest eigenvalue of M. Then for any $k = 1, \ldots, n$:*

$$\lambda_k(A)\lambda_n(B) \leq \lambda_k(AB) \leq \lambda_k(A)\lambda_1(B)$$

Lemma G.5. *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 $k = 1, \ldots, n$:*

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

Proof. Consider any vector $a \in \mathbb{R}^n$ and define $\boldsymbol{a} = a'H$. Then

$$
\begin{aligned}
\alpha' HDH\alpha = \boldsymbol{a}' D\boldsymbol{a} = \sum_{i=1}^{n} d_{ii}(\boldsymbol{a}_i)^2 &\in \left[u\sum_{i=1}^{n}(\boldsymbol{a}_i)^2,\ U\sum_{i=1}^{n}(\boldsymbol{a}_i)^2 \right] \\
&= \left[u \times a'H^2a,\ U \times a'H^2a \right]
\end{aligned}
$$

The result then follows from an application of Courant-Fischer-Weyl min-max principle. □

Lemma G.6. *Let $X_1, \dots, X_n$ denote i.i.d standard normal random variables and $a_1, \dots, a_n$ denote weakly positive constants. Then*

$$\Pr\left(\sum_{i=1}^{n} a_i X_i^2 \leq \epsilon \sum_{i=1}^{n} a_i\right) \leq \sqrt{e\epsilon}$$

G.3 Miscellaneous Lemmas

Lemma G.7. *Let $a_1, \dots, a_n$ and $b_1, \dots, b_n$ be two sequences of real numbers. If $a_i \leq U b_i$ for some $U > 0$, then $\sum_i a_i / \sum_i b_i \leq U$. Conversely if $a_i \geq L b_i$ for some $L > 0$ then $\sum_i a_i / \sum_i b_i \geq L$.*

Proof. Replace $a_i \leq U b_i$ for the upper bound and $a_i \geq L b_i$ for the lower bound. □

The following is a standard bound, but it is used a lot so it is restated here.

Lemma G.8. *Let $a_1, \dots, a_m$ be constants and $p > 1$. Then*

$$|a_1 + \dots a_m|^p \leq m^{p-1} \sum_{i=1}^{m} |a_i|^p$$

Proof. Apply Hölder's inequality with $\frac{1}{p} + \frac{p-1}{p} = 1$ to the vectors $(a_1, \dots, a_m) \in \mathbb{R}^m$ and $(1, \dots, 1) \in \mathbb{R}^m$ □

H Assorted Results from Literature

H.1 Concentration Inequalities and Tail Bounds

Theorem H.1 (Gotze et al. (2021)*Theorem 1.2)**.** *Let $X_1, \dots, X_n$ 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 $t > 0$;*

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

In particular, if $\|\mathbb{E} f^{(d)}(X)\|_{HS} \leq 1$ for $d = 1, \dots D$, then

$$\mathbb{E} \exp\left(\frac{C_{D,a}}{M^a} |f(X)|^{\frac{a}{D}}\right) \leq 2,$$

or equivalently

$$\|f(X)\|_{\Psi_{\frac{a}{D}}} \leq C_{d,a} M^D$$

Theorem H.2 (Hoeffding's Inequality)**.** *Let $X_1, \dots, X_n$ be independent, mean-zero sub-gaussian random variables, and let $a = (a_1, \dots, a_n) \in \mathbb{R}^n$. Then, for every $t \geq 0$, we have*

$$\Pr\left\{\left|\sum_{i=1}^{n} a_i X_i\right| \geq t\right\} \leq 2 \exp\left(-\frac{ct^2}{K^2 \|a\|_2^2}\right)$$

where $K = \max_i \|X_i\|_{\psi_2}$.

Theorem H.3 (Burkholder-Davis-Gurdy for Discrete Time Martingales)**.** *For any $1 \leq k < \infty$ there exist positive constants c_k and C_k such that for all local martingales with $X_0 = 0$ and stopping times τ*

$$c_k \mathbb{E}\Big[\Big(\sum_{t=1}^{\tau}(X_t - X_{t-1})^2\Big)^{k/2}\Big] \leq \mathbb{E}\Big[(\sup_{t\leq\tau} X_t)^k\Big] \leq C_k \mathbb{E}\Big[\Big(\sum_{t=1}^{\tau}(X_t - X_{t-1})^2\Big)^{k/2}\Big]$$

H.2 ANTICONCENTRATION BOUNDS

Let $\xi \in \mathbb{R}^n$ follow a normal distribution on $\mathbb{R}^n$ with mean zero and covariance matrix Σ_ξ. Order the eigenvalues of Σ_ξ in non-increasing order $\lambda_{1\xi} \geq \lambda_{2\xi} \geq ... \geq \lambda_{n\xi}$. Define the quantities

$$\Lambda_{k\xi}^2 = \sum_{j=k}^{\infty} \lambda_{j\xi}^2, \quad k = 1, 2$$

Theorem H.4 (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 $\|\xi - \boldsymbol{a}\|^2$.

We use the following anticoncentration lemma from Nazarov (2003) noted in Chernozhukov et al. (2017).

Lemma H.1. *Let $Y = (Y_1, \ldots, Y_p)'$ be a centered Gaussian random vector in $\mathbb{R}^p$ such that $\mathbb{E}[Y_j^2] \geq b$ for all $j = 1, \ldots, p$ and some constant $b > 0$. Then for every $y \in \mathbb{R}^p$ and $a > 0$,*

$$\Pr(Y \leq y + a) - \Pr(Y \leq y) \leq Ca\sqrt{\log(p)}$$

where C is a constant only depending on b.

H.3 GAUSSIAN COMPARASIONS AND APPROXIMATIONS

We also use the following gaussian approximation results from Belloni et al. (2018), Chernozhukov et al. (2017). Let $X_1, \ldots, X_n \in \mathbb{R}^p$ be independent, mean zero, random vectors and let $Y_1, \ldots, Y_n \in \mathbb{R}^p$ be independent random vectors such that $Y_i \sim N(0, \mathbb{E}[X_i X_i'])$. Suppose that the researcher does not directly observe $X_1, \ldots, X_n$ but instead observes noisy estimates $\widehat{X}_1, \ldots, \widehat{X}_n \in \mathbb{R}^p$.

Define the sums

$$S_n^X = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \widehat{X}_i \qquad\qquad S_n^Y = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i$$

Let $\mathcal{A}^{\text{re}}$ be the class of all hyperrectangles in $\mathbb{R}^p$; that is, $\mathcal{A}^{\text{re}}$ consists of all sets A of the form

$$A = \{w \in \mathbb{R}^p : a_j \leq w_j \leq b_j \text{ for all } j = 1, \ldots, p\}$$

for some $-\infty \leq a_j \leq b_j \leq \infty$, $j = 1, \ldots, p$. Define

$$\rho_n(\mathcal{A}^{\text{re}}) := \sup_{A\in\mathcal{A}^{\text{re}}} \left|\Pr(S_n^X \in A) - \Pr(S_n^Y \in A)\right|$$

Bounding $\rho_n(\mathcal{A}^{\text{re}})$ relies on the following moment conditions:

Assumption H.1. *Suppose there are constants $B_n \geq 1$, $b > 0$, $q > 0$ such that*

(i) $n^{-1} \sum_{i=1}^n \mathbb{E}[X_{ij}^2] \geq b$ *for all* $j = 1, \ldots, p$

(ii) $n^{-1} \sum_{i=1}^n \mathbb{E}[|X_{ij}|^{2+k}] \leq B_n^k$ *for all* $j = 1, \ldots, p$ *and* $k = 1, 2$.

(iii) $\mathbb{E}[(\max_{1 \leq j \leq p} |X_{ij}|/B_n)^4] \leq 1$ *for all* $i = 1, \ldots, n$ *and* $\left(\frac{B_n^4 \ln^7(pn)}{n} \right)^{1/6} \leq \delta_n$.

as well as the following bounds on the estiamtion error

Assumption H.2. *The estimates $\hat{X}_1, \ldots, \hat{X}_n$ satisfy*

$$\Pr\left(\max_{1 \leq j \leq p} \mathbb{E}_n[(\widehat{X}_{ij} - X_{ij})^2] > \delta_n^2 / \log^2(pn) \right) \leq \beta_n$$

Theorem H.5 (Belloni et al. (2018), Theorem 2.1)**.** *Suppose that Assumptions H.1 and H.2 hold. Then there is a constant C which depends only on b such that*

$$\rho_n(\mathcal{A}^{re}) \leq C\{\delta_n + \beta_n\}$$

Let $e_1, \ldots, e_n \overset{\text{iid}}{\sim} N(0,1)$ be generated independently of the data. A gaussian bootstrap draw is defined

$$S_n^{X,\star} := \frac{1}{\sqrt{n}} \sum_{i=1}^n e_i \widehat{X}_i$$

Theorem H.6 (Belloni et al. (2018), Theorem 2.2)**.** *Suppose that Assumptions H.1 and H.2 hold. Then there is a constant C which depends only on b such that*

$$\sup_{A \in \mathcal{A}^{re}} \left| \Pr_e(S_n^{X,\star} \in A) - \Pr(S_n^Y \in A) \right| \leq C\delta_n$$

with probability at least $1 - \beta_n - (\log n)^{-2}$ where $\Pr_e(\cdot)$ denotes the probability measure only taken with respect to the variables $e_1, \ldots, e_n$ conditional on the data used to estimate $\widehat{X}$.