Journal of Econometrics 258 (2026) 106348



Contents lists available at ScienceDirect

Journal of Econometrics

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





An identification-and dimensionality-robust test for instrumental variables models ⋆

Manu Navjeevan

Texas A&M University, USA

ARTICLE INFO

JEL classification:
C12
C36
C55

Keywords:
Instrumental variables
Weak identification
High-dimensional

ABSTRACT

Using modifications of Lindeberg's interpolation technique, I propose a new identification-robust test for the structural parameter in a heteroskedastic instrumental variables model. My analysis accommodates both scenarios where the number of instruments is much larger than the sample size and scenarios with few instruments, thus making the test broadly applicable. The proposed test statistic has a limiting chi-squared distribution so long as an auxiliary parameter can be consistently estimated. This is possible with standard methods when the number of instruments is small or using machine learning methods when the intruments are high-dimensional. To improve power, a simple combination with the sup-score statistic is proposed. The flexibility of my asymptotic approach allows for higher quality first-stage estimators to be used in the construction of the test statistic which provides noticeable power improvements over existing tests, both in Monte Carlo simulations calibrated to empirical data and in an application to the returns to education. Motivated by empirical practice, I also point out that first-stage F-statistics calculated on LASSO-selected variables may be misleading indicators of identification strength.

1. Introduction

When instruments are suspected to be weak, researchers may want to test hypotheses about structural parameters using testing procedures that are robust to identification strength. These procedures typically rely on some conditions on the rate of growth of the number of instruments, d_z, in relation to the sample size, n. The initial identification robust tests developed in Staiger and Stock (1997), Moreira (2003), and Kleibergen (2005) are shown by Andrews and Stock (2007) to control size under heteroskedasticity when the number of instruments cubed is small relative to the sample size, $d_z^3/n \to 0$; Newey and Windmeijer (2009) establish the validity of K- and conditional likelihood ratio-type statistics in GMM models with many weak moment conditions under the same $d_z^3/n \to 0$ condition. Meanwhile, recent and interesting "many-instrument" tests (Crudu et al., 2021; Mikusheva and Sun, 2021; Matsushita and Otsu, 2022; Lim et al., 2022) allow the number of instruments to be proportional to the sample size, $d_z/n \to \rho \in [0, 1)$, but require that the number of instruments itself be large, $d_z \to \infty$.

In practice, these conditions can be difficult to interpret and the variety of tests available under alternate regimes may make it difficult for the researcher to know which test, if any, should be applied in her exact setting. As examples, consider settings such as that of Derenoncourt (2022), where $d_z = 9$ and $n = 130$, Paravisini et al. (2014), where $d_z = 10$ and $n = 5{,}995$, and Gilchrist and Sands

⋆ 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.

E-mail address: mnavjeevan@tamu.edu

https://doi.org/10.1016/j.jeconom.2026.106348
Received 13 November 2025; Received in revised form 24 July 2026; Accepted 11 October 2026
Available online 23 October 2026
0304-4076/© 2026 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

(2016) where $d_z = 52$ and $n = 1{,}671$. In all three cases, the number of instruments cubed, $d_z^3 = 729$, 1,000 and 140,608, respectively, cannot be treated as negligible relative to the sample size. Indeed, all of these papers use post-LASSO estimates of the first-stage, suggesting concern about the large number of instruments by the authors. However as the number of instruments is only moderately large in all of these settings, asymptotic approximations that rely on $d_z \to \infty$ may not accurately resemble the finite sample distribution of the test statistic. This can make size control of many-instrument tests questionable.[1]

By comparison, the test proposed in this paper can be applied in any of the settings listed above as well as when the instruments are high-dimensional, $d_z \gg n$, a setting for which there has been limited progress on identification robust testing to this point. The main problem in these settings has been that the limiting behaviors of the regularized first-stage estimators used when $d_z \gg n$ are difficult to analyze and usually unknown. Existing analyzes sidestep this issue by assuming strong identification and exploiting a certain orthogonality – termed "Neyman Orthogonality" by Chernozhukov et al. (2018) – of the structural parameter estimate to the first-stage estimation error (Belloni et al., 2012). This approach is explicitly not applicable under weak identification where the first-stage estimation error is of a weakly greater order than the signal from the instruments and thus relevant to asymptotic analysis. As such, the few existing proposals for identification robust testing that allow $d_z \gg n$ either fail to fully incorporate first-stage information (Belloni et al., 2012; Doví et al., 2024; Lim et al., 2024) or rely on sample splitting (Mikusheva, 2023), both of which may reduce power.

To construct the test statistic I adapt an idea from Kleibergen (2002, 2005) and leverage a conditional slope parameter, which can be simply estimated using regularized methods even when $d_z \gg n$, to partial out the structural error from the endogenous variable. I then use these partialled out versions of the endogenous variables to construct first-stage estimates. The K-statistic of Kleibergen (2002) was originally proposed to remove the weak-instrument bias in the Lagrange Multiplier statistic. Matsushita and Otsu (2022) recognize that jackknifing the Lagrange Multiplier statistic removes this bias as well, and extend the resulting statistic to many-instrument asymptotics, a point emphasized in the recent review of Mikusheva and Sun (2024). The approach taken in this paper instead extends the K-statistic itself, so that the proposed test can be viewed as a direct extension of the K-tests of Kleibergen (2002, 2005) to settings with many, or even high-dimensional, instruments. The key insight is that, if variables in the model followed a jointly Gaussian distribution, these first-stage estimates would be independent of the structural errors under the null – a fact resulting from the statistical independence of uncorrelated jointly Gaussian random variables. Thus, even under weak identification where their behavior is relevant to the distribution of the proposed test statistic, the limiting χ^2 distribution of the proposed test statistic could be easily derived by conditioning on these estimates. Deriving the limiting distribution of the test statistic in the general model then reduces to showing that one can treat all observations as if they followed a jointly Gaussian distribution.

Because the number of instruments is allowed to be large, standard asymptotic approaches based on central limit theorems cannot be easily applied. My asymptotic analysis thus takes a different tack and directly shows that variables can be treated as if they followed a jointly Gaussian distribution via modifications of Lindeberg's interpolation argument (Lindeberg, 1922). Roughly, I proceed by showing to be negligible the total change in distribution from a one-by-one replacement of each observation in the expression of the test statistic with a Gaussian version. The modifications of Lindeberg's original interpolation method are non-trivial and required to deal with the fact that derivatives of the test statistic with respect to individual observations may be unbounded, an issue arising from the "divide-by-zero" problem of weak identification (Andrews et al., 2019).[2] This interpolation argument requires some conditions on the first-stage estimates. In particular, I require that the first-stage estimates take on a "jackknife-linear" form and suggest using a jackknife ridge regression in practice to allow $d_z \gg n$. Interestingly, though, these first-stage estimates are not required to converge to limiting values which allows the researcher some flexibility if she wanted to use a different approach.

Through the Gaussian approximation result, I examine the power properties of my proposed testing procedure in local neighborhoods of the null. These local neighborhoods are characterized by a bounded local power index. In the case of a single endogenous variable I show that, under an additional regularity condition, the local power index diverging implies that the test is consistent. Unfortunately, the process of partialling out the structural error can introduce bias into the first-stage estimate under the alternative hypothesis. Against certain alternatives, this bias can be particularly pronounced and erase the first-stage signal from the instruments, a problem pointed out by Andrews (2016) in the context of the Kleibergen (2005) K-statistic. To address this, I propose a simple combination with the sup-score test of Belloni et al. (2012), which can also allow $d_z \gg n$. As with the Anderson-Rubin test, while the sup-score statistic does not incorporate first-stage information it does not face a power decline against particular alternatives.

Identification-robust testing procedures may be of particular interest in high-dimensional settings due to a lack of clarity on how to pretest for weak identification. When using post-LASSO estimates of the first stage, current empirical practice appears to be conducting standard t-test inference if the first stage F-statistic calculated on the LASSO-selected variables is larger than 10 (Paravisini et al., 2014; Gilchrist and Sands, 2016; Derenoncourt, 2022). Using a simple numerical demonstration, I argue that first stage F-statistics calculated on LASSO selected variables may not be reliable indicators of identification strength, so that identification-robust inference is warranted even when the reported first-stage F-statistic is large. With this motivation, I apply the newly proposed testing procedures to the returns-to-education data of Angrist and Krueger (1991) and find that the resulting weak instrument-robust confidence intervals

[1] Under fixed d_z with homoskedastic errors, Mikusheva and Sun (2021) show their proposed test statistic's 95th quantile is close, but not equal, to that of the standard normal. Lim et al. (2024) propose a bootstrap critical value that controls size under heteroskedasticity when the number of instruments are fixed, however, as discussed in Section 5, this test does not fully incorporate first-stage information when the number of instruments is large.

[2] Existing interpolation results require these derivatives to be bounded while in this setting they may not even have finite moments. Given this, these modified approaches may be of independent interest to a growing literature on direct Gaussian approximation techniques (Chatterjee, 2006; Chernozhukov et al., 2013; Pouzo, 2015; Celentano et al., 2020).

are about half the size of those obtained by inverting the many-instrument tests, or smaller. A companion confidence-interval analysis of the social-spillovers data of Gilchrist and Sands (2016), presented in Appendix L, shows the same pattern. I discuss how these improvements in power compared to existing tests can be explained by the proposed test statistic's use of higher quality first-stage estimators and individual scores that are uncorrelated across observations.

Finally, I examine the applicability of the theoretical results in this paper through Monte Carlo simulations calibrated to the data of Gilchrist and Sands (2016). The calibrated designs hold the observed instruments and an empirically estimated first stage fixed while growing the instrument dictionary from 48 to 1182 instruments (d_z/n from 0.03 to 0.71). Errors are drawn both under a baseline design in which the auxiliary parameter is approximately sparse and under a design with dense, nonlinear heteroskedasticity of the form $\sigma(z) \propto \exp(\gamma' z)$, which purposefully strains the sparsity conditions used to estimate the auxiliary parameter. While existing tests seem to face size distortions in one dimensionality regime or another, the test based on my proposed test statistic has nearly exact size in every setting considered, under both designs, and its power advantage over the many-instrument tests grows with the ratio of instruments to sample size. While this new test may have diminished power against certain alternatives, this deficiency is ameliorated through combination with the sup-score test, which preserves nearly all of the power of the new test in the calibrated designs.

The outline of this paper is as follows. Section 2 formally defines the model considered and introduces the new test as well as it's combination with the sup-score test. Section 3 evaluates the proposed testing procedures in Monte Carlo designs calibrated to the data of Gilchrist and Sands (2016) and demonstrates their usefulness in an application to the returns-to-education data of Angrist and Krueger (1991) (Section 3.4). Section 4 provides an overview of the Gaussian approximation approach and characterizes the limiting behavior of the test statistic. Section 5 uses this characterization to examine the power properties of the test, establishes the validity of the combination test, and provides potential explanations for observed power improvements compared to the many-instrument tests. Proofs of the main results as well as the presentation of some auxiliary results are deferred to the Appendix.

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.

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 only grow slowly with the sample size, this is without loss of generality.[3] Along with the first stage, the IV model can then be written as a system of simultaneous equations:

$$\begin{aligned} y_i &= x_i'\beta + \epsilon_i \\ x_i &= \Pi_i + v_i \end{aligned} \tag{1}$$

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 $\epsilon_i \in \mathbb{R}$ 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, $\mathbb{E}[\epsilon_i | z_i] = 0$. 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 $\{(z_i, \epsilon_i, v_i)\}_{i=1}^n$ 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:

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

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 $z := (z_1', \dots, z_n') \in \mathbb{R}^{n \times d_z}$ 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:

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

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 $\epsilon_i(\beta_0) := y_i - x_i'\beta_0$. Using these, I construct a variable, r_i, that is a "partialed-out" version of the endogenous variable satisfying $Cov(r_i, \epsilon_i(\beta_0)) = 0$:

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

[3] Section G provides a formal treatment of the model with controls allowing for the number of controls, d_c, to grow so long as $d_c^3/n \to 0$. For settings with high-dimensional controls, Doví (2025) proposes a modification of the sup-score test that controls size under a sparsity assumption. However, at present it is unclear how a test could be constructed that allows for high-dimensional controls and incorporates first stage information.

The construction of r_i can be viewed as an observation-by-observation analog of the recentering that underlies the heteroskedasticity-robust K-statistic of Kleibergen (2005). In a GMM formulation of the model, that statistic recenters the moment Jacobian $z_i \otimes x_i$ by subtracting its projection onto the null moment vector, replacing $z_i \otimes x_i$ with

$$(z_i \otimes x_i) - \bar{\mathbb{E}}\left[(z_i \otimes x_i)(z_i \otimes \epsilon_i(\beta_0))'\right]\left(\bar{\mathbb{E}}\left[(z_i \otimes \epsilon_i(\beta_0))(z_i \otimes \epsilon_i(\beta_0))'\right]\right)^{-1}(z_i \otimes \epsilon_i(\beta_0)).$$

Carrying out this recentering directly requires estimating the $d_x d_z \times d_z$ covariance matrix $\bar{\mathbb{E}}[(z_i \otimes x_i)(z_i \otimes \epsilon_i(\beta_0))']$ and inverting an estimate of the $d_z \times d_z$ matrix $\bar{\mathbb{E}}[(z_i \otimes \epsilon_i(\beta_0))(z_i \otimes \epsilon_i(\beta_0))']$, both of which are high-dimensional objects when the number of instruments is large; indeed, Newey and Windmeijer (2009) establish the validity of the K-statistic with many weak moment conditions under the condition that $d_z^3/n \to 0$. The construction above instead carries out the recentering observation by observation, $z_i \otimes r_i = (z_i \otimes x_i) - (z_i \otimes \rho(z_i))\epsilon_i(\beta_0)$, replacing these high-dimensional matrix estimation problems with the estimation of the single d_x-dimensional function $\rho(\cdot)$. This is what allows the test developed in this paper to accommodate $d_z \gg n$: as described below, $\rho(\cdot)$ can be estimated accurately with ℓ_1-penalized methods whenever it admits an approximately sparse representation. Remark 2.3 discusses further differences from the K-statistic of Kleibergen (2005).

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:[4]

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

While other estimators of $\rho(\cdot)$ are possible (see Section 4.3), I focus on ℓ_1-penalized/LASSO estimators. These estimators are consistent under the assumption that $\rho(z_i)$ has an approximately sparse representation in some 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 low-dimensional case and the high dimensional cases under a unified estimation procedure. Interestingly, if errors are homoskedastic, $\rho_\ell(z_i)$ is a constant function and thus has a sparse representation in any basis that contains a constant term. In general, however, the approximate sparsity assumption can either be interpreted as an assumption that there are only a few instruments that are important for explaining variation in the covariance matrix $\Omega(z_i)$ or as an assumption that the function $\rho(z_i)$ can be accurately approximated using only a smaller set of basis terms in $b(z_i)$.

The parameter ϕ_ℓ can be estimated via LASSO:

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

or via post-LASSO, refitting an unpenalized version of (2) using only the basis terms associated with nonzero coefficients in the initial LASSO regression. The estimating procedure in (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 4.3. With $\hat{\rho}_\ell(z_i) := b(z_i)'\hat{\phi}_\ell$, I construct the estimated version of $r_{\ell i}$, $\hat{r}_{\ell i} := x_{\ell i} - \hat{\rho}_\ell(z_i)\epsilon_i(\beta_0)$ 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,

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

where the weights h_{ij}, which may depend only on the instruments $\boldsymbol{z}$, are collected in a "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, $\boldsymbol{z}(\boldsymbol{z}'\boldsymbol{z})^{-1}\boldsymbol{z}'$, is sometimes referred to as the hat matrix in the sense that $\hat{x} = \boldsymbol{z}(\boldsymbol{z}'\boldsymbol{z})^{-1}\boldsymbol{z}'x$. While the formal analysis allows for any hat matrix satisfying a balanced-design condition described in Section 4, the recommended choice — and the one implemented in the simulations and empirical application of Section 3 — is the jackknife ridge procedure described in (3) below. It is important to note that while $\hat{\Pi}_{\ell i}$ does not depend on $\hat{r}_{\ell i}$, 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 $\hat{\Pi}_{\ell i}$ can be written as $\hat{\Pi}_{\ell i} = \sum_{j=1}^n h_{ij}r_{\ell j}$.

For reasons explained in Section 5, it may be optimal to introduce some regularization in estimating the first-stage models $\hat{\Pi}_{\ell i}$, so I suggest using a jackknife ridge regression procedure setting:

$$\hat{\Pi}_i = z_i'\hat{\pi}_{(-i)}(\lambda^\star) \tag{3}$$

where $\hat{\pi}_{(-i)}(\lambda)$ is the coefficient estimate from a ridge regression of $\hat{r}$ on $\boldsymbol{z}$, leaving out observation i and with penalty parameter set equal to λ.[5] Following recommendations in van Wieringen (2023), the penalty parameter $\lambda^\star$ is set so that the effective degrees of

[4] 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.

[5] A ridge regression coefficient estimate from a regression of $\tilde{\boldsymbol{y}} \in \mathbb{R}^n$ on $\tilde{\boldsymbol{x}} \in \mathbb{R}^{k \times n}$ with penalty parameter $\lambda \in \mathbb{R}_+$ solves $\hat{\pi} \in \arg\min_{\pi \in \mathbb{R}^k} \|\tilde{\boldsymbol{y}} - \pi'\tilde{\boldsymbol{x}}\|_2^2 + \lambda\|\pi\|_2^2$. Nyquist (1988) shows how the fitted values from a jackknife ridge regression can be calculated without having to recompute $\hat{\pi}_{(-i)}$ for each observation. Angrist et al. (1999) provides a similar analysis for jackknife OLS.

freedom is no more than a fraction of the sample size:

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

The jackknife ridge estimate has the benefit of being well defined even when the number of instruments is larger than the sample size. I stress, though, that the $\hat{\Pi}_{\ell i}$ estimators are not required to be consistent and the researcher may use any other hat matrices that she believes will lead to plausible first-stage estimates. Other possible choices of first stage estimator include the jackknife OLS procedure of Angrist et al. (1999) or estimators based on selecting instruments via some preliminary unsupervised technique such as principal component analysis (PCA). Remark 4.1 below discusses how the balanced-design condition may be verified for arbitrary choices of hat matrices.

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

$$\begin{aligned} \epsilon(\beta_0) &= \big(\epsilon_1(\beta_0), \dots, \epsilon_n(\beta_0)\big)' \in \mathbb{R}^n \\ \hat{\Pi} &= \big(\hat{\Pi}_1', \dots, \hat{\Pi}_n'\big)' \in \mathbb{R}^{n \times d_x} \\ \hat{\Pi}_\epsilon &= \big(\hat{\Pi}_{\epsilon 1}', \dots, \hat{\Pi}_{\epsilon n}'\big)' \in \mathbb{R}^{n \times d_x} \end{aligned} \tag{4}$$

The jackknife K-statistic can then be defined

$$JK(\beta_0) = \epsilon(\beta_0)'\hat{\Pi}\big(\hat{\Pi}_\epsilon'\hat{\Pi}_\epsilon\big)^{-1}\hat{\Pi}'\epsilon(\beta_0) \times \mathbf{1}\{\lambda_{\min}(\hat{\Pi}_\epsilon'\hat{\Pi}_\epsilon) > 0\} \tag{5}$$

If all variables were jointly Gaussian and $\rho(\cdot)$ was known to the researcher, the first stage estimates, $\hat{\Pi}$, would be independent of the implied errors, $\epsilon(\beta_0)$.[6] The researcher could then easily derive the limiting $\chi^2_{d_x}$ distribution of $JK(\beta_0)$ by conditioning on $\hat{\Pi}$ and noticing that the result looks like a self normalized sum (Peña et al., 2008). Deriving the limiting $\chi^2_{d_x}$ distribution of $JK(\beta_0)$ in the general model thus reduces to showing that all variables can be treated as if they were jointly Gaussian and that estimation error in $\hat{\rho}(\cdot)$ can be safely ignored. I show this is possible under appropriate moment bounds and conditions on the hat matrix, H. For exposition, I will largely focus on the case where $d_x = 1$, in which case the argument can be simplified. The extension to $d_x > 1$ is not immediate but is possible under strengthened moment conditions.

Remark 2.1. Notice that, if errors are Gaussian to begin with, essentially any estimator could be used to construct the first-stage estimates without the need for jackknifing at all. Since the instruments are fixed, the first-stage estimator is a function only of the vector $r := (r_1, \dots, r_n)'$ which is independent of the implied errors $\epsilon(\beta_0)$ under joint Gaussianity. Thus, we can condition on the first-stage estimates to derive the χ^2 distribution of the test-statistic. Both jackknifing and ridge regression are used only to show Gaussian approximation. Jackknifing ensures that certain cross-terms, whose moments cannot be matched by Gaussian analogs, do not enter the test-statistic, while ridge regression ensures that the estimator has the analytically tractable linear form described at the top of this subsection. Remark 4.4 discusses further.

Remark 2.2. Importantly, the first-stage jackknife ridge estimate suggested in (3) is a "true" jackknife ridge estimate — that is it is an estimate based on recomputing the first-stage coefficients after leaving out each observation. This contrasts to the pseudo-jackknife first-stage estimates used in the many-instruments procedures and related papers by Doví et al. (2024) and Lim et al. (2024) which are based on simply deleting the diagonal of the standard, full-sample OLS or Ridge projection matrices. Using true jackknife ridge estimates complicates theoretical analysis because their asymptotic behavior in high-dimensional settings is less straightforward to characterize. The ability to use the true jackknife ridge estimates of (3) is thus a feature of the novel asymptotic approach in this paper, which allows a measure of flexibility in how the first-stage estimates are constructed. As discussed further in Section 5 and demonstrated in the calibrated simulations and empirical application of Section 3, use of the true vs. pseudo-jackknife ridge estimates appears to have substantial implications for the power properties of the relevant tests.

Remark 2.3. While use of first-stage estimates that are uncorrelated with the structural error is inspired by Kleibergen (2002, 2005), the form of the jackknife K-statistic is distinct from that of the original K-statistics. Beyond the recentering discussed at the beginning of this section, the two statistics also differ in how heteroskedasticity is accounted for in the variance estimate: whereas the K-statistic of Kleibergen (2005) again requires an estimated $d_z \times d_z$ matrix, which cannot be consistently estimated when d_z is large, the jackknife K-statistic uses the heteroskedasticity robust variance estimate $(\hat{\Pi}_\epsilon'\hat{\Pi}_\epsilon)^{-1} \in \mathbb{R}^{d_x \times d_x}$. Showing that these variance estimates can be used to account for heteroskedasticity is a feature of the direct Gaussian approximation approach. Under weak identification the distribution of the variance estimate is relevant to the distribution of the test-statistic. However, even when $d_z \ll n$, the distribution of this variance estimate would be difficult to analyze using traditional methods as it is not a continuous function of a sample mean or even of a quadratic form.

Remark 2.4. The many-instrument settings analyzed in this paper may be subject to misspecification of the linear IV model. In recent work, Kleibergen and Zhan (2025) extend the GMM K-statistic of Kleibergen (2005) to deliver identification-robust inference on the pseudo-true value of the structural parameter under misspecification with a fixed number of instruments, while Yap (2025)

[6] This is because $Cov(r_i, \epsilon_i(\beta_0)) = 0$ and uncorrelated jointly Gaussian random variables are independent. Since the vector $\hat{\Pi}$ is a function of $\{r_i : i \in [n]\}$ and $\{r_i : i \in [n]\} \perp \epsilon(\beta_0)$ we then have $\hat{\Pi} \perp \epsilon(\beta_0)$.

develops inference procedures that are robust to treatment effect heterogeneity in a many-instrument setting. Extending the jackknife K-statistic to provide identification-robust inference on a pseudo-true value under misspecification, particularly when the number of instruments is allowed to be larger than the sample size, is an interesting avenue for future research. Such an extension would require some care, however, as the nuisance parameter $\rho(z_i)$ is defined through the null-implied residual $\epsilon_i(\beta_0)$, whose interpretation changes when the linear IV model is misspecified.

2.2. Combination with sup-score test

As will be discussed in Section 5, the test based on the $JK(\beta_0)$ statistic can have deficient power against certain alternatives. This loss of power is shared with tests based on the K-statistics of Kleibergen (2002, 2005). For those tests, the power decline can be understood by recalling that the K-statistic is a quadratic form in the score of an objective function; since the score vanishes at every stationary point of the objective – maxima and saddle points as well as the minimizer – the K-statistic is equal to zero at each of these points, no matter how far they lie from the true parameter. In the present setting the same phenomenon arises through the construction of r_i: partialling out the null errors, $\epsilon(\beta_0)$, from the endogenous variables introduces bias into the first stage estimates, $\hat{\Pi}$, under the alternative hypothesis. Against certain alternatives this bias can "erase" the first stage signal from the instruments, in which case the first-stage estimates select a direction unrelated to the true deviation and the test has trivial power.

The power deficiency in tests based on the K-statistic is typically addressed by combining the K-statistic with the Anderson-Rubin statistic based on a constructed conditioning variable. Prominent examples of such combinations include the celebrated conditional likelihood test of Moreira (2003), GMM-M test of Kleibergen (2005), and minimax regret tests of Andrews (2016). In a setting with many weak instruments, Ayyar et al. (2025) recently propose a conditional likelihood ratio test that combines jackknife versions of the Anderson-Rubin and Lagrange Multiplier statistics. I take a similar approach here in combining the newly proposed tests with tests based on the sup-score statistic of Belloni et al. (2012),

$$S(\beta_0) := \sup_{\ell \in [d_z]} \left| \frac{\sum_{i=1}^n \epsilon_i(\beta_0) z_{\ell i}}{(\sum_{i=1}^n z_{\ell i}^2)^{1/2}} \right|. \tag{6}$$

which have correct asymptotic size even when the number of instruments is much larger than the sample size, $d_z \gg n$. A level $\alpha \in (0,1)$ test based on the sup-score statistic rejects whenever $S(\beta_0) > c^S_{1-\alpha}$ where, for $e_1, \ldots, e_n$ iid standard normal and generated independently of the data, $c^S_{1-\alpha}$ is the simulated multiplier bootstrap critical value:[7]

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

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 decides which test to run based on an attempt to detect whether the alternative β is such that $\mathbb{E}[\hat{\Pi}^I_{\ell,i}] = 0$ for all $i \in [n]$ and *some* $\ell \in [d_x]$, where $\hat{\Pi}^I_{\ell,i} := \sum_{j \neq i} h_{ij} r_{\ell j}$ denotes the infeasible version of the first-stage estimate that could be constructed if $\rho(\cdot)$ were known to the researcher. When this happens, tests based on the $JK(\beta_0)$ statistic have trivial power against deviations in the ℓ^{th} coordinate of β, so in local neighborhoods of these values of β, tests based on the sup-score statistic may be preferable. Detection of whether $\mathbb{E}[\hat{\Pi}^I_{\ell,i}] = 0$ is based on the conditioning statistic:

$$C = \inf_{\ell \in [d_x]} \sup_{i \in [n]} \left| \frac{\sum_{j \neq i} h_{ij} \hat{r}_{\ell j}}{(\sum_{j \neq i} h_{ij}^2)^{1/2}} \right|. \tag{7}$$

Under the assumption that $\mathbb{E}[\hat{\Pi}^I_i] = 0$ 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 $\{(e_{\ell 1}, \ldots, e_{\ell n}) : \ell \in [d_x]\}$ iid standard normal and generated independently of the data, and for any $\theta \in (0,1)$, define the conditional quantile

$$c^C_{1-\theta} := (1-\theta) \text{ quantile of } \inf_{\ell \in [d_x]} \sup_{i \in [n]} \left| \frac{\sum_{j \neq i} e_{\ell j} h_{ij} \hat{r}_{\ell j}}{(\sum_{j \neq i} h_{ij}^2)^{1/2}} \right| \text{ conditional on } \{(y_i, x_i, z_i)\}_{i=1}^n \tag{8}$$

The thresholding test decides which test to run by comparing the conditioning statistic C to a threshold value τ,

$$T(\beta_0; \tau) = \begin{cases} \mathbf{1}\{JK(\beta_0) > \chi^2_{d_x, 1-\alpha}\} & \text{if } C \ge \tau \\ \mathbf{1}\{S(\beta_0) > c^S_{1-\alpha}\} & \text{if } C < \tau \end{cases}. \tag{9}$$

In principle, the thresholding statistic has correct size for any (preset) choice of parameter τ. In practice, the calibrated simulations use $\tau = c^C_{0.3}$, while the supplementary simulations also report results with $\tau = c^C_{0.75}$.

[7] This conditional quantile can also be approximated using an empirical bootstrap procedure as demonstrated by Deng and Zhang (2020).

Table 1
Comparasions of first-stage F-statistics and 95% confidence interval coverage Probability using selected and oracle instruments.

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

3. Simulation evidence and empirical application

In empirical settings where the number of instruments is large, current practice is often to estimate the first stage with LASSO or post-LASSO and to gauge the strength of identification using the first-stage F-statistic calculated on the selected instruments. A leading example is Gilchrist and Sands (2016), who study social spillovers in movie consumption using $d_z = 48$ linearly independent nationally aggregated weather instruments with $n = 1{,}671$ opening weekend days and employ a post-LASSO estimate of the first stage (Belloni et al., 2012). This strategy is also shown to work well by Angrist and Frandsen (2022) in the returns-to-education data of Angrist and Krueger (1991), where $d_z = 180$ or $d_z = 1{,}530$ instruments are available with $n = 329{,}509$. This section uses these two settings to examine the finite-sample behavior of the tests proposed in this paper. Using a simple simulation experiment, I first demonstrate that first-stage F-statistics calculated on LASSO-selected variables can be misleading indicators of identification strength, so that identification-robust inference may be warranted even when the reported first-stage F-statistic is large. I then examine the size and power of the proposed tests in Monte Carlo designs calibrated to the Gilchrist and Sands (2016) data and apply the tests to the data of Angrist and Krueger (1991).

The calibrated designs hold the observed instrument matrix and a first stage estimated from the data fixed while growing the instrument dictionary from the initial 48 instruments ($d_z/n \approx 0.03$) to 1182 instruments ($d_z/n \approx 0.71$). Errors are simulated under two designs for the conditional covariance of the errors: a baseline design under which the nuisance parameter $\rho(z_i)$ is approximately sparse, in line with the conditions that motivate the estimation of $\rho(z_i)$ in Section 2, and a design with dense, nonlinear heteroskedasticity of the form $\sigma(z_i) \propto \exp(\gamma' z_i)$, under which $\rho(z_i)$ is itself a dense, nonlinear function of the instruments and the approximate sparsity conditions are purposefully strained. The test based on the $JK(\beta_0)$ statistic appears to have nearly exact size in every instrument set under both designs, without knowledge of the relevant asymptotic regime. Moreover, its power advantage over the "many-instrument" jackknife AR and jackknife LM tests grows with d_z/n, that is, in the settings where a researcher may wish to expand the instrument dictionary. A similar pattern appears in the data of Angrist and Krueger (1991), where confidence intervals obtained by inverting the jackknife K-test are about half the length of their many-instrument counterparts, or smaller. The original confidence-interval analysis of the Gilchrist and Sands (2016) data, which motivated these calibrated designs, is presented in Appendix L; a complementary simulation study on fully synthetic data is presented in Appendix M.

3.1. First-stage F-statistics on LASSO-selected instruments

To handle their 48 instruments, Gilchrist and Sands (2016) set the first-stage LASSO penalty so that the number of instruments selected is one, two, or three. The resulting first-stage F-statistics computed on the selected instrument(s), 38.80, 25.86, and 20.95, respectively, seem to indicate strong identification. However, the first-stage F-statistic on the full set of instruments is only 3.80. Since the LASSO objective is an ℓ_1-penalized version of the OLS loss, computing F-statistics on the variables selected by LASSO may mechanically lead to higher F-statistics even if the underlying relationship between the instruments and the endogenous variable is weak.

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

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

Given this lack of clarity on the strength of identification, I evaluate the proposed testing procedures in Monte Carlo designs calibrated to the Gilchrist and Sands (2016) data.

3.2. Empirically calibrated Monte Carlo design

The calibrated designs are built directly on the Gilchrist and Sands (2016) sample of $n = 1{,}671$ opening weekend days.[8] As in the original study, the outcome, the endogenous variable, and the instruments are residualized against a vector of date controls and a constant; the instrument columns are then standardized. Section G provides a formal treatment of the model with exogenous controls that justifies this partialling out. The residualized instrument matrix is held fixed across simulation replications, matching the conditional-on-instruments analysis in the rest of the paper; only the errors are redrawn. Data are generated according to the model

$$y_i = x_i\beta + \epsilon_i, \qquad x_i = \Pi_i + v_i, \tag{10}$$

with $\beta = 0$, and each test is evaluated for the hypothesis $H_0 : \beta = \beta_0$ over a grid of values $\beta_0 - \beta$.

The first stage is taken from the data: Π_i is the fitted value from an OLS regression of the residualized endogenous variable on the 48 initial instruments, rescaled once so that the expected first-stage F-statistic on the initial instrument set equals the value of 3.80 observed in the data. Importantly, Π_i is then held fixed as the instrument dictionary grows. Adding constructed instruments therefore adds no identifying information; it only dilutes the signal across a larger dictionary, mimicking the empirical practice of expanding a first-stage specification with technical terms. The expected first-stage F-statistic computed on the full tested dictionary falls from 3.80 to 0.25 to 0.05 across the instrument sets described below.

I consider three instrument sets. The first, initial, consists of the 48 initial weather instruments ($d_z/n \approx 0.03$). The second, expanded, adds the squares and cubes of all initial instruments and the pairwise interactions of the 28 instruments most correlated with the endogenous variable, so that $d_z = 522$ ($d_z/n \approx 0.31$). Because the weather instruments are shared by the three days of each opening weekend, any weather-only dictionary has rank at most 556, which bounds d_z/n by roughly one third and forces nearly uniform leverage. The third set, calendar, therefore interacts the 330 strongest terms of the expanded dictionary with Saturday and Sunday indicators, which vary within opening weekend, so that $d_z = 1{,}182$ ($d_z/n \approx 0.71$). With the calendar instrument set the design reaches the high-dimensional regime, with median hat-matrix leverage of 0.73 and maximal leverage near one.[9]

The error processes are calibrated to the empirical residuals. Writing u_{1i}, u_{2i} for independent standardized Laplace innovations,[10] errors are generated

$$\epsilon_i = \sigma_\epsilon(z_i)\,u_{1i}, \qquad v_i = \sigma_v(z_i)\big(\kappa(z_i)\,u_{1i} + \sqrt{1-\kappa(z_i)^2}\,u_{2i}\big),$$

and I consider two designs for the conditional variance surfaces $(\sigma_\epsilon(z_i), \sigma_v(z_i))$ and the conditional correlation $\kappa(z_i)$. In the first, approximately sparse design, the heteroskedasticity index is a quadratic in the three instruments most correlated with the endogenous variable, $\sigma_\epsilon(z_i) \propto 1 + 0.5\,(z_{1i}^2 + z_{2i}^2 + z_{2i}z_{3i})$ with $\sigma_v(z_i)$ analogous, and the conditional correlation varies linearly in z_{1i} alone, clipped to $[-0.95, 0.95]$. Under this design the nuisance parameter $\rho(z_i)$ is approximately sparse, in line with the conditions that motivate the LASSO estimate of Section 2. In the second, dense design, the conditional standard deviations take the exponential form $\sigma_\epsilon(z_i) \propto \exp(a\,\gamma_1' z_i/2)$ and $\sigma_v(z_i) \propto \exp(a\,\gamma_2' z_i/2)$ with $a = 0.5$, normalized so that the marginal variances match those of the empirical residuals. The loading vector γ_1 places nonzero weight on every column of the instrument dictionary, with weights proportional to each column's sample correlation with a residual-based volatility index, and γ_2 is a permutation of γ_1.[11] The conditional correlation is likewise dense, $\kappa(z_i) = \bar\kappa + 0.6 \cdot (\gamma_1' z_i + \gamma_2' z_i)/2$, clipped to $[-0.95, 0.95]$, where $\bar\kappa$ is the empirical residual correlation. Under the dense design the nuisance parameter $\rho(z_i)$ is a dense, highly nonlinear function of all of the instruments – the $\Omega(z) = \exp(\gamma' z)$ form of conditional variance – so the LASSO estimate $\hat\rho(z_i)$ can at best recover a sparse approximation. The dense design thus strains the approximate sparsity conditions underlying the estimation of $\rho(z_i)$ and provides a check on the robustness of the proposed tests to failures of exact sparsity.

I compare the jackknife K-test based on $JK(\beta_0)$, the thresholding test $T(\beta_0; \tau_{0.3})$ of Section 5.2 with cutoff set to the 30$^{\text{th}}$ quantile of the conditioning statistic, and the sup-score test $S(\beta_0)$ of Belloni et al. (2012) against the jackknife AR test (JAR) of Crudu et al. (2021) and Mikusheva and Sun (2021), the jackknife LM test (JLM) of Matsushita and Otsu (2022), cross-fit versions of both (JAR_{CF}, JLM_{CF}) that use the variance estimator of Mikusheva and Sun (2021), and the bootstrap ridge-regularized Anderson–Rubin test of Lim et al. (2024) (LWZ), with its ridge penalty selected by the authors' recommended diagnostics. The $JK(\beta_0)$ statistic is implemented with the true jackknife ridge estimates of (3). The nuisance parameter $\rho(z_i)$ is estimated by the ℓ_1-penalized regression (2) over the full tested dictionary with the plug-in penalty $\lambda \propto \sqrt{\log(d_b)/n}$, implemented via `glmnet` (Friedman et al., 2010). Critical values for the sup-score and conditioning statistics, and for the LWZ statistic, are simulated with 499 multiplier bootstrap draws. Size is computed from 1000 replications per instrument set and design, and power over 41-point grids of $\beta_0 - \beta$, with common random numbers across tests.

[8] Appendix L describes the setting, the outcome and endogenous variables, and the weather instruments in detail.

[9] A fourth, weather-only instrument set that uses all pairwise interactions and attains the 556 rank cap yields results nearly identical to the expanded set; it is reported in Appendix N.

[10] As in Appendix M, Laplace errors are used because the jackknife K-statistic has a nearly exact χ_1^2 distribution under jointly Gaussian errors; non-Gaussian errors provide a sharper check on the quality of the Gaussian approximation.

[11] Concretely, γ_1 is proportional to the vector of sample correlations between each standardized dictionary column and the index $\hat\epsilon_i^2 + \hat v_i^2 + |\hat\epsilon_i \hat v_i|$ formed from the calibration residuals, offset so that every entry is bounded away from zero and normalized to unit Euclidean length; γ_2 reverses the order of γ_1. The linear indices $\gamma_1' z_i$ and $\gamma_2' z_i$ are standardized before exponentiation.

Table 2
Simulated size at the 5% nominal level in the empirically calibrated designs, under the approximately sparse and dense specifications of the nuisance parameter $\rho(z_i)$. The design matrix and calibrated first stage are held fixed while the tested instrument dictionary grows. Each instrument set is simulated 1000 times per design; simulation standard errors are approximately 0.007.

	Sparse $\rho(z_i)$			Dense $\rho(z_i)$		
Testing Procedure	d_z=48	d_z=522	d_z=1,182	d_z=48	d_z=522	d_z=1,182
$JK(\beta_0)$	0.039	0.044	0.044	0.046	0.046	0.046
$T(\beta_0; \tau_{0.3})$	0.039	0.043	0.044	0.046	0.046	0.046
$S(\beta_0)$	0.040	0.012	0.003	0.029	0.010	0.002
JAR	0.051	0.053	0.051	0.049	0.048	0.049
JAR_{CF}	0.051	0.087	0.070	0.052	0.076	0.072
JLM	0.043	0.058	0.053	0.048	0.059	0.051
JLM_{CF}	0.043	0.069	0.051	0.049	0.059	0.045
LWZ	0.035	0.046	0.054	0.039	0.054	0.060

3.3. Size and power results

Table 2 reports simulated null rejection probabilities at the 5% nominal level under both error designs. The jackknife K-test has nearly exact size in every instrument set under both designs: its rejection rate stays between 0.039 and 0.046 whether the dictionary contains 48, 522, or 1182 instruments and whether $\rho(z_i)$ is approximately sparse or dense; the thresholding test inherits this behavior. The many-instrument tests appear less stable as the dimensionality regime changes: the cross-fit jackknife AR test overrejects with the expanded and calendar instrument sets under both designs (rejection rates between 0.070 and 0.087), and the plug-in JLM test drifts to 0.058–0.059 with the expanded set. The sup-score test becomes severely conservative as the dictionary grows, with null rejection falling to 0.003 and 0.002 with the calendar set, a consequence of its critical value growing with the number of instruments. The LWZ test controls size reasonably well, though its rejection rate drifts upward with the dictionary (0.054 and 0.060 with the calendar set). It is notable that moving from the sparse to the dense design has essentially no effect on the size of the jackknife K-test, even though under the dense design the LASSO estimate $\hat{\rho}(z_i)$ is necessarily only a sparse approximation to the true nuisance parameter. This size control may partly be an artifact of the similarity between the $JK(\beta_0)$ statistic and the jackknife LM statistic, however. To the extent that the LASSO penalty shrinks $\hat{\rho}(z_i)$ toward zero under the dense design, the estimated partialled-out variable $\hat{r}_i = x_i - \hat{\rho}(z_i)\epsilon_i(\beta_0)$ stays close to x_i, in which case the numerator of the $JK(\beta_0)$ statistic resembles that of the jackknife LM statistic of Matsushita and Otsu (2022), whose size control does not rely on estimation of $\rho(z_i)$.

To examine power, the critical value of each test is set to the simulated 95$^{\text{th}}$ quantile of the corresponding test statistic under H_0, so that differences in power cannot be driven by differences in size. I focus on the approximately sparse design, under which the maintained conditions of the formal analysis hold; power results under the dense design are similar and are reported in Appendix N. Figs. 1–3 plot the resulting calibrated power curves for the initial, expanded, and calendar instrument sets, and Table 3 summarizes all three instrument sets by the average calibrated rejection probability over the non-null grid points.

With the initial instrument set, where $d_z/n \approx 0.03$ and identification is weak but the dictionary is small, the jackknife K-test and the JLM test perform essentially equivalently (Fig. 1; average calibrated power 0.509 and 0.499), and both dominate the AR-type procedures. Little appears to be lost from using the jackknife K-test when the number of instruments is small. As the dictionary grows, however, the performance of the tests diverges. With the expanded instrument set the average calibrated power of the JLM test falls to 0.305 while that of the jackknife K-test rises to 0.562; with the calendar set the gap widens to 0.642 against 0.144, and at the endpoints of the $\beta_0 - \beta$ grid the jackknife K-test rejects with probability 0.914 and 0.965 while the JLM test rejects with probability below 0.27. The cross-fit variance estimator improves the power of the JLM test only modestly. This pattern is consistent with the discussion in Section 5: the deleted-diagonal first-stage estimates underlying the JAR and JLM statistics degrade as leverage grows, whereas the true jackknife ridge estimates remain informative. With the calendar instrument set the median leverage is 0.73, so deleting the diagonal discards most of each observation's first-stage information. Indeed, the correlation between the jackknife ridge first-stage estimates and the true first stage exceeds that of the deleted-diagonal estimates by 0.22 for the expanded instrument set and by 0.38 for the calendar instrument set.

Among the alternatives that allow $d_z \gg n$, the LWZ test appears to be the strongest competitor with the calendar instrument set (average calibrated power 0.360), consistent with its AR-type construction spreading power over many directions rather than concentrating it in the estimated first-stage direction; however, its power remains well below that of the jackknife K-test in all instrument sets. The thresholding test tracks the jackknife K-test closely in every instrument set, selecting the $JK(\beta_0)$ branch in the vast majority of replications, while retaining the sup-score branch as insurance against the distant alternatives discussed in Section 5. Appendix N reports two further exercises: results for the fourth, weather-only 556-instrument set and the power results under the dense design. The conclusions are unchanged in both cases; in particular, the power comparisons above are not sensitive to whether the nuisance parameter is sparse or dense.

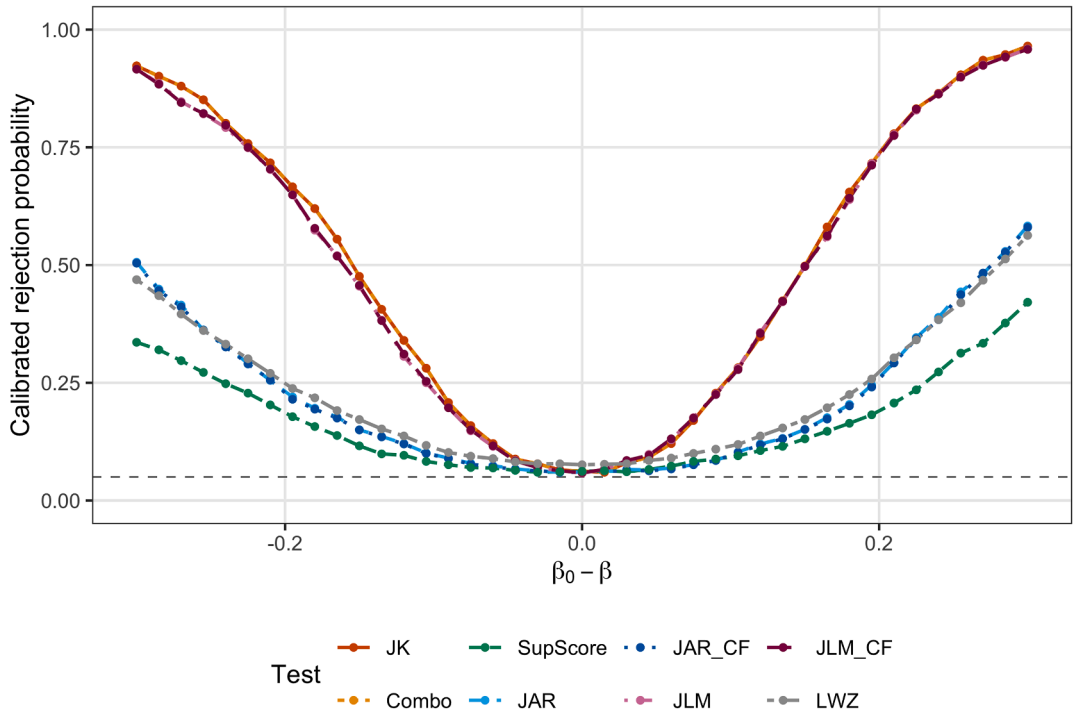


Fig. 1. Calibrated power curves for the initial instrument set ($d_z = 48$, $d_z/n \approx 0.03$) of the empirically calibrated design with an approximately sparse nuisance parameter. Rejection probabilities are computed over a 41-point grid of ($\beta_0 - \beta$) values with 1000 simulations per point.

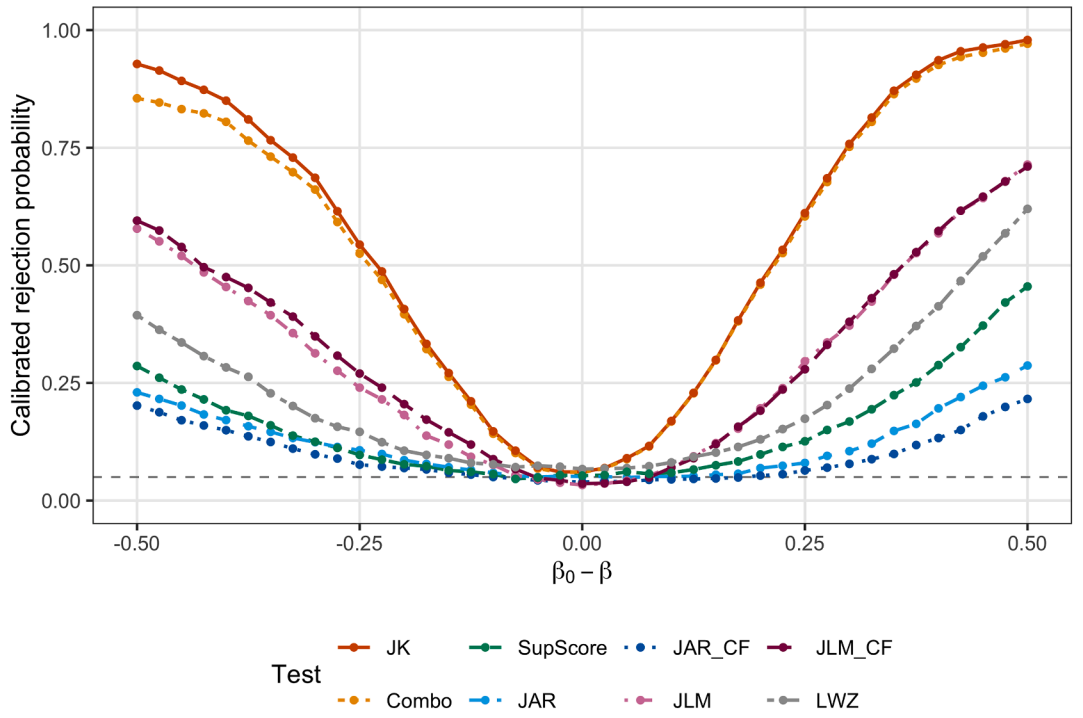


Fig. 2. Calibrated power curves for the expanded instrument set ($d_z = 522$, $d_z/n \approx 0.31$) of the empirically calibrated design with an approximately sparse nuisance parameter. Rejection probabilities are computed over a 41-point grid of ($\beta_0 - \beta$) values with 1000 simulations per point.

Table 3
Average calibrated rejection probability over the non-null grid points in the empirically calibrated design with an approximately sparse nuisance parameter, by instrument set. Power grids span $\beta_0 - \beta \in [-0.30, 0.30]$, $[-0.50, 0.50]$, and $[-0.80, 0.80]$ for the initial, expanded, and calendar instrument sets, respectively.

Testing Procedure	Initial $d_z = 48$ $d_z/n \approx 0.03$	Expanded $d_z = 522$ $d_z/n \approx 0.31$	Calendar $d_z = 1{,}182$ $d_z/n \approx 0.71$
$JK(\beta_0)$	0.509	0.562	0.642
$T(\beta_0; \tau_{0.3})$	0.509	0.546	0.615
$S(\beta_0)$	0.168	0.155	0.227
JAR	0.219	0.121	0.086
JAR_{CF}	0.217	0.096	0.087
JLM	0.499	0.305	0.144
JLM_{CF}	0.500	0.316	0.186
LWZ	0.228	0.218	0.360

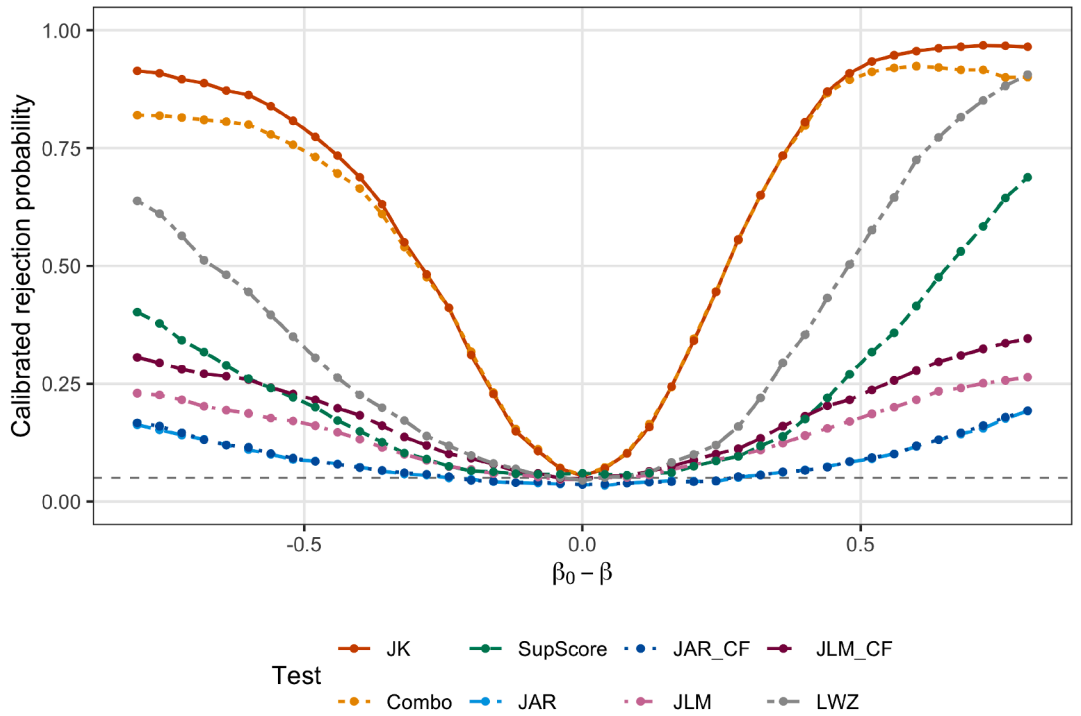


Fig. 3. Calibrated power curves for the calendar instrument set ($d_z = 1{,}182$, $d_z/n \approx 0.71$) of the empirically calibrated design with an approximately sparse nuisance parameter. Rejection probabilities are computed over a 41-point grid of $(\beta_0 - \beta)$ values with 1000 simulations per point.

3.4. Application to returns to education

For additional comparison, I revisit the setting of Angrist and Krueger (1991), who study the effect of education on log-wages, instrumenting for education using various combinations of quarter-of-birth (QOB), year-of-birth (YOB), and place-of-birth indicators (POB). This dataset has the benefit of being used in the empirical studies of both Mikusheva and Sun (2021) and Matsushita and Otsu (2022), facilitating an easy comparison of the performance of the newly proposed testing procedure to these existing "many-instrument" methods. Moreover, Angrist and Frandsen (2022) report reduced bias compared to 2SLS in this dataset when using post-LASSO estimates in the first stage, suggesting that high-dimensional or machine-learning techniques can be useful in this setting.

Table 4 displays the resulting identification robust confidence intervals using two set of instruments. The first set, which contains all QOB × YOB and all QOB × POB interactions for $d_z = 180$, corresponds to the specification in Table VII of Angrist and Krueger (1991). The second set of instruments, $d_z = 1{,}530$, contains all QOB × YOB × POB interactions. The confidence intervals for the jackknife Anderson-Rubin (JAR) and jackknife Lagrange-Multiplier (JLM) tests are taken directly from Mikusheva and Sun (2021) and Matsushita and Otsu (2022), respectively. To implement the jackknife K-test I estimate $\rho(z_i)$ as in (2) and set $\lambda = 0$ when constructing

Table 4
95% Confidence Intervals and Interval Lengths in the data of Angrist and Krueger (1991).

Number of Instruments	JAR	JLM	$JK(\beta_0)$	$S(\beta_0)$
Initial Instrument Set (d_z=180)	⟵ 0.193 ⟶ [0.008, 0.201]	⟵ 0.066 ⟶ [0.067, 0.133]	⟵ 0.034 ⟶ [0.067, 0.101]	∅
All Interactions (d_z=1,530)	⟵ 0.249 ⟶ [−0.047, 0.202]	⟵ 0.098 ⟶ [0.025, 0.123]	⟵ 0.034 ⟶ [0.008, 0.042]	∅

the $\hat{\Pi}_i$ values. However, for computational reasons I opt to split the data into 11 pieces when estimating $\hat{\Pi}_i$ rather than using a true "leave-one-out" jackknife approach. Confidence intervals from the combination test are not reported as the sup-score confidence interval is empty in both specifications.

The results in Table 4 are in line with those of the calibrated simulation study. In both specifications the confidence intervals obtained from inverting the jackknife K-test are substantially smaller than those obtained by inverting the "many-instrument" JAR and JLM tests. Indeed, the confidence intervals obtained from inverting the jackknife K-statistic are nearly half the length of those obtained from inverting the JLM test and less than a quarter the length of those obtained from inverting the JAR test. It is notable that the range of plausible returns to education values implied by the jackknife K-confidence interval with $d_z = 1{,}530$ are strictly below that implied by the jackknife-K confidence interval with $d_z = 180$. This may be related to misspecification of the linear model as indicated by the empty sup-score confidence interval. "Downward bias" of the confidence intervals when using the larger instrument set is also seen for the JAR and JLM confidence intervals, though to a lesser extent. Remark 5.3, below, provides reasoning for why confidence intervals based on the $JK(\beta_0)$ statistic may be tighter than those based on the JAR and JLM statistics.

Both here and throughout this paper, my results should not be interpreted as a critique of Belloni et al. (2012), Mikusheva and Sun (2021), or Matsushita and Otsu (2022), whose prior work I relied upon and was inspired by.

4. Limiting behavior of the test statistic

In this section, I establish the validity of tests based on the $JK(\beta_0)$ statistic. The limiting behavior of the test statistic is analyzed via a direct Gaussian approximation technique. I describe the approach and characterize the limiting behavior of the test statistic under the null and in local neighborhoods of the null. 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 first stage estimates. When there is only a single endogenous variable, $d_x = 1$, the argument can be considerably simplified. The extension to $d_x > 1$ uses a more involved argument which relies on stronger moment conditions.

Formally, I show that quantiles of the jackknife K-statistic can be approximated by analogous quantiles of the Gaussian statistic:

$$JK_G(\beta_0) := \tilde{\epsilon}(\beta_0)\tilde{\Pi}(\tilde{\Pi}_\epsilon'\tilde{\Pi}_\epsilon)^{-1}\tilde{\Pi}'\tilde{\epsilon}(\beta_0); \tag{11}$$

where, for each $i \in [n]$, $(\tilde{\epsilon}_i(\beta_0), \tilde{r}_i')'$ are generated independently across indices following a Gaussian distribution with the same mean and covariance matrix as $(\epsilon_i(\beta_0), r_i')'$. Further, for $\tilde{\Pi}_{\ell i} = \sum_{j\neq i} h_{ij}\tilde{r}_{\ell j}$ define $\tilde{\Pi}_i := (\tilde{\Pi}_{1i}, \ldots, \tilde{\Pi}_{d_x i})' \in \mathbb{R}^{d_x}$, $\tilde{\Pi}_{\epsilon i} := (\mathbb{E}[\epsilon_i^2(\beta_0)])^{1/2}\tilde{\Pi}_i$,

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

As uncorrelated jointly Gaussian 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 $\chi^2_{d_x}$.

To outline the argument, consider functions $\varphi_\gamma(\cdot) \in C_b^3(\mathbb{R})$ that approximate the indicators $\mathbf{1}\{\cdot \leq a\}$, where $a \in \mathbb{R}$ is arbitrary and as $\gamma \to 0$ the quality of the approximation improves but the derivatives of φ_γ become larger in magnitutde. A primary goal is to show, for a sequence γ_n tending to zero, that

$$\mathbb{E}[\varphi_{\gamma_n}(JK_I(\beta_0)) - \varphi_{\gamma_n}(JK_G(\beta_0))] \to 0 \tag{12}$$

for a version of the test statistic, $JK_I(\beta_0)$, that could be constructed if $\rho(\cdot)$ was known to the researcher. In order to establish (12), existing interpolation methods cannot be applied as they require the derivative of the test statistic with respect to individual obserations are bounded. In this setting, the derivative of the test statistics with respect to terms in the denominator matrix, $\hat{\Pi}_\epsilon'\hat{\Pi}_\epsilon$, may be as large as the inverse of the minimum eigenvalue of the denominator matrix. When identification is sufficiently weak, the eigenvalues of the denominator matrix can be arbitrarily close to zero and thus inverse of its minimum eigenvalue may not even 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.[12] Geometrically, this approach can be thought of as "stretching out" the function $\varphi_{\gamma_n}(\cdot)$ in directions where the minimum eigenvalue of the denominator matrix is close to zero. Through the chain rule, this allows for control of the overall derivative of $\varphi_{\gamma_n}(JK_I(\beta_0))$ with respective to an individual observation.

4.1. High level assumptions

I now detail the assumptions needed for the argument, starting with the assumptions that are commmon to the cases with $d_x = 1$ and $d_x > 1$. 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 4.1 (Balanced Design). *(i) Let $s_{\ell,n}^{-2} = \max_{1\le i\le n} \mathbb{E}[(\hat{\Pi}^I_{\ell i})^2]$ for each $\ell \in [d_x]$; then, the minimum eigenvalue of the following matrix is bounded away from zero:*

$$c^{-1} \le \lambda_{\min}\mathbb{E}\Big(\tfrac{s_{\ell,n}s_{k,n}}{n} \textstyle\sum_{i=1}^n (\hat{\Pi}^I_{\ell i})(\hat{\Pi}^I_{ki}) \Big)_{\substack{1\le \ell\le d_x \\ 1\le k\le d_x}}$$

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

Assumption 4.1(i) requires that the average second moment of the infeasible first-stage estimators be of 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 4.1 provides further intuition for this assumption and below discusses how this assumption and Assumption 4.1(ii) may be verified in practice. Remark 4.2 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 4.1(ii) requires that the maximum leverage of any observation be bounded. When H is symmetric, it is automatically satisfied.[13] Assumption 4.1(iii) can be viewed as a mild technical requirement that there be more than one "effective" instrument in the hat matrix.[14] This condition can be easily verified in practice by examining the eigenvalues of HH'.

Next, I make a high level assumption that the estimation error in $\hat{\rho}(\cdot)$ can be treated as negligible in both the numerator and denominator. Later, I will verify this assumption for the particular choice of $\hat{\rho}(\cdot)$ described in Section 2. For each $\ell \in [d_x]$ define $\hat{\Pi}^I_{\ell,i} := \sum_{j\neq i} h_{ij} r_{\ell j}$, the version of the first stage estimates that could be constructed if $\rho(\cdot)$ was known to the researcher. Using these, define the magnitude of estimation error in the numerator and denominator as

$$\Delta_N := \max_{\ell\in[d_x]} \Big| \frac{s_{\ell,n}}{\sqrt{n}} \sum_{i=1}^n \epsilon_i(\beta_0)\big(\hat{\Pi}_{\ell,i} - \hat{\Pi}^I_{\ell,i}\big)\Big|$$

$$\Delta_D := \max_{\ell\in[d_x]} \frac{s_{\ell,n}^2}{n} \sum_{i=1}^n \epsilon_i^2(\beta_0)\big(\hat{\Pi}_{\ell,i} - \hat{\Pi}^I_{\ell,i}\big)^2$$

Assumption 4.2 (Estimation Error). *Estimation error in both the numerator and denominator of the test statistic can be treated as negligible, $(\Delta_N, \Delta_D) \to_p 0$.*

Showing that $(\Delta_N, \Delta_D) \to_p 0$ implies that estimation error can be treated as negligible for the test statistic $JK(\beta_0)$ requires some care. In a standard approach, this would straightforwardly follow from application of the continous mapping theorem. However, this approach requires that the scaled numerator and denominator each have well defined distributional limits, something that is not required by the direct Gaussian approximation. Instead, I establish and make use of anticoncentration bounds to show that the basic results of the continuous mapping theorem still apply even when the numerator and denominator do not have weak limits.

Finally, 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 $\mathbb{E}[r_{\ell i}]$.

Assumption 4.3 (Local Identification). *(i) The local power index is bounded $P \le c$ for*

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

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

[12] 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.

[13] To see this for $d_x = 1$ notice that $s_n^{-2} = \max_i \mathbb{E}[(\hat{\Pi}^I_i)^2] \ge \max_i Var(\hat{\Pi}^I_i) = \max_i \sum_{j\neq i} h_{ij}^2 Var(r_j)$, while $Var(r_j)$ will be assumed 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 $i \in [n]$.

[14] In the case of a standard projection matrix (no deleted diagonal), Assumption 4.1(iii) would be satisfied whenever $\text{rank}(z(z'z)^{-1}z) > 1$, which occurs whenever there are at least two linearly independent instruments.

Under H_0, Assumption 4.3 is trivially satisfied since $(\beta - \beta_0) = 0$ and $\sum_{j \neq i} s^2_{\ell,n} h^2_{ji} \leq c$ for each $\ell \in [d_x]$. The local power index is the second moment of the scaled numerator of the test statistic and is a measure of the association between the true first stage Π_i and the first-stage estimates $\widehat{\Pi}_i$. In Section 5, I discuss how the strength of this association is related to the power of the test under local alternatives. Assumption 4.3(ii) can be roughly interpreted as requiring the local neighborhoods of H_0 considered to be those in which the means of $(\epsilon_1(\beta_0), \dots, \epsilon_n(\beta_0))$ are of the same or lesser order than the means of $(r_1, \dots, r_n)$.

4.2. Limiting behavior of the test statistic

Under these assumptions I present results showing that, in local neighborhoods of H_0, the distribution of the test statistic can be uniformly approximated by the distribution of the Gaussian statistic described in (11). As mentioned, the argument can be simplified to require lighter moment bounds when $d_x = 1$. These moment bounds will be made on the model primitives, $\eta_i := (\beta - \beta_0)v_i + \epsilon_i = \epsilon_i(\beta_0) - \mathbb{E}[\epsilon_i(\beta_0)]$ and $\zeta_{\ell i} := v_{\ell i} - \rho_\ell(z_i)\eta_i = r_{\ell i} - \mathbb{E}[r_{\ell i}]$ for each $\ell \in [d_x]$.

Theorem 4.1 (Single Endogenous Variable). *Suppose that Assumptions 4.1, 4.2, 4.3 hold. In addition suppose that (i)* $\{|\Pi_i| + |\rho(z_i)| + |(\beta - \beta_0)|\} \leq c$ *and (ii) for any* $r, s \in \mathbb{Z}_+$ *satisfying* $r + s \leq 6$ $c^{-1} \leq \mathbb{E}[|\eta_i|^r |\zeta_i|^s] \leq c$. *Then, for* $d_x = 1$,

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

In particular, under H_0, $JK(\beta_0) \rightsquigarrow \chi^2_1$.

In the case of $d_x = 1$, 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 4.3(ii) holds.

Theorem 4.2 (Consistency). *Suppose that Assumptions 4.1, 4.2, and 4.3(ii) hold along with the moment conditions of Theorem 4.1. Then, if* $P \to \infty$ *the test based on* $JK(\beta_0)$ *is consistent; i.e for any fixed* $a \in \mathbb{R}$, $\Pr(JK(\beta_0) \leq a) \to 0$.

The dependence of the consistency result on Assumption 4.3(ii) is a nontrivial restriction because of the bias taken on in constructing r_i. In particular, if errors are homoskedastic, against certain alternatives it is possible that $\mathbb{E}[\widehat{\Pi}^I_i] = 0$ for all $i \in [n]$ even under strong identification (see Section 5.1). This is an extreme case, however. In general, bias in $\mathbb{E}[r_i]$ does not imply a violation of Assumption 4.3(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 4.5, Theorem 4.2 does not necessarily rule out consistency when $P \to \infty$ but Assumption 4.3(ii) fails.

Regardless, bias taken on in constructing r_i has consequences for the power of the test in finite samples. The combination with the sup-score test described in Section 2 is an attempt to rectify this. While this attempt is not perfect, it appears to work well both in the empirically calibrated simulations of Section 3 and in the empirical applications to the data of Angrist and Krueger (1991) (Section 3.4) and of Gilchrist and Sands (2016) (Appendix L).

The argument when $d_x > 1$ is considerably more involved than the case where $d_x = 1$ and requires strengthened moment condition on the variables η_i and ζ_i. Given a random variable X and $\upsilon > 0$ the Orlicz (quasi-)norm is defined

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

Random variables with a finite Orlicz norm for some $\upsilon \in (0, 1] \cup \{2\}$ are termed α-sub-exponential random variables (Gotze et al., 2021; Sambale, 2022). This class of encompasses a wide range of potential distributions including all bounded and sub-Gaussian random random variables (with $\upsilon = 2$), all sub-exponential random variables such as Poisson or noncentral χ^2 random variables (with $\upsilon = 1$), as well as random variables with "fatter" tails such as Weibull distributed random variables with shape parameter $\upsilon \in (0, 1]$. Thus, while imposing that the variables η_i and ζ_i are α-sub-exponential is notably stronger than the finite sixth moments required by Theorem 4.1, it may still be plausible in a wide range of empirical settings.

Theorem 4.3 (Multiple Endogenous Variables). *Suppose that Assumptions 4.1, 4.2, 4.3 hold. In addition suppose that (i)* $c^{-1} \leq \lambda_{\min}(\mathbb{E}[\eta_i \eta_i']) \leq \lambda_{\max}(\mathbb{E}[\eta_i \eta_i']) \leq c$ *and (ii) for some* $\nu \in (0, 1] \cup \{2\}$ *both* $\|\eta_i\|_{\psi_\nu} \leq c$ *and* $\|\zeta_{\ell i}\|_{\psi_\nu} \leq c$. *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, $JK(\beta_0) \rightsquigarrow \chi^2_{d_x}$.

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 Section 5, 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).

4.3. Controlling estimation error

The final step is to show that estimation error in $\hat{\rho}$ can be treated as negligible, that is verify Assumption 4.2. Establishing that this holds even when identification is weak makes use of the fact that estimation error in $\hat{\rho}(\cdot)$ enters the test statistic only through its

interaction with the implied error $\epsilon_i(\beta_0)$. When identification is weak, the implied error is nearly conditional mean independent of the instruments and thus the product, $\epsilon_i(\beta_0)\hat{\rho}(\cdot)$, is insensitive to small pertubations in $\hat{\rho}(\cdot)$.[15]

In principle, this orthogonality could be combined with cross-fitting to allow for the use of other machine learning methods to estimate $\hat{\rho}(\cdot)$, as in Chernozhukov et al. (2018). This possibility is explored briefly in Section H. The use of other machine learning methods may be useful if sparsity of $\rho(\cdot)$ is not a plausible assumption in the researcher's empirical setting since estimators such as random forests and neural networks can be consistent in high dimensional settings under alternate assumptions.[16] However, I focus on the ℓ_1-penalized procedure proposed in Section 2 both for expositional simplicity and because the sparsity assumption required for the consistency of this procedure mirrors that needed for the popular post-Lasso first-stage estimator.

Assumption 4.4 (Estimation Error). *For each $\ell \in [d_x]$ (i) there is a fixed constant $\upsilon \in (0,1] \cup \{2\}$ such that $\|\eta_i\|_{\psi_\upsilon} \leq c$; (ii) the basis terms $b(z_i)$ are bounded, $\|b(z_i)\|_\infty \leq c$; (iii) the approximation error satisfies $(\mathbb{E}_n[\xi_{\ell i}^2])^{1/2} = o(n^{-1/2})$; (iv) the researcher has access to an estimator $\hat{\phi}$ of ϕ satisfying $\log(d_b n)^{2/(\upsilon \wedge 1)}\|\hat{\phi}_\ell - \phi_\ell\|_1 \to_p 0$; (v) the following moment bounds hold*

$$\textit{(va)}\ \max_{1\leq l\leq d_b}\left|\mathbb{E}\left[\frac{s_n}{\sqrt{n}}\sum_{i=1}^n\sum_{j\neq i}h_{ij}\epsilon_i(\beta_0)b_l(z_j)\epsilon_j(\beta_0)\right]\right| \leq c$$
$$\textit{(vb)}\ \max_{\substack{1\leq i\leq n\\ 1\leq l\leq d_b}}|\mathbb{E}[s_n\textstyle\sum_{j\neq i}h_{ij}b_l(z_j)\epsilon_j(\beta_0)]| \leq c.$$

Assumption 4.2(i) ensures that η_i has finite exponential moments (i.e has a well defined moment generating function), which is required to allow the number of basis terms used to approximate $\rho(\cdot)$ to grow at a near exponential rate compared to the sample size ($d_b \gg n$). When using fewer basis terms this assumption may be relaxed. Assumption 4.2(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 4.2(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 4.2(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) or post-LASSO procedures based on refitting an unpenalized version of (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) 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 4.2(v) is a strengthening of the definition of local neighborhoods and can be interpreted similarly to Assumption 4.3(ii). Since the moment conditions in Assumption 4.2(va,vb) hold with $b_\ell(z_j)\epsilon_j(\beta_0)$ replaced with r_j, Assumption 4.2(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 $i = 1,\dots,n$ and $\ell = 1,\dots,d_b$. As with Assumption 4.3(ii), it is trivially satisfied under H_0 or, using the fact that $\max_i\sum_{j\neq i}s_n^2h_{ij}^2 \leq c$, whenever $\mathbb{E}[\epsilon_i(\beta_0)] = \Pi_i(\beta-\beta_0)$ is in a $\sqrt{n}$-neighborhood of zero.

Theorem 4.4 (Estimation Error). *Suppose that Assumptions 4.1 and 4.4 hold. Then $(\Delta_N, \Delta_D) \to_p 0$.*

Theorem 4.4 verifies Assumption 4.2 for the ℓ_1-penalized estimator proposed in Section 2. Thus, under Assumptions 4.1 and 4.4(i)-(iv), the limiting null distribution of $JK(\beta_0)$ is $\chi^2_{d_x}$.

Remark 4.1. When $d_x = 1$, a sufficient condition for Assumption 4.1(i) is that there is some fixed quantile $q \in (0,100)$ such that $(cq)^{-1} \leq \frac{q^{\text{th}}\text{-quantile of } \mathbb{E}[(\hat{\Pi}_i^I)^2]}{\max_i \mathbb{E}[(\hat{\Pi}_i^I)^2]}$. In practice this can be verified by checking that there is some quantile q such that both

$$\frac{q^{\text{th}}\text{-quantile of } \sum_{j\neq i}h_{ij}^2}{\max_i\sum_{j\neq i}h_{ij}^2} \quad\text{and}\quad \frac{q^{\text{th}}\text{-quantile of } (\sum_{j\neq i}h_{ij}\hat{r}_j)^2}{\max_i(\sum_{j\neq i}h_{ij}\hat{r}_j)^2} \tag{13}$$

are bounded away from zero. Similarly, Assumption 4.1(ii) can be verified by checking that

$$\frac{\max_i\sum_{j\neq i}h_{ji}^2}{\max_i\sum_{j\neq i}h_{ij}^2}$$

is bounded from above. 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 ($\Pi_i \lesssim n^{-1/2}$).

Remark 4.2. The balanced-design condition in Assumption 4.1(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

[15] In the langauge of Chernozhukov et al. (2018), this is termed "Neyman Orthogonality." Under strong identification, Neyman orthogonality allows a large range of machine learning techniques to be used in estimating the first stage. This (approximate) orthogonality also holds under structural parameter sequences that are local to the null.

[16] Chi et al. (2022) provide results for random forests under a "sufficient impurity decrease" assumption while Farrell et al. (2021) and Schmidt-Hieber (2020) provide results for neural networks under a generalized heirarchical model assumption.

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 $j \neq i$.[17] This would not violate Assumption 4.1 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.

Remark 4.3. The modified Lindeberg interpolation method allows me to give a nearly uniform explicit bound on the Gaussian approximation error in the case where $d_x = 1$. In particular, 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) and $JK_I(\beta_0)$ is the version of the test statistic that could be constructed if $\rho(\cdot)$ was known to the researcher. While it does not account for estimation error in $\hat{\rho}(\cdot)$, obtaining an explicit bound reflects an improvement over the original analyzes 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 4.4. 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 the OLS projection matrix, $H = z(z'z)^{-1}z$, and the number of instruments satisfies $d_z^3/n \to 0$. See Section Appendices A, and G for details.

Remark 4.5. Theorem 4.2 does not necessarily rule out that a test based on $JK_I(\beta_0)$ is consistent when $P \to \infty$ but Assumption 4.3(ii) fails to hold. 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 4.6. 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 $\rho(z_i)$, then the OLS estimate satisfies $\|\hat{\phi} - \phi\|_1 \to 0$ under $d_z^2/n \to 0$ even if ϕ is fully dense. This requirement is substantially weaker than the $d_z^3/n \to 0$ requirement of the standard K-statistic. For example, d_z^2/n may be plausibly satisfied in the setting of Paravisini et al. (2014), $d_z^2 = 100$ and $n = 5{,}995$, whereas $d_z^3/n \to 0$ may be questionable.

5. Power Properties and Improvements

Using the characterization of the limiting behavior of the test statistic derived in Section 4, 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). I describe how the combination with the sup-score statistic, described in Section 2, attempts to remedy this and formally establish its validity. Finally, I compare the $JK(\beta_0)$ statistic to statistics used in the many-instrument literature and provide reasoning for the power improvements seen in Section 3.

5.1. Local power properties

For exposition, I focus on the case where $d_x = 1$. In local neighborhoods of H_0, as defined in Assumptions 4.3 and 4.2, Theorem 4.1 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 $\mu(\tilde{r})$, $JK_G(\beta_0)|\tilde{r} \sim A^2(\tilde{r}) \cdot \chi_1^2(\mu(\tilde{r}))$:

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

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

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

$$\mu_\infty^2(\tilde{r}) = (\beta - \beta_0)^2 \frac{\left(\sum_{i=1}^n \Pi_i\tilde{\Pi}_i\right)^2}{\sum_{i=1}^n Var(\eta_i)\tilde{\Pi}_i^2}. \tag{14}$$

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

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

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

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

This deficiency is inherited from the similarity of the $JK(\beta_0)$ statistic to the K-statistic. As pointed out by Moreira (2001), this need not be an issue as long as there is a fixed constant $C \neq 0$ such that $\mathbb{E}[r_i] = C\Pi_i$ for all $i \in [n]$. However, in general, this will introduce bias into the first-stage estimates $\hat{\Pi}_i$ 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.[18]

5.2. A simple combination test

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

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 (15) is similar in spirit to the Anderson-Rubin test but has correct asymptotic size even when d_z grows near exponentially as a function of the sample size.

$$S(\beta_0) := \sup_{1\leq \ell \leq d_z} \left| \frac{\sum_{i=1}^n \epsilon_i(\beta_0) z_{\ell i}}{(\sum_{i=1}^n z_{\ell i}^2)^{1/2}} \right| \tag{15}$$

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

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

As with the Anderson-Rubin test, tests based on the sup-score statistic may have suboptimal power properties in overidentified models as 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 $\mathbb{E}[\hat{\Pi}^I_{\ell,i}] = 0$ for all $i = 1, \ldots, n$ and some $\ell \in [d_x]$. When this is the case, the researcher would prefer to test the null hypothesis using the sup-score statistic. As mentioned in Section 2, detection of whether $\mathbb{E}[\hat{\Pi}^I_{\ell,i}] = 0$ for some $i \in [n]$ is based on the conditioning statistic:

$$C = \inf_{\ell \in [d_x]} \sup_{i \in [n]} \left| \frac{\sum_{j \neq i} h_{ij} \hat{r}_j}{(\sum_{j \neq i} h_{ij}^2)^{1/2}} \right|. \tag{16}$$

Under the assumption that the first-stage estimates do not contain any information about the true first-stage, $\mathbb{E}[\hat{\Pi}^I_i] = 0$ 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 $\{(e_{\ell 1}, \ldots, e_{\ell n}) : \ell \in [d_x]\}$ iid standard normal and generated independently of the data, and for any $\theta \in (0,1)$, define the conditional quantile

$$c^C_{1-\theta} := (1-\theta) \text{ quantile of } \inf_{\ell \in [d_x]} \sup_{i \in [n]} \left| \frac{\sum_{j \neq i} e_{\ell j} h_{ij} \hat{r}_{\ell j}}{(\sum_{j \neq i} h_{ij}^2)^{1/2}} \right| \text{ conditional on } \{(y_i, x_i, z_i)\}_{i=1}^n \tag{17}$$

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

$$T(\beta_0; \tau) = \begin{cases} \mathbf{1}\{JK(\beta_0) > \chi^2_{1;1-\alpha}\} & \text{if } C \geq \tau \\ \mathbf{1}\{S(\beta_0) > c^S_{1-\alpha}\} & \text{if } C < \tau \end{cases} \tag{18}$$

for some cutoff τ. The calibrated simulations use the 30^{th} quantile of the distribution of C under the assumption that $\mathbb{E}[\hat{\Pi}^I_i] = 0, \forall i \in [n]$, while the supplementary simulations also report results using the 75^{th} quantile.

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

To show that the thresholding test is correctly sized, I compare the rejection probability to that of a Gaussian analog. In addition to $JK_G(\beta_0)$, defined in Section 4, define the Gaussian analogs of $S(\beta_0)$ and the conditioning statistic C:

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

where, as in Section 4, $(\tilde{e}_i(\beta_0), \tilde{r}_i)'$ 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 $Cov(\tilde{e}_i(\beta_0), \tilde{r}_i) = 0$ 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 $(\tilde{r}_1, \dots, r_n)$; 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 (18).

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

Assumption 5.1(i) is a strengthening of the moment bound on r_i similar to that of Assumption 4.2(i). As discussed, while more restrictive than the condition in Theorem 4.1, this still allows for a wide range of potential distributions for r_i. Assumption 5.1(ii) requires that the number of observations used to test $\mathbb{E}[\hat{\Pi}_i] = 0$ 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 5.1(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 5.1. *Suppose Assumptions 4.1, 4.3, 4.4, and 5.1 hold along with the additional moment bounds of Theorem 4.3. Then,*

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

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

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

The first part of Theorem 5.1 establishes the asymptotic validity of the thresholding test $T(\beta_0; \tau)$ for any choice of cutoff τ. While not explicitly stated in the statement of the theorem, this result is uniform in the choice of τ; for any sequence $\{\tau_n\} \subset \mathbb{R}_+$ the sequence of testing procedures $T(\beta_0; \tau_n)$ will also have asymptotic size α. The proof of this statement follows the logic outlined above. The second part of Theorem 5.1 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 be treated as negligible under Assumption 4.4.

A related question is whether the $JK(\beta_0)$ statistic itself admits bootstrap calibration, as do the Anderson-Rubin-type statistics of Lim et al. (2024). Because the $JK(\beta_0)$ statistic is asymptotically pivotal, with a limiting $\chi^2_{d_x}$ distribution that does not depend on unknown features of the data generating process, resampling is not needed for the validity of its critical value. Nevertheless, a multiplier bootstrap version of the statistic could be constructed in the same manner as the critical values above, and the direct Gaussian couplings developed in this paper are the type of strong approximation used by Lim et al. (2024) to establish the validity of their bootstrap procedure. I conjecture that such a scheme would be valid, and possibly deliver finite-sample refinements over the $\chi^2_{d_x}$ critical value, but leave its formal analysis to future work.

In the case of a single endogenous variable, $d_x = 1$, Theorem 5.1 could be established under the lighter conditions of Theorem 4.1 along with Assumption 5.1. However, for brevity, I do not seperate the two cases here.

Remark 5.1. As mentioned by Andrews (2016) in the context of the standard K-statistic, the 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 $\mathbb{E}[\hat{\Pi}_i^I] = \sum_{j \neq i} h_{ij} \{\Pi_j - \rho(z_j)\Pi_j(\beta - \beta_0)\}$ to be related to the true first stage Π_i and for the test to have nontrivial power. Moreover, in light of the dependence of the consistency result in Theorem 4.2 on Assumption 4.3(ii), in the case where $\mathbb{E}[\hat{\Pi}_i] = 0$ for all $i \in [n]$ it may be particularly important to avoid using the jackknife K-statistic to test H_0.

5.3. Comparison to many-instrument procedures

It is useful to compare the $JK(\beta_0)$ statistic to the JLM statistic of Matsushita and Otsu (2022), which also converges to a limiting χ^2 distribution when $d_z \to \infty$. In the case where $d_x = 1$ the JLM statistic can be expressed

$$\text{JLM}(\beta_0) := \frac{\left(\sum_{i=1}^n \epsilon_i(\beta_0)\sum_{j\neq i} P_{ij}x_j\right)^2}{\sum_{i=1}^n \epsilon_i^2(\beta_0)\left(\sum_{j\neq i} P_{ij}x_j\right)^2 + \sum_{i=1}^n \sum_{j\neq i} P_{ij}^2\epsilon_i(\beta_0)\epsilon_j(\beta_0)x_i x_j} \tag{19}$$

where $P = z(z'z)^{-1}z'$ is the standard OLS projection matrix and $P_{ij} = [P]_{ij}$ denotes its ij^{th} element. This expression looks similar to that of the $JK(\beta_0)$ statistic with first stage estimates $\hat{\Pi}_i = \sum_{j\neq i} P_{ij}x_j$. From this, one can posit two main reasons for the increased power of tests based on the $JK(\beta_0)$ statistic seen in the calibrated simulations of Section 3 and the empirical applications of Section 3.4 and Appendix L.

The first is that, when the bias of r_i is not too adverse, first stage estimates based on the "true" jackknife ridge or jackknife OLS described in (3) may be closer to the true first stage, Π_i than those based on the deleted-diagonal projection matrix. As seen in Section 5.1, higher quality first stage estimates can improve the power of the test by increasing the correlation between these estimates and $\epsilon_i(\beta_0)$ under H_1. This loss in quality of first-stage estimates based on the deleted-diagonal projection matrix may be negligible when the diagonal elements, P_{ii}, are small in which case the deleted diagonal estimates, $\hat{\Pi}_i = \sum_{j\neq i} P_{ij}x_j$, closely resemble standard OLS estimates. However, when either the number of instruments is large relative to the sample size or the instruments are highly correlated the diagonal elements P_{ii} will be large in which case estimates of Π_i based on the deleted-diagonal projection matrix may not be accurate. This pattern can be seen in both empirical applications; in the data of Angrist and Krueger (1991) (Section 3.4) and of Gilchrist and Sands (2016) (Appendix L) the improvements in power from using the $JK(\beta_0)$ statistic become more pronounced as the number of instruments increases, as they do across the instrument sets of the calibrated simulations in Section 3.3.

A second potential reason for improved power is that the $JK(\beta_0)$ statistic uses individual scores, $\epsilon_i(\beta_0)\hat{\Pi}_i$, that are uncorrelated with each other under H_0. That is, for $j \neq i$, $\mathbb{E}[\epsilon_i(\beta_0)\hat{\Pi}_i\epsilon_j(\beta_0)\hat{\Pi}_j] = 0$. Thus, the second term in the denominator of (19), which accounts for the covariance between individual scores in the numerator of the JLM statistic, does not appear in the expression of $JK(\beta_0)$. Note that this second term has a positive expectation under both positive selection, $\mathbb{E}[\epsilon_i(\beta_0)x_i] > 0$ for all $i \in [n]$, and negative selection, $\mathbb{E}[\epsilon_i(\beta_0)x_i] < 0$ for all $i \in [n]$. If this term is large, it can substantially increase the denominator of the JLM statistic relative to that of the $JK(\beta_0)$ statistic, reducing power of tests based on the JLM statistic. This again may be likely when the diagonal elements, P_{ii}, are large due to idempotency of the projection matrix: $P_{ii} = \sum_{j=1}^n P_{ij}^2$. Moreover, by construction $Var(r_i) \leq Var(x_i)$, so a large second term in the denominator of the JLM statistic may not be offset by a smaller first term, at least in local regions of H_0.

In sum, when bias taken on in constructing r_i is not too adverse, the $JK(\beta_0)$ statistic may have a larger numerator than the JLM statistic due to the use of higher quality first-stage estimates and a smaller denominator due to the use of uncorrelated individual scores. Since both the $JK(\beta_0)$ and JLM statistics are compared to the same χ^2 quantile, both of these properties may lead to more likely rejection of tests based on the $JK(\beta_0)$ statistic under H_1. Matsushita and Otsu (2022) note that the power properties of the JLM statistic are similar to those of the JAR statistic of Mikusheva and Sun (2021), suggesting that the improvements in power compared to the JAR test seen in Section 3 and Section 3.4 may be explained similarly.

The tests of Doví et al. (2024) and Lim et al. (2024) are very similar to the many instrument tests but replace estimators based on the deleted diagonal OLS projection matrix with those based on the deleted diagonal Ridge projection matrix. That is, they use $P_{ij} = [z(z'z + \lambda I_p)^{-1}z']_{ij}$ in the expression of (19). While this nominally allows for their analyzes to incorporate a number of instruments that is larger than the sample size, the power properties of the tests based on these statistics run into the same problems as described above.[19] Moreover, in these more recent papers, the penalty parameter λ must be chosen to satisfy certain technical properties that allow them to apply the many-instruments central limit theorem of Chao et al. (2012) used in Mikusheva and Sun (2021) and Matsushita and Otsu (2022). This contrasts to the construction of the $JK(\beta_0)$ statistic – in my analysis the researcher is afforded the flexibility to choose a penalty parameter that maximizes the predictive power of the first-stage estimates.

The difference in construction between the $JK(\beta_0)$ statistic and the Anderson-Rubin-type statistics of Doví et al. (2024) and Lim et al. (2024) also has implications for local power that can be understood using the analysis of Section 5.1. Let $P_\lambda := z(z'z + \lambda I_{d_z})^{-1}z'$ denote the ridge projection matrix used by these tests, with $\lambda = \lambda_n$ chosen according to the corresponding procedure. The deleted-diagonal quadratic-form statistic uses the weights $\xi_{ij} := [P_\lambda]_{ij}\mathbf{1}\{i \neq j\}$. After normalization, it therefore has the form

$$\frac{\sum_i \sum_{j\neq i} \xi_{ij}\epsilon_i(\beta_0)\epsilon_j(\beta_0)}{\sqrt{k_{\text{eff}}}}, \qquad k_{\text{eff}} := \sum_i \sum_{j\neq i} \xi_{ij}^2,$$

where k_{eff} is the effective number of off-diagonal ridge-projection weights. Since $\mathbb{E}[\epsilon_i(\beta_0)] = \Pi_i(\beta - \beta_0)$, the standardized mean shift of such a statistic under the alternative is of order

$$(\beta - \beta_0)^2 \frac{\sum_i \sum_{j\neq i} \xi_{ij}\Pi_i\Pi_j}{\sqrt{k_{\text{eff}}}}.$$

The deviation signal is therefore spread over all k_{eff} effective directions rather than concentrated in an estimated first-stage direction. For the resulting test to have nontrivial asymptotic power, this deviation signal must grow at least at the rate $\sqrt{k_{\text{eff}}}$. By contrast, the

[19] Indeed, in both Doví et al. (2024) and Lim et al. (2024) the power properties of the ridge-regularized tests appear to be similar to those of the many instrument tests.

local power calculation for $JK(\beta_0)$ has no analogous k_{eff} normalization. The analysis of Section 5.1 suggests that tests based on $JK(\beta_0)$ have nontrivial local power whenever the noncentrality parameter $\mu^2_\infty(\tilde{r})$ is bounded away from zero; since this parameter is of the same order as the local power index of Assumption 4.3, the requirement does not scale with the number of instruments. When the effective number of instruments diverges, there are thus sequences of local alternatives against which tests based on $JK(\beta_0)$ retain nontrivial power while the power of the Anderson-Rubin-type tests tends to their size, mirroring the classical comparison between the Anderson-Rubin and K-statistics when the number of instruments is fixed. I stress that this comparison is heuristic; a formal statement would require characterizing the behavior of the bootstrap critical values of Lim et al. (2024) along these local sequences, which is beyond the scope of this paper. It is, however, consistent with the calibrated simulations of Section 3, in which the test of Lim et al. (2024) controls size well but its power lies well below that of the jackknife K-test in each of the calibrated designs.

6. 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 may not fully address the power decline, I show that it works well in simulations calibrated to empirical data and leave further refinements to future work.

Declaration of Competing interest

We declare that we have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Proof of Theorem 4.1

Theorem 4.1 follows from the following two main technical lemmas, the proofs of which will comprise the majority of this appendix section. Let $JK_I(\beta_0)$ be the version of the test statistic that could be constructed if $\rho(\cdot)$ was known to the researcher, defined in more detail shortly.

Lemma A.1 (Infeasible Uniform Approximation). *Suppose that Assumptions 4.1 and 4.3 hold as well as the moment bounds of Theorem 4.1. Then,*

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

Lemma A.2 (Negligible Estimation Error). *Suppose that Assumptions 4.1 and 4.3 hold as well as the moment bounds of Theorem 4.1. Then, if* $(\Delta_N,\Delta_D)\to_p 0$,

$$\sup_{a\in\mathbb{R}}\Big|\Pr(JK(\beta_0)\le a)-\Pr(JK_I(\beta_0)\le a)\Big|\to_p 0$$

A.1. Proof of Lemma A.1

Before proceeding, we will introduce some notation. Let $\tilde{H}=s_nH$ and $\tilde{h}_{ij}=s_nh_{ij}$, where s_n is as in Assumption 4.1. Recall that $\tilde{h}_{ii}=0$ and define

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

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

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 Theorem A.1 is then equivalent to showing that $\sup_a |\Pr(JK^a \le 0) - \Pr(\tilde{JK}^a \le 0)| \to 0$. 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

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

We do so in a few lemmas, the final result being shown in Lemma A.10 at the bottom of this subsection.

Lemma A.3 (Lindeberg Interpolation). *Suppose that Assumptions 4.1 and 4.3 hold along with the conditions of Theorem 4.1. 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. (Proof of Lemma A.3)

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

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

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

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

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

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

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

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

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

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 I.1 and I.3, Cauchy-Schwarz, and the generalized Hölder inequality,[20] will allow us to bound for a fixed constant M that depends only on c,

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

and

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

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

Lemma A.4 (Gaussian Denominator Anti-Concentration). *Suppose that the conditions of Theorem 4.1 and Assumption 4.1 hold. Then, for any sequence $\delta_n \searrow 0$,*

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

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

Proof. (Proof of Lemma A.4) Since $\kappa_i^2(\beta_0) \in [c^{-1}, c]$ for all $i = 1, \dots, n$ we have that

$$\tilde{D} \geq \frac{c^{-1}}{n} \sum_{i=1}^{n} \Big(\sum_{j=1}^{n} \tilde{h}_{ij} r_j \Big)^2.$$

Then

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

where $\tilde{r} := (\tilde{r}_1, \dots, \tilde{r}_n)' \in \mathbb{R}^n$ 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 $c^{-1} < \mathbb{E}[r' \bar{H} r] < c$:

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

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

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 $\Sigma_r = \bar{H}^{1/2} \mathcal{R} \bar{H}^{1/2}$ where $\mathcal{R} = \mathrm{diag}(Var(r_1), \dots, Var(r_n))$. 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 4.1 and Lemma J.5, Λ_2/Λ_1 is bounded away from zero. Using Theorem K.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

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

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

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

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, $\lambda_1(M) \leq \sqrt{\mathrm{trace}(M^2)}$ 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.5. *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 $\delta > 0$:*

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

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

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

The second statement of the lemma follows symmetrically. □

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 $x \in \mathbb{R}$*

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

Then $X_n = O_p(1)$.

Lemma A.7. *Pick any $\epsilon > 0$, and let $M_{\epsilon/2}$ be such that $\Pr(Y_n > M_{\epsilon/2}) \le \epsilon/2$ for all $n \ge N_\epsilon$. In addition, let $\tilde{N}_\epsilon$ be such that $|\Pr(X_n \le M_{\epsilon/2}) - \Pr(Y_n \le M_{\epsilon/2})| \le \epsilon/2$ for all $n \ge \tilde{N}_\epsilon$. Then for all $n \ge N_\epsilon \vee \tilde{N}_{\epsilon/2}$,*

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

Lemma A.8. *Suppose that X_n and Y_n are sequences of (real-valued) random variables such that $Y_n = O_p(1)$ and for any $\Delta \in \mathbb{R}$*

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

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

Lemma A.9. *Pick an $\epsilon > 0$. By Lemma A.6, $X_n = O_p(1)$. Pick a constant $M_{\epsilon/3}$ such that $\Pr(X_n > M_{\epsilon/3}) \le \epsilon/3$ and $\Pr(Y_n > M_{\epsilon/3}) \le \epsilon/3$. Then for any $x \in \mathbb{R}$ we can bound $|\Pr(X_n \le x) - \Pr(Y_n \le x)|$ by considering two cases:*

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

$$|\Pr(X_n \le x) - \Pr(Y_n \le x)| \le \sup_{x \le M_{\epsilon/3}} |\Pr(X_n \le x) - \Pr(Y_n \le x)| \tag{A.10}$$

by hypothesis, there is an N_ϵ such that for $n \ge N_\epsilon$ the RHS of (A.10) is less than ϵ.

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

$$\begin{aligned}|\Pr(X_n \le x) - \Pr(Y_n \le x)| &\le |\Pr(X_n \le M_{\epsilon/3}) - \Pr(Y_n \le M_{\epsilon/3})| \\ &\quad + |\Pr(M_{\epsilon/3} < X_n \le x) - \Pr(M_{\epsilon/3} < Y_n \le x)| \\ &\le |\Pr(X_n \le M_{\epsilon/3}) - \Pr(Y_n \le M_{\epsilon/3})| + \epsilon/3 + \epsilon/3\end{aligned} \tag{A.11}$$

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

WLOG $N_{\epsilon/3} \ge N_\epsilon$. Combining the bounds in (A.10) and (A.11), for any $n \ge N_{\epsilon/3}$ and any $x \in \mathbb{R}$,

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

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

Lemma A.10 (Approximate Distribution). *Under Assumptions 4.1 and 4.3 and the conditions of Theorem 4.1*

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

Proof. (Proof of Lemma A.10) First, fix a $\Delta \ge 0$ and consider any $a \le \Delta$. As in Lemma A.4, 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 $x \le 0$, $\tilde{\varphi}(x)$ is decreasing if $x \in (0,1)$, and $\tilde{\varphi}(x)$ is zero if $x \ge 1$. 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.3 we can write for some constant M that depends only on Δ:

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

Applying Lemma A.5 and $\{\tilde{JK}^a \le 0\} = \{JK_G(\beta_0) \le a\}$ gives:

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

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

$$\Pr(JK_G(\beta_0) \le a) - e \le \Pr(JK_I(\beta_0) \le a) \le \Pr(JK_G(\beta_0) \le a) + e$$

where

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

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

$$\sup_{a \le \Delta} \Big|\Pr(JK_I(\beta_0) \le a) - \Pr(JK_G(\beta_0) \le a)\Big| \le C\Big(\frac{\gamma_n^{-2} + \gamma_n^{-3}}{\sqrt{n}} + \gamma_n^{1/2} + \gamma_n^{1/4}\Big) = o(1) \tag{A.12}$$

where C is a constant that depends only on Δ. Noting that the numerator $JK_G(\beta_0)$ is $O_p(1)$ under Assumption 4.3 while the inverse of the denominator of $JK_G(\beta_0)$ is $O_p(1)$ by Lemma A.4, we can apply Lemma A.8. This step shows that the result in (A.12) 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.12) yields the rate of decay in Remark 4.3. □

A.2. Proof of Lemma A.2

Proof. (Proof of Lemma A.2) For N and D defined at the top of Section A.1 define $\hat{N} = N + \Delta_N$ and $\hat{D} = D + \Delta_D$. We can then write $JK(\beta_0) = \hat{N}^2/\hat{D}$ 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 I.3 to see that $N^2 = O_p(1)$ while under Assumption 4.1, $D = O_p(1)$. Thus, $2ND\Delta_n + D\Delta_n - N^2\Delta_D = o_p(1)$. Meanwhile, by Lemma A.13, $\Pr(D^2 \le \delta_n) \to 0$ for any sequence $\delta_n \to 0$. Apply Lemma A.11 to obtain that $|JK(\beta_0) - JK_I(\beta_0)| \to_p 0$.

Finally, apply Lemma A.14 with $X_n = JK(\beta_0)$, $Y_n = JK_I(\beta_0)$ and $Z_n = JK_G(\beta_0)$ to show that the distribution of $JK(\beta_0)$ may be uniformly approximated by the distribution of $JK_G(\beta_0)$. The density of Z_n is uniformly bounded by Lemma I.4. □

Lemma A.11. *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 $\delta_n \to 0$, $\Pr(|Y_n| \le \delta_n) \to 0$, then,*

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

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

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

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

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

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

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

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

Lemma A.12 (Denominator Interpolation). *Suppose that the moment bounds of Theorem 4.1 and Assumption 4.1 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:*

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

Proof. (Proof of Lemma A.12) We inherit the definitions of D_{-i}, Δ^a_{2i}, Δ^b_{2i}, $\tilde{\Delta}^a_{2i}$, and $\tilde{\Delta}^b_{2i}$ from the proof of Lemma A.3 with $a = 1$. Then, as before we can write

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

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

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

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:

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

Using Lemma I.1 we can bound

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

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

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

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

Lemma A.13 (Denominator anti-concentration). *Suppose that the moment bounds of Theorem 4.1 and Assumption 4.1 hold. Then, for any sequence $\delta_n \searrow 0$,*

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

Proof. (Proof of Lemma A.13) 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 $x \le 0$, $\tilde{\varphi}(x)$ is decreasing if $x \in (0, 1)$, and $\tilde{\varphi}(x)$ is zero if $x \ge 1$. 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.12 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

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

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.4. □

Lemma A.14. *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 $\sup_{a\in\mathbb{R}} |\Pr(Y_n \le a) - \Pr(Z_n \le a)| \to 0$. Then $\sup_{a\in\mathbb{R}} |\Pr(X_n \le a) - \Pr(Z_n \le a)| \to 0$.*

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

$$\begin{aligned}\Pr(X_n \le a) &\le \Pr(Y_n \le a + \epsilon) + \Pr(|Y_n - X_n| > \epsilon) \\ &\le \Pr(Z_n \le a + \epsilon) + |\Pr(Y_n \le a + \epsilon) - \Pr(Z_n \le a + \epsilon)| \\ &\quad + \Pr(|Y_n - X_n| > \epsilon) \\ \text{so that } \Pr(X_n \le a) - \Pr(Z_n \le a) &\le \Pr(a < Z_n \le a + \epsilon) + |\Pr(Y_n \le a + \epsilon) - \Pr(Z_n \le a + \epsilon)| \\ &\quad + \Pr(|Y_n - X_n| > \epsilon)\end{aligned}$$

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

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

$$+ \Pr(|Y_n - X_n| > \epsilon_n)$$

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 $\sup_{a\in\mathbb{R}} |\Pr(Y_n \le a) - \Pr(Z_n \le a)| \to 0$ 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 $\sup_{a\in\mathbb{R}} \Pr(Z_n \le a) - \Pr(X_n \le a) \le o(1)$ which completes the claim of the lemma. □

Appendix B. Proof of Theorem 4.2

Proof. (Proof of Theorem 4.2) As at the top of Section 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 $\tilde{h}_{ij} = s_n h_{ij}$. A primary goal is to show that tests based on the infeasible statistic, $JK_I(\beta_0)$, are consistent. That is, $\Pr(JK_I(\beta_0) \le a) \to 0$ for any fixed $a \in \mathbb{R}_+$. The event $\{JK_I(\beta_0) \le a\}$ is equivalently expressed $\{N^2 - aD \le 0\}$ so that $\Pr(JK(\beta_0) \le a) = \Pr(N^2 - aD \le 0)$. Under the moment bounds of Theorem 4.1 and Assumption 4.1, $aD = O_p(1)$ so by Lemma B.2 it suffices to show that $\Pr(|N| \le M) \to 0$ for any fixed $M \ge 0$. By assumption, $P = \mathbb{E}[N^2] \to \infty$ so we move to show that $Var(N) = O(1)$ and then apply Lemma B.1 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

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

Notice that $N = N_1 + N_2$. To show that $Var(N_1) = O(1)$, 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 $Cov(\boldsymbol{a}_i, \boldsymbol{a}_j) = 0$ for $i \ne j$. Thus,

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

where the final inequality follows from the upper bound on $Var(\eta_i)$ and by definition of $\tilde{h}_{ij} = s_n h_{ij}$ from Assumption 4.1.

To show that $Var(N_2) = O(1)$ 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 4.3(ii), $|\boldsymbol{b}_i| = |\mathbb{E}[\sum_{j=1}^{n} \tilde{h}_{ji} \epsilon_j(\beta_0)]| \le c^{1/2}$, so we can bound

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

Since $Var(N) \le 2Var(N_1) + 2Var(N_2)$, we can conclude that tests based on $JK_I(\beta_0)$ are consistent.

Finally, we want to show that this fact, along with $(\Delta_N, \Delta_D) \to_p 0$ implies that tests based on $JK(\beta_0)$ are consistent. To do this, notice that we can write

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

and thus that $JK(\beta_0) \le a$ if and only if

$$\widehat{JK}_a := (N + \Delta_N)^2 - a(D + \Delta_D) = N^2 - aD + 2N\Delta_N + \Delta_N^2 - a\Delta_D \le 0.$$

Define $JK_a = N^2 - aD$. Using that $\{\widehat{JK}_a \le 0\} \subseteq \left\{ \frac{\widehat{JK}_a}{JK_a} JK_a \le 0 \right\} \cup \{JK_a \le 0\}$ we can write

$$\begin{aligned} \Pr(\widehat{JK}_a \le 0) &\le \Pr\left(\frac{\widehat{JK}_a}{JK_a} JK_a \le 0 \right) + \Pr(JK_a \le 0) \\ &\le 2\Pr(JK_a \le 0) + \Pr\left(\frac{\widehat{JK}_a}{JK_a} \le \frac{1}{2} \right) \end{aligned}$$

By consistency of the test based on the infeasible $JK_I(\beta_0)$ statistic, we have that $\Pr(JK_a \le 0) \to 0$. Thus, it only remains to show that $\Pr(\widehat{JK}_a/JK_a \le 1/2) \to 0$. This, in turn, follows if

$$\frac{\widehat{JK}_a - JK_a}{JK_a} = \frac{2N\Delta_N + \Delta_N^2 - a\Delta_D}{N^2 - aD} \to_p 0.$$

The above results can be used to show that $\Pr(|N^2 - aD| \le \delta_n) \to 0$ for any sequence $\delta_n \searrow 0$ so that $\frac{1}{JK_a} = O_p(1)$. Combined with $(\Delta_N, \Delta_D) \to_p 0$ this implies that $\{\Delta_N^2 - a\Delta_D\}/JK_a \to_p 0$. What remains is to show that $2N\Delta_N/(N^2 - aD) \to_p 0$. Write

$$\frac{2N\Delta_N}{N^2 - aD} = \frac{\frac{2N\Delta_N}{N^2}}{1 - a\frac{D}{N^2}}.$$

Since $D = O_p(1)$ while $\Pr(N^2 \le M) \to 0$ for any M we have that $D/N^2 \to_p 0$. Moreover, $\Pr(|N| \le M) \to 0$ for any fixed M implies $N/N^2 = O_p(1)$ so that $2N\Delta_N/N^2 \to_p 0$. We can apply continuous mapping theorem to conclude. □

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

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

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

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

Lemma B.2. *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 $n \geq N_X$):

$$\begin{aligned}\Pr(X_n^2 - Y_n > M) &\geq \Pr(X_n^2 > M_1 + M_Y, |Y_n| \leq M_Y) \\ &= 1 - \Pr(\{X_n^2 < M_1 + M_Y\} \cup \{|Y_n| > M_Y\}) \\ &\geq 1 - \epsilon/2 - \epsilon/2 = 1 - \epsilon\end{aligned}$$

□

Appendix C. Proof of Theorem 4.4

I provide the proof for the case where $d_x = 1$ with the general case following symmetrically. For any $j = 1, \dots, d_b$ define the matrix $B_j = \text{diag}(b_j(z_1), \dots, b_j(z_n))$ and collect observations $\epsilon(\beta_0) = (\epsilon_1(\beta_0), \dots, \epsilon_n(\beta_0))' \in \mathbb{R}^n$, $r = (r_1, \dots, r_n)' \in \mathbb{R}^n$, $\hat{r} = (\hat{r}_1, \dots, \hat{r}_n)' \in \mathbb{R}^n$, and $\xi = (\xi_1, \dots, \xi_n)' \in \mathbb{R}^n$. In addition, collect $b_\epsilon = (b_{\epsilon 1}, \dots, b_{\epsilon n}) \in \mathbb{R}^{d_b \times n}$ 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 K.1 to the quadratic form $\epsilon(\beta_0)'(\boldsymbol{H} B_j)\epsilon(\beta_0)$. First notice that, under Assumption 4.2(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 K.1 this give us an upper bound on $\|\mathbb{E} f^{(1)}(X)\|_{\text{HS}}$. Next, Assumption 4.1 gives us that the Frobenius norm of $\boldsymbol{H} = \frac{s_n}{\sqrt{n}} \boldsymbol{H}$ is bounded, since the rows of $s_n H$ are square summable, $\sum_{j \neq i} (s_n h_{ij})^2 \leq c$ for all $i = 1, \dots, n$. In the notation of Theorem K.1 this gives us an upper bound on $\|\mathbb{E} f^{(2)}(X)\|_{\text{HS}}$. Applying Theorem K.1 and a union bound then gives us that

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

Since $\max_{1 \leq j \leq d_b} |\mathbb{E}[\epsilon(\beta_0)' \boldsymbol{H} B_j \epsilon(\beta_0)]| \leq c$ under Assumption 4.2(v), (C.1) 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 $\log^{2/a}(d_b) \|\hat{\gamma} - \gamma\|_1 \to_p 0$ 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 4.3(ii) while under Assumption 4.2 $\|\xi\|_2 = o(1)$.

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

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

Since both $\mathbf{E} = O_p(1)$ and $\mathbf{F} = O_p(1)$ under the moment bounds of Theorem 4.1 and Assumption 4.1, it suffices to show that

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

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 4.2(v) $\max_{i,j}|\mathbb{E}[\sum_{j\neq i}\tilde{h}_{ij}b(z_j)\epsilon_j(\beta_0)| \leq c$. Under Assumptions 4.1 and 4.2(ii), $\max_{i,j}\sum_{j\neq i}\tilde{h}_{ij}^2 b^2(z_j) \leq c^2$ so we can apply Theorem K.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 4.2(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 4.1 and 4.2(ii), and $\sum_{i=1}^n \xi_i^2 = o(1)$ by Assumption 4.2(iii).

Appendix D. Proof of Theorem 4.3

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

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

As with Theorem 4.1, the result Theorem 4.3 follows directly from combining the following lemmas. The first is the main technical lemmma, and shows that the distribution of the infeasible statistic $JK_I(\beta_0)$ can be uniformly approximated by that of $JK_G(\beta_0)$. The proof of this technical lemma is involved and deferred to Section F. The second lemma establishes that estimation error can be treated as negligible. As with Lemma A.2, the main difficulty hear is in dealing with the fact that neither the numerator vector nor denominator matrix of the $JK(\beta_0)$ statistic may have stable limiting distributions.

Lemma D.1. *Suppose that Assumptions 4.1 and 4.3 hold as well as the moment conditions of Theorem 4.3. Then,*

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

Proof. (Proof of Theorem D.1) Theorem D.1 follows as a consequence of the joint gaussian approximation with the combination statistic established in Section F. □

Lemma D.2. *Suppose that Assumptions 4.1 and 4.3 hold along with the moment conditions of Theorem 4.3. Then, if* $(\Delta_N, \Delta_D) \to_p 0$, $\Big|JK(\beta_0) - JK_I(\beta_0)\Big| \to_p 0$.

Proof. (Proof of Lemma D.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

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

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

By assumption 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 F.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

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

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

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

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

Appendix E. Proofs of results in Section 5

The statement of Theorem 5.1 relies on showing

$$\sup_{(a_1,a_2)\in\mathbb{R}^2}\Big|\Pr(JK(\beta_0)\le a_1, C\le a_2) - \Pr(JK_G(\beta_0)\le a_1, C_G\le a_2)\Big| \to 0$$

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

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 K.5 after verifying Assumption K.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.

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

This joint interpolation argument is rather involved however, and deferred to Section F. The interpolation argument for the conditioning statistic very closely follows the results in Chernozhukov et al. (2013). The results of Section 5 rely on showing that the difference between C and C_I can be treated as negligible. This in turn reduces to verifying Assumption K.2, which is done in Lemma E.1, below.

Lemma E.1. *Suppose that Assumption 4.2 holds. Then there are sequences $\delta_n \searrow 0$, $\beta_n \searrow 0$ such that*

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

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

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

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

Notice that we can bound

$$\begin{aligned} \max_{1\le i\le n}\frac{1}{n}\sum_{j=1}^n(\hat{r}_i - r_i)^2 &= \max_{1\le i\le n}\Big|(\hat{\gamma}-\gamma)'n^{-1}\sum_{j=1}^n \epsilon_j^2(\beta_0)b(z_i)b(z_j)'(\hat{\gamma}-\gamma)\Big| \\ &\quad + \max_{1\le i\le n}|n^{-1}\sum_{j=1}^n \dot{h}_{ij}^2\xi_j^2| \end{aligned}$$

$$\leq \max_{\substack{1\leq i\leq n\\ 1\leq j,k\leq d_b}} \underbrace{\left|n^{-1}\sum_{j=1}^{n}\epsilon_j^2(\beta_0)b_j(z_j)b_k(z_j)\right|}_{\mathbf{A}_{ijk}} \|\widehat{\gamma}-\gamma\|_1^2$$

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

Under Assumption 4.2(i,ii) each $\mathbf{A}_{ijk}$ is υ-sub-exponential by Theorem K.1 (that is $\|\mathbf{A}_{ijk}\|_{\psi_\upsilon}$ is bounded). An application of Lemma J.2 then yields that $\max_{i,j,k}|\mathbf{A}_{ijk}| = O_p(\log^{1/\upsilon}(d_b n))$. Along with Assumption 4.2(iv) this gives that $\max_{i,j,k}|\mathbf{A}_{ijk}|\|\widehat{\gamma}-\gamma\|_1 = O_p(\log^{-3/(\upsilon\wedge 1)}(d_b n)) = o_p(\log^{-2}(n))$. 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 4.2(iii) $(\sum_{j=1}^n \xi_j^4)^{1/2} = o(1)$. Since $\log^2(n)/\sqrt{n}\to 0$ this shows (E.2). □

E.1. Proof of Theorem 5.1

The first result in Theorem 5.1 with $JK(\beta_0)$ and C replaced with their infeasible analogs $JK_I(\beta_0)$ and C_I follows from the argument in Section F. After verifying that $|JK(\beta_0) - JK(\beta_0)| \to_p 0$ via Theorem 4.4 and that Assumption K.2 is satisfied via Lemma E.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 K.6 are satisfied. To see that Assumption K.1(i,ii) is satisfied under the moment assumptions of Theorem 4.1 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 K.1(iii) is satisfied with $B_n = \log^{1/\upsilon}(n)$ by Assumption 5.1(i,iii) and Lemma J.2. Finally Assumption K.2 is satisfied by applying Lemma E.1. Apply Theorem K.6 to conclude.

Appendix F. Joint Gaussian approximation of $JK(\beta_0)$ and C

Theorems 4.3 and 5.1 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 Section F.2 after introducing some notation in Section F.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 $JK(\beta_0)$ statistic 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.

F.1. Notation

Jackknife Statistic Definitions. Define $\tilde{h}_{\ell,ij} = s_{n,\ell}h_{ij}$ for each $\ell = 1,\dots,d_x$ 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 Theorem A.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

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

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

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

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

where

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

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

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

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

Notice that in this notation we can write the test statistic and gaussian test statistics, after scaling by $\mathrm{diag}(s_{n,1},\dots,s_{n,d_x})$, 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 $\ell = 1,\dots,d_x$ and $i = 1,\dots,n$. 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:

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

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 $C = \max_{1\le l\le 2nd_x}(C_{-1} + \frac{1}{\sqrt{n}}\Delta_{C1})_l$, while $\tilde{C} = \max_{1\le l\le 2nd_x}(C_{-n} + \Delta_{Cn})_l$.

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

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

Section I.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 $\tilde{\varphi}(\cdot,\cdot,\cdot) : \mathbb{R}^{d_x} \times (\mathbb{R}^{d_x \times d_x}) \times \mathbb{R}^{2nd_x} \to \mathbb{R}$ via

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

where

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

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.

F.2. Main argument

Lemma F.1 (Joint Lindeberg Interpolation). *Suppose that Assumptions 4.1 and 4.3 hold as well as the moment conditions of Theorem 4.3. Then there are fixed constants M_1, M_2 such that*

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

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

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

By second degree Taylor expansion, we break each of the summands in (F.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 $i = 1, \dots, n$ and (ii) the mean and covariance matrix of $(\epsilon_i(\beta_0), r_i)$ being equal to the mean and covariance matrix of $(\tilde{\epsilon}_i(\beta_0), r_i)$

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

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

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

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

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

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

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 I.13 and I.18. Under the moment conditions of Theorem 4.3 the absolute value of terms $(\Delta_{Ui})_\ell, |\Delta_{Di}^a|_{mn}$, and $(\Delta_{Di}^b/\sqrt{n})_{no}$ can also be shown to have bounded third moments via the exact same steps as in the proof of Lemma I.1. Putting these together with generalized Holder's inequality will yield a finite constants M_1 and M_2 such that $|\mathbf{A}_{lmn}| \le M_1 \log^{M_2}(n)(\gamma^{-1} + \gamma^{-2})$, $\mathbf{B}_{\ell mno} \le M_1 \log^{M_2}(n)(\gamma^{-1} + \gamma^{-2})$, and $|\mathbf{C}_{\ell mno}| \le M_1 \log^{M_2}(n) n^{1/2}(\gamma^{-1} + \gamma^{-2})$. To bound $\mathbf{D}_{\ell mn}$ terms notice that

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

Apply Lemma I.1 to bound Δ_{Di}^a, and Lemmas I.13 and I.18 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 I.9 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 J.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

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

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},\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},\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},\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},\vec{(}\bar{D}),\bar{C})(\Delta_{Ui})_\ell(\Delta_{Ui})_m(\Delta_{Di})_{no}\right]\\
&+\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},\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},\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},\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},\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},\vec{(}\bar{D}),\bar{C})(\Delta_{Ci})_\ell(\Delta_{Ci})_m(\Delta_{Ui})_n\right]
\end{aligned}$$

where $\bar{U},\vec{(}\bar{D}),$ and $\bar{C}$ vary term by term but are always in the hyper-rectangles $[U_{-i}, U+\Delta_{Ui}]$, $[\vec{(}D_{-i}),\vec{(}D_{-i}+\Delta_{Di})]$, 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 I.1 to bound moments of Δ_{Ui} and $(\Delta_{Di}/\sqrt{n})$, Lemma I.18 to bound moments of the second and third derivatives of $\phi(\tilde{U},\vec{(}\tilde{D}))$, Lemma I.12 to bound the sums of derivatives of $\tau(\tilde{C})$, and Lemma J.2 to bound moments of $\max_{1\leq\ell\leq n}(\Delta_{Ci})_\ell$ will yield that

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

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

Lemma F.2 (Denominator Anticoncentration). *Suppose that Assumptions 4.1 and 4.3 hold as well as the moment conditions of Theorem 4.3. Then for any sequence $\delta_n \to 0$ we have that* $\Pr(\lambda_{\min}(\bar{D}) \leq \tilde{\delta}_n) \to 0$.

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

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

$$\ge \frac{1}{cn}\sum_{i=1}^{n}\Big(\sum_{\ell=1}^{d_x}\sum_{j=1}^{n}a_\ell\tilde{h}_{\ell,ij}r_{\ell,j}\Big)^2 \text{Define } \dot{s}_{n,j} = \max_{\{\ell:a_\ell\neq 0\}} s_{n,\ell} \text{ and } \dot{h}_{ij} = s_n h_{ij} \qquad = \frac{1}{cn}\sum_{i=1}^{n}\Big(\sum_{j=1}^{n}\dot{h}_{ij}\sum_{\ell=1}^{d_x}\frac{a_\ell s_{n,\ell}}{s_n}r_{\ell,j}\Big)^2$$

By the moment conditions required by Theorem 4.3 we have that $\lambda_{\min}(\mathbb{E}[D]) \ge \underline{c}$, so that

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{\ell=1}^{d_x}\sum_{j=1}^{n}a_\ell\tilde{h}_{\ell,ij}r_{\ell,j}\right)^2\right] \ge c^{-1}.$$

Moreover, by assumption, $Var(\sum_{\ell=1}^{d_x}\frac{a_\ell s_{n,\ell}}{s_n})$ is bounded from above and below. Define the matrix $\tilde{H} = [\dot{h}_{ij}]_{ij}$ and follow the same steps as Lemma F.2 to conclude. □

Lemma F.3 (Gaussian Approximation). *Suppose that Assumptions 4.1 and 4.3 hold as well as the moment conditions of Theorem 4.3. Then,*

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

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

$$\Pr(N'D^{-1}N \le a_1, C \le a_2) \le \mathbb{E}[\tilde{\phi}_{\gamma,a}(U,\vec{(}D),C)]$$

$$\le \mathbb{E}[\tilde{\phi}_{\gamma,a}(\tilde{U},\vec{(}\tilde{D}),\tilde{C})] + \frac{M_1\log_2^M(n)}{\sqrt{n}}(\gamma^{-1}+\gamma^{-2})$$

$$\le \Pr(\tilde{N}'\tilde{D}^{-1}\tilde{N} \le a_1, \tilde{C} \le a_2) + \Pr(a_1 \le \tilde{N}'\tilde{D}^{-1}N \le a_1 + \gamma\lambda_{\min}^5(D))$$

$$+ \Pr(a_2 \le C \le a_2+\gamma) + \frac{M_1\log_2^{M_2}(n)}{\sqrt{n}}(\gamma^{-1}+\gamma^{-2}+\gamma^{-3})$$

$$\le \Pr(\tilde{N}'\tilde{D}^{-1}\tilde{N} \le a_1, \tilde{C} \le a_2) + \Pr(a_1 \le \tilde{N}'\tilde{D}^{-1}N \le a_1 + \gamma\lambda_{\min}^5(D))$$

$$+ \Pr(a_2 \le C \le a_2+\gamma) + \frac{M_1\log_2^{M_2}(n)}{\sqrt{n}}(\gamma^{-1}+\gamma^{-2}+\gamma^{-3})$$

Let $\gamma \to 0$ at a rate such that $\frac{\log^{M_2}(n)}{\sqrt{n}}\gamma^{-3} \to 0$ and apply Lemmas F.1 and F.2 to conclude as in the proof of Lemma A.10. A symmetric argument shows that the lower bound tends to zero.

□

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

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 $\Pr(\lambda_{\min}^2(\Sigma_n) \le \tilde{\delta}_n) \to 0$. For any $m \in \mathbb{N}$ let $\mathcal{A}_m$ be a set of points in S^{d-1} such that

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

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

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

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

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

and notice that, by definition $\Pr(\min_{a\in\mathcal{A}_{\tilde{n}_j}} a'\Sigma_n a \le 2\tilde{\delta}_n) < \delta_{\tilde{n}_j}$. 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

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

Define the events

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

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

Appendix G. 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 $y := (y_1, \ldots, y_n)' \in \mathbb{R}^n$ and $x := (x'_1, \ldots, x'_n)' \in \mathbb{R}^{n\times}$ define $y^\perp := M_2 y$ and $x^\perp := M_2 x$ 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 $\epsilon(\beta_0) := y - x\beta_0$, $\epsilon^\perp(\beta_0) = M_2\epsilon(\beta_0)$ and $r^\perp := M_2 r$ 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 $M_2 H$ and the vector $\hat{\Pi} \in \mathbb{R}^n$ is defined $\hat{\Pi} = (M_2 H)(M_2 r)$. 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,

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

where $\boldsymbol{h}_{ij} = [M_2 \tilde{H} M_2]_{ij}$, $\tilde{H} = s_n H$, 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

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

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 $Var(\tilde{\epsilon}(\beta_0)) = Var(\epsilon_i)$ 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^2_1 distribution conditional on any realization of $\tilde{r}$ and thus its unconditional distribution is also χ^2_1.

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 A.2, it suffices to show for the first step that $\Delta_D \to_p 0$, where

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

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

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

Both of these terms will tend to zero by the consistency $\hat{\Gamma}$ 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 Section 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 $M_2 H M_2$ 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 Section 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.3. 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 $\Big(\sum_{j\neq i} H_{ji}^2\Big) \lesssim 1$ by Assumption 4.1(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

$$\begin{aligned}\Delta_{2i} &= \epsilon_i^2(\sum_{j\neq i}\boldsymbol{h}_{ij}\dot{r}_j)^2 + r_i^2\sum_{j\neq i}\boldsymbol{h}_{ji}^2\ddot{\epsilon}_j^2 + r_i\sum_{j\neq i}\ddot{\epsilon}_j\Big(\sum_{k\neq j,i}\boldsymbol{h}_{ji}\boldsymbol{h}_{jk}r_k\Big)\\ &\quad + \epsilon_i^2\big(\boldsymbol{h}_{ii}^2r_i^2 + 2\boldsymbol{h}_{ii}r_j\sum_{j\neq i}\boldsymbol{h}_{ij}r_j\big)^2\\ \tilde{\Delta}_{2i} &= \tilde{\epsilon}_i^2(\sum_{j\neq i}\boldsymbol{h}_{ij}\dot{r}_j)^2 + \tilde{r}_i^2\sum_{j\neq i}\boldsymbol{h}_{ji}^2\ddot{\epsilon}_j^2 + \tilde{r}_i\sum_{j\neq i}\ddot{\epsilon}_j\Big(\sum_{k\neq j,i}\boldsymbol{h}_{ji}\boldsymbol{h}_{jk}r_k\Big)\\ &\quad + \epsilon_i^2\big(\boldsymbol{h}_{ii}^2r_i^2 + 2\boldsymbol{h}_{ii}r_j\sum_{j\neq i}\boldsymbol{h}_{ij}r_j\big)^2\end{aligned}$$

where $\ddot{\epsilon}_j$ is equal to $Var(\epsilon_j)$ 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.3. 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 Section A.

Appendix H. Alternative construction of test statistic via cross fitting

To accomodate a general class of estimators for $\hat{\rho}(z_i)$ I propose a cross-fit form of the jackknife K-statistic. In this section I detail the cross-fitting procedure and present high level conditions needed for estimation error to be treated as negligible. These conditions can be satisfied by a large class of machine learning estimators under alternate conditions.

H.1. Cross fit test statistic

To construct the cross fit test statistic, evenly (and randomly) split the sample into two subsets, $\mathcal{I}_1$ and $\mathcal{I}_2$ such that $\mathcal{I}_1 \cap \mathcal{I}_2 = \emptyset$ and $\mathcal{I}_1 \cup \mathcal{I}_2 = [n]$. For each $k = 1, 2$ construct an estimator $\hat{\rho}^{(k)}(\cdot)$ using only the observations in $\mathcal{I}_k$. For observations in $i \in \mathcal{I}_1$ form the first-stage estimates

$$\hat{\Pi}_i = \sum_{j\in\mathcal{I}_1\setminus\{i\}} h_{ij}^{(1)}(x_j - \epsilon_j(\beta_0)\hat{\rho}^{(2)}(z_j))$$

where the weights $h_{ij}^{(1)}$ come from a hat matrix $H^{(1)}$ that only depends on the observations in $\mathcal{I}_1$, for example the ridge regression hat matrix of Section 2 using only the instruments $(z_i)_{i\in\mathcal{I}_1}$. First stage estimates for observations in $\mathcal{I}_2$ symmetrically, using the estimator $\hat{\rho}^{(1)}(\cdot)$. The test statistic is then constructed as before

$$JK(\beta_0) = \frac{\big(\sum_{i=1}^n \epsilon_i(\beta_0)\hat{\Pi}_i\big)^2}{\sum_{i=1}^n \epsilon_i^2(\beta_0)(\hat{\Pi}_i)^2}$$

Notice that while only half the sample is being used to construct each first stage estimate, the full sample is still being used to test the null hypothesis.

H.2. Controlling estimation error

After writing the overall hat matrix as

$$H = \begin{pmatrix} H^{(1)} & \mathbf{0} \\ \mathbf{0} & H^{(2)}\end{pmatrix},$$

analysis of the infeasible statistic proceeds as before. Thus, to show that the crossfit test statistic has a limiting χ^2_1 distribution by Lemma A.2 it suffices to state high-level conditions under which $(\Delta_N, \Delta_D)' \to_p 0$.

Assumption H.1 (Cross-fit Conditions). *Suppose (i) that there is a constant $v \in (0,1] \cup \{2\}$ such that for all $i \in [n]$, $\|\epsilon_i\|_{\Psi_v} \leq c$ and that (ii) for $k = 1, 2$*

$$\max_{i \in I_k} (\hat{\rho}^{(-k))}(z_j) - \rho(z_j))^2 = o_p(\log^{1/v}(n)).$$

where $\hat{\rho}^{(-k)}(\cdot)$ indicates the estimator of $\rho(\cdot)$ computed using observations in $[n] \setminus I_k$.

Lemma H.1 (Negligible Cross-fit error). *Suppose that Assumptions 4.1, and H.1 hold as well as the moment conditions of Theorem 4.1. Then, under H_0, $(\Delta_N, \Delta_D)' \to_p 0$.*

Proof. We consider the statements $\Delta_N \to_p 0$ and $\Delta_D \to_p 0$ seperately.

$\Delta_N \to_p 0$:. It suffices to show that $\Delta_{N,1} \to_p 0$ for

$$\Delta_{N,1} = \frac{1}{\sqrt{n}} \sum_{i \in I_1} \epsilon_i(\beta_0) \sum_{j \in I_1 \setminus \{i\}} \epsilon_j(\beta_0) \tilde{h}_{ij} (\hat{\rho}^{(2)}(z_j) - \rho(z_j)).$$

The corresponding statement for $\Delta_{N,2}$ follows from the same logic and $\Delta_N = \Delta_{N,1} + \Delta_{N,2}$. Define the event

$$\Omega(\epsilon) := \{\max_{i \in I_k} (\hat{\rho}^{(2)}(z_j) - \rho(z_j))^2 \leq \epsilon\}$$

and consider a sequence $\epsilon_n \to 0$ such that $\Pr(\Omega(\epsilon_n)) \geq 1 - \epsilon_n$. Noting that $\Omega(\epsilon_n) \perp (\epsilon_i(\beta_0))_{i \in I_k}$ we write

$$\Delta_{N,1} = \epsilon(\beta_0) \boldsymbol{H} \epsilon(\beta_0)$$

where $\boldsymbol{H} \in \mathbb{R}^{|I_1| \times |I_1|} = \frac{1}{\sqrt{n}} \left(\tilde{h}_{ij} (\hat{\rho}^{(1)}(z_j) - \rho(z_j))\right)_{i,j \in I_1}$. Since $\mathbb{E}[\Delta_{N,1} | \Omega(\epsilon_n)] = 0$, an application of the generalized Hanson-Wright inequality, Theorem K.1, gives us that there is a sequence $\delta_n \to 0$ such that

$$\Pr(\Delta_{N,1} \geq \epsilon_n | \Omega(\epsilon_n)) \leq \delta_n$$

which allows us to conclude.

$\Delta_D \to_p 0$:. As before, it suffices to show that $\Delta_{D,1} \to_p 0$ where

$$\Delta_{N,1} = \frac{1}{n} \sum_{i \in I_1} \epsilon_i^2(\beta_0) \Big\{ \Big(\sum_{j \in I_1 \setminus \{i\}} \tilde{h}_{ij} \hat{r}_j \Big)^2 - \Big(\sum_{j \in I_1 \setminus \{i\}} \tilde{h}_{ij} r_j \Big)^2 \Big\}$$

From the proof of Theorem 4.4, it suffices to show that

$$\max_{i \in I_1} \Big| \sum_{j \in I_1 \setminus \{i\}} \tilde{h}_{ij} \epsilon_j(\beta_0) (\hat{\rho}^{(2)}(z_j) - \rho(z_j)) \Big| \to_p 0$$

Consider a sequence $\epsilon_n \to 0$ such that $\log^{1/v}(n) \epsilon_n \to 0$ and $\Pr(\Omega(\epsilon_n)) \geq 1 - \epsilon_n$ and apply Theorem K.1 to conclude that there is a sequence $\delta_n \to 0$ such that

$$\Pr\Big(\max_{i \in I_1} \Big| \sum_{j \in I_1 \setminus \{i\}} \tilde{h}_{ij} \epsilon_j(\beta_0) (\hat{\rho}^{(2)}(z_j) - \rho(z_j)) \Big| \geq \epsilon_n \mid \Omega(\epsilon_n) \Big) \leq \delta_n$$

Since $\Pr(\Omega(\epsilon_n)) \to 1$ this gives the result. □

Appendix I. Relevant moment bounds

I.1. Moment bounds for Section 4

Here I provide some lemmas that are useful in the proof of Lemmas A.3–A.10

Lemma I.1. *Let $\Delta_{1i}, \tilde{\Delta}_{1i}, \Delta^a_{2i}, \tilde{\Delta}^a_{2i}, \Delta^b_{2i}, \tilde{\Delta}^b_{2i}$ be as in (A.2). Then under Assumption 4.1 and the moment conditions of Theorem 4.1 there is a constant $M > 0$ such that for any $k = 1, \ldots, 6$:*

$$\mathbb{E}[|\Delta_{1i}|^k] \leq M \qquad \mathbb{E}[|\tilde{\Delta}_{1i}|^k] \leq M \text{and for any } k = 1, \ldots, 3: \mathbb{E}[|\Delta^a_{2i}|^k] \leq M\alpha^k \mathbb{E}[|\tilde{\Delta}^k_{2i}|] \leq M\alpha^k$$

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

Lemma I.2. *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 I.5 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 $k = 1, \ldots, 6$

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

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 $\epsilon_i(\beta_0)$. The bounds on $\mathbb{E}[|\Delta_{2i}^a|^k]$ and $\mathbb{E}[|\tilde{\Delta}_{2i}^a|^k]$ also follow from (I.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 I.7 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 (I.1).

Lemma I.3. *Let N and N_{-i} be defined as in Section A.1. Under Assumptions 4.1 and 4.3 and the moment conditions in Theorem 4.1, 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 $N = N_1 + N_2 + N_3$:

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

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

Since via Assumption 4.1, $\sum_{i=1}^n h_{ji}^2 \leq c$, and $|\gamma_j| \leq c$, 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 4.3, $\mathbb{E}[N_2^2] \leq c$ while Assumption 4.1 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 I.5 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 I.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 I.8.

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

Lemma I.4. *Let $\tilde{N}$ and $\tilde{D}$ be defined as in Section A.1. Let $f(\cdot, \tilde{r})$ be the density function of $\frac{\tilde{N}}{\tilde{D}^{1/2}}|\tilde{r}$. Under Assumption 4.3 and the moment bounds of Theorem 4.1, there is a constant $M > 0$ such that $\sup_x |f(x, \tilde{r})| \leq M$ for almost all $\tilde{r}$.*

Proof. Recall that

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

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

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

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

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

The maximum of the normal density is proportional to the inverse of the standard deviation so it suffices to show that the variance in (I.2) is bounded away from zero. To this end, notice that under the moment bounds of Theorem 4.1 and Assumption 4.3

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

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

□

Lemma I.5. *Let $X_1, \dots, X_n$ be random variables such that $\mathbb{E}[X_i] = \mu_i$ and $\mathbb{E}[(\sum_{i=1}^n X_i)^2] \le C$. Suppose that for any $i = 1, \dots, n$ there is a constant U such that*

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

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

Proof. First write

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

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

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

where the implied constant in the second line is 32 by an application of Lemma J.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 I.6. *Let $h = (h_1, \dots, h_n) \in \mathbb{R}^n$ be such that $\sum_{i=1}^n h_i^2 \le b$. Suppose that $X_1, \dots, X_n$ are such that $\mathbb{E}[|X_i|^k] \le M$ for all $k = 1, 2, 3$. Then*

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

Proof. We can expand out

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

$$\leq M^3\Big(\sum_{i=1}^n h_i^2\Big)^3 \leq c^3M^3$$

□

Lemma I.7. *Let $v_1,\dots,v_n$ be random variables such that $\mathbb{E}[|v_i|^3]\leq M$ for all $i=1,\dots,n$. Let $h=(h_1,\dots,h_n)\in\mathbb{R}^n$ 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

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

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 I.8. *Let $v_1,\dots,v_n$ be a martingale difference array such that $\mathbb{E}[|v_i|^l]\leq M$ for all $l=1,\dots,k$. Then there is a fixed constant C_k that only depends on k such that*

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

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

$$\mathbb{E}[(\frac{1}{\sqrt{n}}\sum_{i=1}^n v_i)^k]\leq\mathbb{E}[(\max_{s\leq n}\sum_{i=1}^s X_s)^k]$$
$$\leq C_k\mathbb{E}\Big[\Big(\sum_{i=1}^n v_i^2/n\Big)^{k/2}\Big]\leq C_k\mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^n v_i^k\Big]\leq C_kM$$

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

I.2. Useful properties of smooth max

Lemma I.9 (Chernozhukov et al. (2013), Lemma A.2). *For every $1\leq j,k,l\leq p$,*

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

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

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

Moreover,

$$\pi_j(z)\geq0,\qquad\sum_{j=1}^p\pi_i(z)=1,\qquad\sum_{j,k=1}^p|w_{jk}(z)|\leq2,\qquad\sum_{j,k,l=1}^p|q_{jkl}|\leq6$$

Lemma I.10 (Chernozhukov et al. (2013), Lemma A.3). *For every $x,z\in\mathbb{R}^p$,*

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

Lemma I.11 (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*

$$\partial_jm(z)=(\partial g(F(\beta))\pi_j)(z)$$
$$\partial_j\partial_km(z)=(\partial^2g(F_\beta)\pi_j\pi_k+\partial g(F_\beta)\beta w_{jk})(z)$$
$$\partial_j\partial_k\partial_lm(z)=(\partial^3g(F_\beta)\pi_j\pi_k\pi_l+\partial^2g(F_\beta)\beta(w_{jk}\pi_l+w_{jl}\pi_k+w_{kl}\pi_j)+\partial g(F_\beta)\beta^2q_{jkl})(z)$$

where π_j,w_{jk},q_{jkl} are as described in Lemma I.9.

Lemma I.12 (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* $1 \le j, k, l \le p$,

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

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,

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

I.3. Moment bounds for Theorems 4.3 and 5.1

Lemma I.13. *Suppose that the moment conditions of Theorem 4.3 hold and let* N *and* D *be as defined at the top of Section F.2 Then under* H_0, *for any* k *there is a fixed constant* C_k *such that for any* $\ell = 1, \dots, d_x$

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

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

I.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 I.14. *Let* $D \in \mathbb{R}^{d \times d}$ *be a symmetric, real matrix such that* $\det(D) \ne 0$. *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:

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

Second Order:

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

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

Third Order:

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

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

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

Lemma I.15. *The derivative of an element of the inverse of a matrix $\boldsymbol{X}$ can be expressed (Petersen and Pedersen, 2012)*

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

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 I.16. *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*

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

Lemma I.17. *We can express the derivative of the detrminant (Petersen and Pedersen, 2012),*

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

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

Lemma I.18. *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 I.14 and I.16 as well as the inequality (I.5) below. For the moment bounds, we may extensive use of following bounds on elements of D^{-1} for a positive-definite D^{-1}:

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

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 (I.5) and the expressions in Lemmas I.14 and I.16. 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.

$$\begin{aligned} \mathbb{E}\left|A\frac{\partial\gamma}{\partial N_l}\right|^k &\lesssim \sum_{j=1}^{d}\mathbb{E}|(\text{trace}(D))^{kdp}N_j^kA^k| \\ &\lesssim \sum_{j=1}^{d}\sum_{\imath=1}^{d}\mathbb{E}[D_{\imath\imath}^{kdp}N_j^kA^k] \\ &\leq \sum_{j=1}^{d}\sum_{\imath=1}^{d}\gamma^{2kdp}\mathbb{E}[N_j^{2k}A^{2k}] \end{aligned}$$

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

□

Appendix J. Technical lemmas

J.1. Probability lemmas

Lemma J.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:

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

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

Lemma J.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

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

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 $\Pr(|X_i|^\alpha \geq t) \leq 2\exp(-t/K)$.

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

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

J.2. Matrix lemmas

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

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

J.3. Miscellaneous lemmas

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

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

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

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

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

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 $(1, \dots, 1) \in \mathbb{R}^m$ □

Appendix K. Assorted results from literature

K.1. Concentration inequalities and tail bounds

Theorem K.1 (Gotze et al. (2021)*Theorem 1.2). *Let $X_1, \dots, X_n$ be independent random variables satisfying $\|X_i\|_{\Psi_a} \le M$ 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 $t > 0$;*

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

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

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

or equivalently

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

Theorem K.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 $t \ge 0$, we have*

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

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

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

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

K.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 $\lambda_{1\xi} \ge \lambda_{2\xi} \ge \dots \ge \lambda_{n\xi}$. Define the quantities

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

Theorem K.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*

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

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

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

where C is a constant only depending on b.

K.3. Gaussian comparasions and approximations

We also use the following gaussian approximation results from Belloni et al. (2018), Chernozhukov et al. (2017). Let $X_1, \dots, X_n \in \mathbb{R}^p$ be independent, mean zero, random vectors and let $Y_1, \dots, Y_n \in \mathbb{R}^p$ 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 $X_1, \dots, X_n$ but instead observes noisy estimates $\hat{X}_1, \dots, \hat{X}_n \in \mathbb{R}^p$.

Define the sums

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

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

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

for some $-\infty \le a_j \le b_j \le \infty$, $j = 1, \dots, p$. 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 K.1. *Suppose there are constants $B_n \ge 1$, $b > 0, q > 0$ such that*

(i) $n^{-1} \sum_{i=1}^n \mathbb{E}[X_{ij}^2] \ge b$ *for all* $j = 1, \dots, p$
(ii) $n^{-1} \sum_{i=1}^n \mathbb{E}[|X_{ij}|^{2+k}] \le B_n^k$ *for all* $j = 1, \dots, p$ *and* $k = 1, 2$.
(iii) $\mathbb{E}[(\max_{1 \le j \le p} |X_{ij}|/B_n)^4] \le 1$ *for all* $i = 1, \dots, n$ *and* $\left(\frac{B_n^4 \ln^7(pn)}{n} \right)^{1/6} \le \delta_n$.

as well as the following bounds on the estimation error

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

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

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

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

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

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

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

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

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 $e_1, \dots, e_n$ conditional on the data used to estimate $\hat{X}$.

Appendix L. Application to social spillovers in movie consumption

This appendix reports the confidence-interval application to the data of Gilchrist and Sands (2016) that motivated the empirically calibrated simulation designs of Section 3.

The Gilchrist and Sands (2016) sample consists of 1671 opening weekend days between January 1, 2002 and January 1, 2012. For each opening weekend, the authors observe gross ticket sales for movies wide released in theaters in the United States with a run in theaters of at least six weeks.[21] The data are obtained through Box Office Mojo, a subsidiary of the Internet Movie Database (IMDb).

The outcome variables of interest are gross ticket sales of movies that opened in a given weekend in the second through sixth weeks of their run, while the endogenous variable is the gross ticket sales of a movie in its opening weekend. To control for seasonal

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

Table L.5
95% Confidence Intervals and Interval Lengths in the data of Gilchrist and Sands (2016).

Parameter	β_2	β_3	β_4	β_5	β_6	β_7
Estimate (s.e.)	0.475 (0.024)	0.269 (0.023)	0.164 (0.017)	0.121 (0.013)	0.093 (0.010)	1.222 (0.074)
	Initial instrument set, $d_z = 48$					
$JK(\beta_0)$	⟵0.114⟶ [0.441, 0.555]	⟵0.114⟶ [0.234, 0.348]	⟵0.074⟶ [0.127, 0.201]	⟵0.074⟶ [0.0936, 0.167]	⟵0.046⟶ [0.0736, 0.120]	⟵0.375⟶ [0.989, 1.365]
$S(\beta_0)$	∅	⟵0.033⟶ [0.294, 0.328]	∅	∅	∅	⟵0.561⟶ [0.989, 1.551]
JLM	⟵0.140⟶ [0.428, 0.569]	⟵0.127⟶ [0.221, 0.348]	⟵0.087⟶ [0.134, 0.221]	⟵0.074⟶ [0.100, 0.174]	⟵0.060⟶ [0.080, 0.140]	⟵0.441⟶ [0.989, 1.384]
	Temperature instruments only, $d_z = 36$					
$JK(\beta_0)$	⟵0.147⟶ [0.462, 0.609]	⟵0.134⟶ [0.268, 0.401]	⟵0.096⟶ [0.158, 0.254]	⟵0.080⟶ [0.114, 0.194]	⟵0.067⟶ [0.094, 0.161]	⟵0.455⟶ [1.117, 1.572]
$S(\beta_0)$	∅	∅	∅	∅	∅	∅
JLM	⟵0.161⟶ [0.448, 0.609]	⟵0.161⟶ [0.248, 0.408]	⟵0.107⟶ [0.147, 0.254]	⟵0.087⟶ [0.114, 0.200]	⟵0.067⟶ [0.094, 0.161]	⟵0.542⟶ [1.070, 1.612]
	Initial instruments plus all interactions with temp. instruments, $d_z = 524$					
$JK(\beta_0)$	⟵0.040⟶ [0.462, 0.502]	⟵0.033⟶ [0.268, 0.301]	⟵0.020⟶ [0.154, 0.174]	⟵0.013⟶ [0.094, 0.107]	⟵0.007⟶ [0.067, 0.074]	⟵0.120⟶ [1.043, 1.164]
$S(\beta_0)$	⟵0.047⟶ [0.415, 0.462]	∅	∅	⟵0.207⟶ [0.040, 0.247]	⟵0.060⟶ [0.161, 0.221]	∅
JLM	⟵0.080⟶ [0.441, 0.522]	⟵0.060⟶ [0.247, 0.308]	⟵0.040⟶ [0.147, 0.187]	⟵0.027⟶ [0.094, 0.120]	⟵0.013⟶ [0.067, 0.080]	⟵0.227⟶ [0.990, 1.217]

periodicity in both the supply of and demand for movies, a vector of date controls are included. Formally, the authors are interested in the parameters β_w, $w = 2, \dots, 7$ from the linear IV model(s):

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

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

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

Putting together the nationally aggregated weather measures leaves Gilchrist and Sands (2016) with 48 linearly independent instrumental variables.[22] To handle the large number of instruments, the authors employ a post-LASSO estimate of the first stage (Belloni et al., 2012); they set the first-stage penalty parameter so that the number of instrument selected is one, two, or three. As discussed in Section 3.1, the resulting first-stage F-statistics on the selected instrument(s) seem to indicate strong identification, while the first-stage F-statistic on the full set of instruments is only 3.80.

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.

I compare the 95% confidence intervals based on the jackknife K-statistic to those based on the jackknife Lagrange-Multiplier (JLM) statistic of Matsushita and Otsu (2022) and the sup-score statistic of Belloni et al. (2012). Confidence intervals based on the jackknife AR statistic of Mikusheva and Sun (2021) are not reported as they were empty for all specifications. Empty confidence intervals could indicate misspecification of the linear model, though they may also reflect finite-sample over-rejection of the jackknife AR test in settings, like this one, with a nonnegligible number of exogenous controls (see Table 3 of Mikusheva and Sun, 2024). Similarly, confidence intervals based on the thresholding statistic, implemented as recommended in Section 5, are also not reported as they

[22] There are 52 instruments in total, but four linearly dependent ones are ignored in the following.

always align with the those of jackknife K-statistic. The narrower confidence bands of the jackknife K-statistic in specifications where the sup-score confidence interval is non-empty indicates higher power from the jackknife K-test in this setting, so the combination test suggesting its use may be expected. The level of penalty $\lambda^\star$ from Section 2 evaluates to zero so that jackknife K-statistic is essentially implemented using the jackknife OLS hat matrix of Angrist et al. (1999).

Tables L.5 reports the 95% confidence intervals for $\beta_1, \dots, \beta_7$ generated by weak-instrument robust confidence intervals for three sets of instruments: the first is the initial set of 48 instruments in Gilchrist and Sands (2016), the second set includes only the temperature instruments for $d_z = 36$, and the final includes the initial instruments as well as all interactions between the temperature instruments and the remaining instruments for $d_z = 524$. For reference, I also provide point estimates and standard errors for $\beta_1, \dots, \beta_7$ 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.

Qualitatively, the results from the weak-instrument robust confidence intervals are similar to that of the author's original analysis; indeed the Gilchrist and Sands (2016) point estimates are always in the identification robust confidence intervals when using either the initial instrument set ($d_z = 48$) or the reduced instrument set ($d_z = 36$). However, when using the larger instrument set of $d_z = 524$ we obtain confidence bands using the jackknife K-test that rule out the author's initial point estimates for the parameters β_5, β_6 and β_7 and suggest somewhat smaller social spillover effects in movie consumption. Across all specifications and instrument sets, confidence intervals obtained from inverting the jackknife K-test are consistently smaller than those obtained from inverting both the sup-score and jackknife Lagrange Multiplier test. These reductions in confidence interval length are most noticeable when using the complete set of interactions, across all parameters the jackknife K-confidence intervals are nearly half the length of their jackknife Lagrange Multiplier counterparts. As with the jackknife Anderson-Rubin test, the sup-score confidence intervals are also often empty which again could suggest misspecification of the linear IV model.

Appendix M. Stylized simulation study

This appendix reports a simulation study on fully synthetic data that complements the empirically calibrated designs of Section 3. I examine the performance of tests based on the $JK(\beta_0)$ statistic and compare it with that of other tests that may be used in settings where the number of instruments is nonnegligible as a fraction of sample size. I consider a reduced-form data-generating process (DGP) similar to that of Matsushita and Otsu (2022). The outcome variable, y_i, and endogenous variable, x_i, are generated according to

$$\begin{aligned} y_i &= x_i + \epsilon_i \\ x_i &= \Pi_i + \upsilon_i \end{aligned} \tag{M.1}$$

where $\Pi_i = \frac{1}{r_n}\sum_{k=1}^{5} \frac{3}{4}\bar{z}_{ki} + \frac{1}{4}\bar{z}_{ki}^2 + \frac{1}{4}\bar{z}_{ki}^3$ is a (dense) transformation of an initial set of instruments $\bar{z}_i \in \mathbb{R}^{15}$ generated as described below. The value of r_n varies depending on the strength of identification considered; under weak identification, $r_n = n^{-1/2}$ while for power curves I consider an intermediate identification strength, $r_n = n^{-1/3}$.[23] To model heteroskedasticity, the errors (e_i, υ_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 $\upsilon_i = \varrho_2(1 + \bar{z}_{1i})\epsilon_i + (1 - \varrho_2)^2 e_{2i}$ where e_{1i} and e_{2i} are generated independently of each other and other variables in the model according to a Laplace distribution with location parameter $\mu = 0$ and scale parameter $b = 1$. Since the jackknife K-statistic has a nearly exact χ^2 distribution when the errors are jointly Gaussian, I purposefully avoid normally distributed errors to investigate the quality of asymptotic approximations. The parameters ϱ_1 and ϱ_2 control the degree of heteroskedasticity and endogeneity, respectively.

I examine the size of the test under three different instrument regimes. In all three regimes, I begin with an initial set of instruments $\bar{z}_i = (\bar{z}_{1i}, \dots, \bar{z}_{15i})'$ generated independently across indices according to a multivariate Gaussian distribution with Toeplitz covariance structure, $Cov(\bar{z}_{\ell i}, \bar{z}_{ki}) = 2^{-|\ell - k|}$. In the first regime, the instruments only include the initial set, $z_i = \bar{z}_i$ so that $d_z = 15$. In the second regime, the full set of instruments z_i additionally includes all quadratic and cubic terms, $(z_{\ell i}^2, z_{\ell i}^3)$, $\ell = 1, \dots, 15$ so that in total $d_z = 45$. In the third regime, the full set of instrument includes the initial set of instruments, $\bar{z}_i$, all quadratic and cubic terms (30 additional terms) and interactions of the initial set of instruments ($\binom{15}{2} = 105$ additional terms), so that in total $d_z = 150$. 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.

In constructing the jackknife K-statistic, I opt to use the deleted diagonal ridge matrix, $H = [h_{ij}]$ where $h_{ij} = [\boldsymbol{z}(\boldsymbol{z}'\boldsymbol{z} + \lambda I_{d_z})\boldsymbol{z}']_{ij}\mathbf{1}\{i \neq j\}$, instead of a true jackknife ridge procedure in order to make the simulations compuationally tractable. Following reccomendations in van Wieringen (2023), the penalty paramter λ is set so that effective degrees of freedom of the resulting hat matrix is no more than $n/5$.[24] To estimate the parameter $\rho(z_i)$, I implement the default cross-validated ℓ_1-penalized procedure of (2) via the `glmnet` package in `R` (Friedman et al., 2010). I use the full vector of instruments as the basis to approximate $\rho(z_i)$.

I compare the simulated size of the jackknife K test and to the performance of the sup-score test, $S(\beta_0)$, of Belloni et al. (2012), the thresholding test introduced in Section 5.2, 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)

[23] Intermediate identification strength is considered to let the power curves come up to one at the boundaries of the considered range of β values. Power curves with $r_n = n^{-1/2}$ look similar, but shrunk towards zero.

[24] Precisely, the penalty parameter is set $\lambda = \max 0, (n/5)^{-1} d_1^2(\boldsymbol{z})(d_z - n/5)$, where $d_1^2(\boldsymbol{z})$ is the square of the maximum singular value of the design matrix $\boldsymbol{z}$.

Table M.6
Simulated Size of Identification and Heteroskedasticity Robust Tests under Weak Identification. Each DGP is simulated 5000 times.

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	15	0.2	0.3	0.0514	0.0356	0.0482	0.0454	0.0234	0.0454	0.0434
		0.2	0.6	0.0500	0.0376	0.0460	0.0412	0.0258	0.0728	0.0436
		0.5	0.3	0.0466	0.0384	0.0430	0.0402	0.0238	0.0784	0.0450
		0.5	0.6	0.0454	0.032	0.0432	0.0394	0.0220	0.0734	0.0458
	45	0.2	0.3	0.0430	0.0102	0.0372	0.0268	0.0062	0.0930	0.0386
		0.2	0.6	0.0422	0.0102	0.0406	0.0302	0.0078	0.0890	0.0414
		0.5	0.3	0.0446	0.0104	0.0372	0.0242	0.0074	0.1058	0.0306
		0.5	0.6	0.0452	0.0110	0.0414	0.0308	0.0040	0.1052	0.0342
	150	0.2	0.3	0.0490	0.0044	0.0416	0.0242	0.0000	0.1066	0.0460
		0.2	0.6	0.0480	0.0068	0.0422	0.0288	0.0000	0.1074	0.0408
		0.5	0.3	0.0482	0.0060	0.0424	0.0244	0.0000	0.1070	0.0458
		0.5	0.6	0.0434	0.0070	0.0404	0.0268	0.0000	0.1120	0.0414
500	15	0.2	0.3	0.0540	0.0448	0.0510	0.0490	0.0374	0.0702	0.0512
		0.2	0.6	0.0516	0.0424	0.0478	0.0488	0.0368	0.0674	0.0444
		0.5	0.3	0.0474	0.0398	0.0452	0.0466	0.0294	0.0690	0.0488
		0.5	0.6	0.0490	0.0392	0.0466	0.0464	0.0320	0.0718	0.0446
	45	0.2	0.3	0.0554	0.0196	0.0480	0.0364	0.0198	0.0840	0.0340
		0.2	0.6	0.0496	0.0206	0.0456	0.0392	0.0202	0.0812	0.0378
		0.5	0.4	0.0552	0.0192	0.0514	0.0368	0.0166	0.0904	0.0330
		0.5	0.6	0.0518	0.0224	0.0472	0.0346	0.0188	0.0950	0.0328
	150	0.2	0.3	0.0476	0.0168	0.0456	0.0380	0.0044	0.0754	0.0432
		0.2	0.6	0.0456	0.0146	0.0428	0.0386	0.0036	0.0730	0.0426
		0.5	0.3	0.0540	0.0116	0.0486	0.0332	0.0052	0.0856	0.0380
		0.5	0.6	0.0456	0.0180	0.0436	0.0364	0.0036	0.0784	0.0418

of Matsushita and Otsu (2022). Critical values of the sup-score and conditioning statistic are simulated with the procedures described in Section 5 with 1000 bootstrap replications. For the combination test cutoff, I report results using two different quantiles of the conditioning statistic under the assumption that $\mathbb{E}[\hat{\Pi}_i^I] = 0$ for all $i \in [n]$; $\tau_{0.3}$ corresponding to the 30^{th} quantile and $\tau_{0.75}$ corresponding to the 75^{th} quantile.

Table M.6 reports 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, both the jackknife AR and jackknife LM test all exhibit moderate size distortions in various regimes. The jackknife AR test in particular appears to overreject in nearly all setups considered. This property also observed in the simulation study of Matsushita and Otsu (2022) and so may be driven by the similarity of this simulation setup to theirs. The jackknife LM statistic appears to have good size properties when $d_z = 10$ and $d_z = 150$ but is consistently (though only moderately) undersized in the intermediate regime where $d_z = 45$. This size distortion does not improve when moving from $n = 200$ to $n = 500$, suggesting that the requirement that $d_z \to \infty$ is important for the quality of finite-sample approximation by its limiting distribution. Though the good performance of the jackknife LM statistic when $d_z = 10$ is notable, it should also be remarked that this is the setup with the least amount of correlation between the instruments.

The sup-score and Anderson-Rubin test seem to be undersized in all regimes, with the Anderson-Rubin test nearly never rejecting when $d_z = 150$. However, and in line with the theory, both of their size properties appear to improve when increasing the sample size from $n = 200$ to $n = 500$. It is possible that the size properties of the sup-score test could be improved by using an empirical bootstrap based approach to simulate the critical value, as proposed by Deng and Zhang (2020), however I do not consider that approach here. The threshholding test appears to inherit the conservative nature of the sup-score test, though to a lesser degree due to the combination with the $JK(\beta_0)$ test.

In addition to examining the size properties of the given tests I also investigate the power properties of the tests in this setting. Figs. M.4 and M.5 plot calibrated local power curves under an intermediate identification strength where the first stage is in a $n^{-1/3}$ neighborhood of zero for $n \in \{200, 500\}$, the number of instrument is plausible large, $d_z = 150$, $\varrho_1 \in \{0.2, 0.5\}$ and $\varrho_2 \in \{0.3, 0.6\}$. The critical value of each test is set to simulated 95^{th} quantile of the distribution of the corresponding test-statistic under H_0. I compare the calibrated local power curves of the $JK(\beta_0)$ test, the combination test with cutoff $\tau_{0.75}$, the jackknife AR test, the Jackknife LM test, and the sup-score test.[25]

[25] For both the jackknife AR and jackknife LM tests, I use cross-fit estimates of test statistic variances proposed and shown to improve power by Mikusheva and Sun (2021).

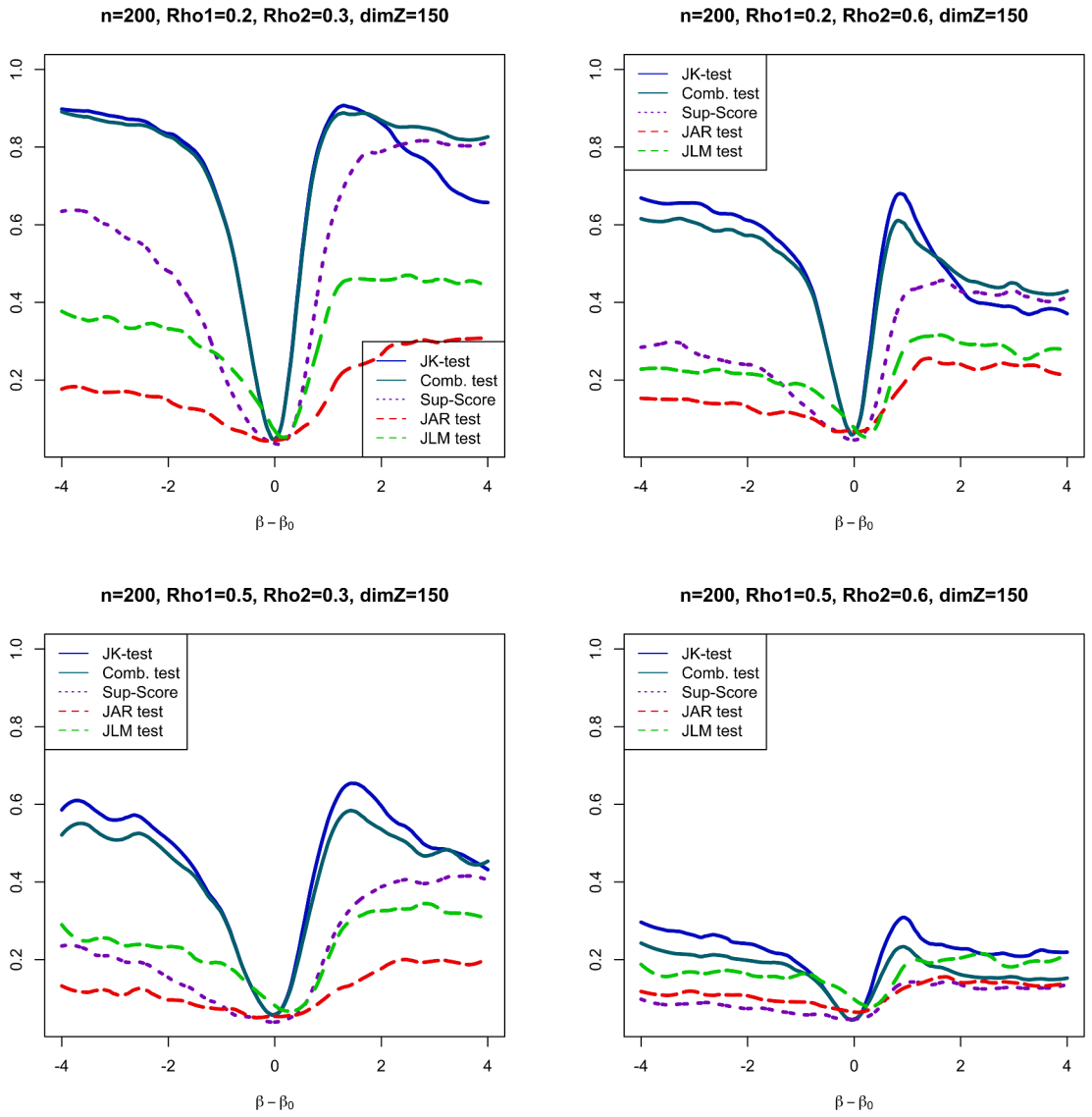


Fig. M.4. Calibrated Local Power Curves under Intermediate Identification Strength and 150 Instruments. Sample size is 200 and rejection probability is calculated on a grid of 100 $(\beta_0 - \beta)$ points between -4 and 4. At each point the DGP is simulated 1000 times.

In all setups considered, the jackknife K and combination tests have substantially stronger power than the jackknife AR, jackknife LM, and sup-score tests in local neighborhoods of the null as well as for negative values of $(\beta_0 - \beta)$. For values of $(\beta_0 - \beta)$ larger than 1.5, tests based on the jackknife K-statistic suffer from a loss of power as described in Section 5. This power decline is largerly ameliorated by combining the jackknife K-statistic with the sup-score statistic and the thresholding test has good power properties over all alternatives considered. However, tests based on the jackknife AR or jackknife LM statistic can still provide better power than the threshholding test for very positive values of $(\beta_0 - \beta)$ in some setups.

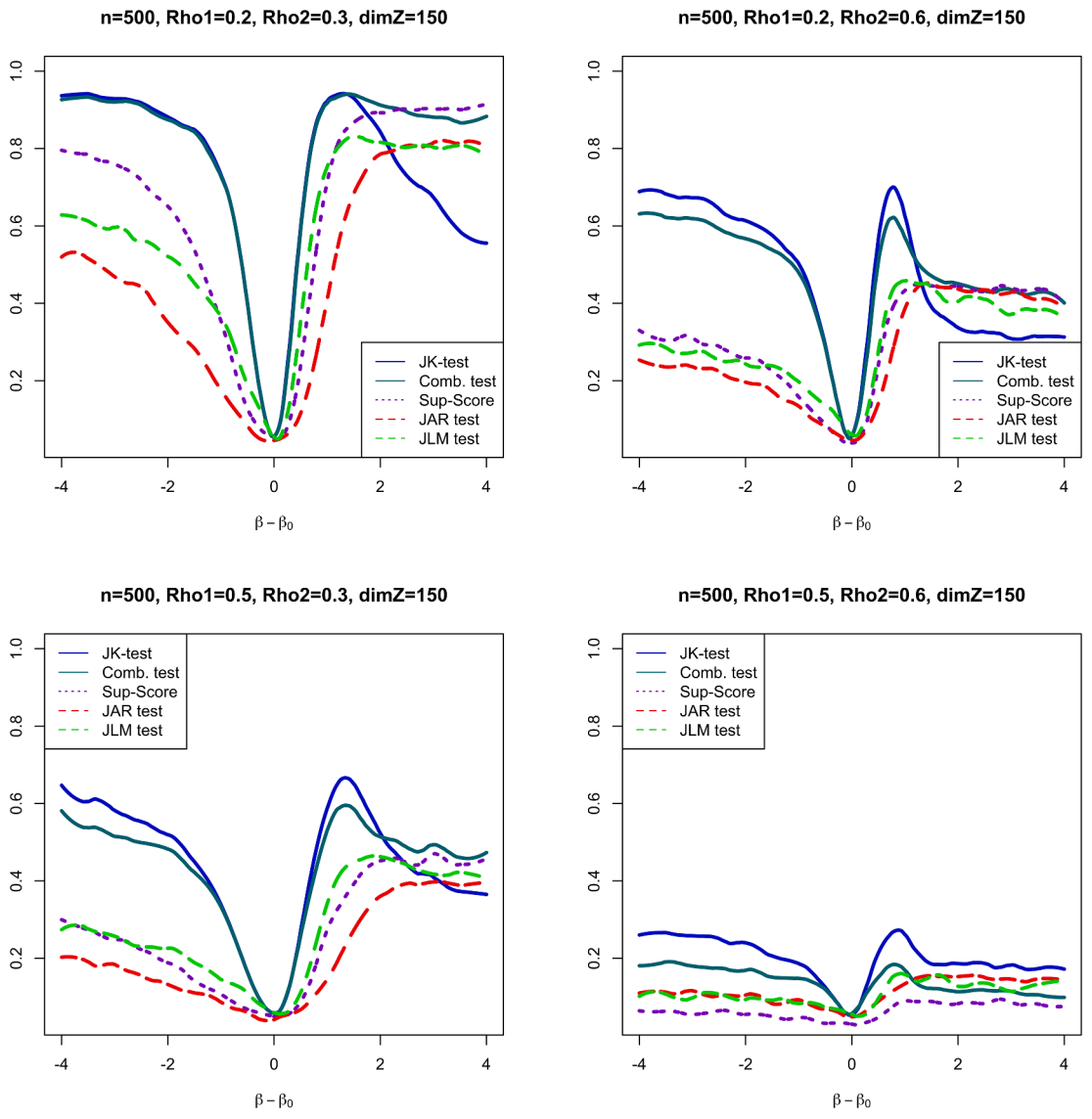


Fig. M.5. Calibrated Local Power Curves under Intermediate identification Strength and 150 Instruments. Sample size is 500 and rejection power is calculated on a grid of 100 ($\beta_0 - \beta$) points between -4 and 4. At each point the DGP is simulated 2000 times.

Appendix N. Additional calibrated simulation results

This appendix reports two additional sets of results for the empirically calibrated simulation study of Section 3: results for a fourth instrument set, and power results under the dense specification of the nuisance parameter $\rho(z_i)$.

The weather-rank instrument set. Because the weather instruments in the Gilchrist and Sands (2016) data are shared by the three days of each opening weekend, any weather-only dictionary has rank at most 556. A fourth instrument set augments the initial instruments with their squares and cubes and *all* pairwise interactions, attaining this rank cap with $d_z = 556$ ($d_z/n \approx 0.33$). Table N.7 reports simulated size and average calibrated power for this instrument set under both error designs of Section 3.2. The results are nearly identical to those for the expanded instrument set ($d_z = 522$) reported in the main text: growing a weather-only dictionary past the expanded set changes neither the size nor the power comparisons for any test. This is why the calendar instrument set, which breaks the rank cap by interacting dictionary terms with within-weekend day indicators, is needed to study the high-dimensional regime.

Table N.7
Simulated size at the 5% nominal level and average calibrated rejection probability for the weather-rank instrument set ($d_z = 556$, $d_z/n \approx 0.33$) under the approximately sparse and dense specifications of the nuisance parameter. 1000 simulations per design; the power grids span $\beta_0 - \beta \in [-0.50, 0.50]$ and $[-0.45, 0.45]$ under the sparse and dense designs, respectively.

	Size		Avg. Calibrated Power	
Testing Procedure	Sparse $\rho(z_i)$	Dense $\rho(z_i)$	Sparse $\rho(z_i)$	Dense $\rho(z_i)$
$JK(\beta_0)$	0.045	0.051	0.565	0.536
$T(\beta_0; \tau_{0.3})$	0.043	0.050	0.563	0.534
$S(\beta_0)$	0.008	0.006	0.160	0.072
JAR	0.051	0.049	0.120	0.102
JAR_{CF}	0.089	0.081	0.089	0.084
JLM	0.055	0.046	0.307	0.354
JLM_{CF}	0.058	0.050	0.321	0.363
LWZ	0.048	0.057	0.186	0.119

Power under the dense nuisance design. The power results in Section 3.3 are computed under the approximately sparse design, in line with the conditions that motivate the LASSO estimate of $\rho(z_i)$. Table N.8 reports the corresponding average calibrated rejection probabilities under the dense design, where $\rho(z_i)$ loads on every instrument and the LASSO estimate can at best recover a sparse approximation. The comparisons in the main text carry over: the power advantage of the jackknife K-test over the many-instrument tests again grows with d_z/n, and moving from the sparse to the dense design costs the jackknife K-test essentially nothing (average calibrated power with the calendar instrument set of 0.642 sparse versus 0.630 dense). The sup-score test loses noticeably more power under the dense design. Fig. N.6 overlays the calibrated power curves of the two designs for each instrument set for a subset of the tests.

Table N.8
Average calibrated rejection probability over the non-null grid points in the empirically calibrated design with a dense nuisance parameter, by instrument set. Power grids span $\beta_0 - \beta \in [-0.30, 0.30]$, $[-0.45, 0.45]$, and $[-0.70, 0.70]$ for the initial, expanded, and calendar instrument sets, respectively.

Testing Procedure	Initial $d_z = 48$ $d_z/n \approx 0.03$	Expanded $d_z = 522$ $d_z/n \approx 0.31$	Calendar $d_z = 1{,}182$ $d_z/n \approx 0.71$
$JK(\beta_0)$	0.533	0.563	0.630
$T(\beta_0; \tau_{0.3})$	0.533	0.534	0.599
$S(\beta_0)$	0.130	0.083	0.150
JAR	0.192	0.105	0.086
JAR_{CF}	0.187	0.082	0.085
JLM	0.529	0.333	0.136
JLM_{CF}	0.529	0.360	0.181
LWZ	0.218	0.193	0.365

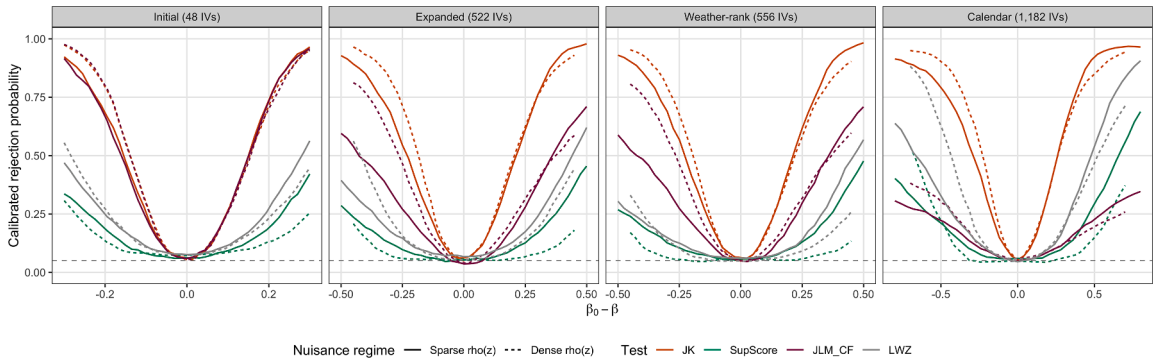


Fig. N.6. Calibrated power curves for the $JK(\beta_0)$, sup-score, cross-fit JLM, and LWZ tests under the sparse (solid) and dense (dashed) nuisance designs, by instrument set.

References

Anderson, T.W., Rubin, H., 1949. Estimation of the parameters of a single equation in a complete system of stochastic equations. Ann. Math. Stat. 20 (1), 46–63. https://doi.org/10.1214/aoms/1177730090

Andrews, D. W.K., Moreira, M., Stock, J.H., 2004. Optimal Invariant Similar Tests for Instrumental Variables Regression. Working Paper 299. National Bureau of Economic Research. https://doi.org/10.3386/t0299

Andrews, D. W.K., Moreira, M.J., Stock, J.H., 2006. Optimal two-sided invariant similar tests for instrumental variables regression. Econometrica 74 (3), 715–752. http://www.jstor.org/stable/4123100

Andrews, D. W.K., Stock, J.H., 2007. Testing with many weak instruments. J. Econom. 138 (1), 24–46. 50th Anniversary Econometric Institute. https://doi.org/10.1016/j.jeconom.2006.05.012

Andrews, I., 2016. Conditional linear combination tests for weakly identified models. Econometrica 84 (6), 2155–2182. https://doi.org/10.3982/ECTA12407

Andrews, I., Stock, J.H., Sun, L., 2019. Weak instruments in instrumental variables regression: theory and practice. Annu. Rev. Econ. 11, 727–753. https://doi.org/10.1146/annurev-economics-080218-025643

Angrist, J.D., Frandsen, B., 2022. Machine labor. J. Labor Econ. 40 (S1), S97–S140.

Angrist, J.D., Imbens, G.W., Krueger, A.B., 1999. Jackknife instrumental variables estimation. J. Appl. Econom. 14 (1), 57–67. http://www.jstor.org/stable/223249

Angrist, J.D., Krueger, A.B., 1991. Does compulsory school attendance affect schooling and earnings? Q. J. Econ. 106 (4), 979–1014. http://www.jstor.org/stable/2937954

Ayyar, S., Matsushita, Y., Otsu, T., 2025. Conditional likelihood ratio test with many weak instruments. Econom. Theory, First View, 42 (3), 604–630.

Belloni, A., Chen, D., Chernozhukov, V., Hansen, C., 2012. Sparse models and methods for optimal instruments with an application to eminent domain. Econometrica 80 (6), 2369–2429. https://doi.org/10.3982/ECTA9626

Belloni, A., Chernozhukov, V., Chetverikov, D., Hansen, C., Kato, K., 2018. High-dimensional econometrics and regularized GMM. arXiv:1806.01888

Celentano, M., Montanari, A., Wu, Y., 2020. The estimation error of general first order methods. In: Abernethy, J., Agarwal, S. (Eds.), Proceedings of Thirty Third Conference on Learning Theory. PMLR, pp. 1078–1141. https://proceedings.mlr.press/v125/celentano20a.html

Chamberlain, G., 1987. Asymptotic efficiency in estimation with conditional moment restrictions. J. Econom. 34 (3), 305–334. https://www.sciencedirect.com/science/article/pii/0304407687900157. https://doi.org/10.1016/0304-4076(87)90015-7

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

Chatterjee, S., 2006. A generalization of the Lindeberg principle. Ann. Probab. 34 (6), 2061 – 2076. https://doi.org/10.1214/009117906000000575

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., Robins, J., 2018. Double/debiased machine learning for treatment and structural parameters. Econom. J. 21 (1), C1–C68. https://academic.oup.com/ectj/article-pdf/21/1/C1/27684918/ectj00c1.pdf, https://doi.org/10.1111/ectj.12097

Chernozhukov, V., Chetverikov, D., Kato, K., 2013. Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. Ann. Stat. 41 (6), 2786 – 2819. https://doi.org/10.1214/13-AOS1161

Chernozhukov, V., Chetverikov, D., Kato, K., 2017. Central limit theorems and bootstrap in high dimensions. Ann. Probab. 45 (4), 2309–2352. http://www.jstor.org/stable/26362255

Chetverikov, D., Sørensen, J. R.-V., 2021. Analytic and bootstrap-after-cross-validation methods for selecting penalty parameters of high-dimensional m-estimators. ArXiv NA, 1–50. arXiv:2104.04716

Chi, C.-M., Vossler, P., Fan, Y., Lv, J., 2022. Asymptotic properties of high-dimensional random forests. Ann. Stat. 50 (6), 3415–3438.

Crudu, F., Mellace, G., Sándor, Z., 2021. Inference in instrumental variable models with heteroskedasticity and many instruments. Econom. Theory 37 (2), 281-310. https://doi.org/10.1017/S026646662000016X

Deng, H., Zhang, C.-H., 2020. Beyond Gaussian approximation. Ann. Stat. 48 (6), 3643–3671.

Derenoncourt, E., 2022. Can you move to opportunity? evidence from the great migration. Am. Econ. Rev. 112 (2), 369–408. https://doi.org/10.1257/aer.20200002

Doví, M.-S., 2025. Robust inference with high-dimensional controls and weak instruments. Preprint.

Doví, M.-S., Kock, A.B., and, S.M., 2024. A ridge-regularized jackknifed anderson-rubin test. J. Bus. Econ. Stat. 42 (3), 1083–1094. https://doi.org/10.1080/07350015.2023.2290739

Farrell, M.H., Liang, T., Misra, S., 2021. Deep neural networks for estimation and inference. Econometrica 89 (1), 181–213. https://doi.org/10.3982/ECTA16901

Friedman, J., Tibshirani, R., Hastie, T., 2010. Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. 33 (1), 1–22. https://doi.org/10.18637/jss.v033.i01

Gilchrist, D.S., Sands, E.G., 2016. Something to talk about: social spillovers in movie consumption. J. Polit. Econ. 124 (5), 1339–1382. https://doi.org/10.1086/688177

Götze, F., Naumov, A., Spokoiny, V., Ulyanov, V., 2019. Large ball probabilities, Gaussian comparison and anti-concentration. Bernoulli 25 (4A), 2538 – 2563. https://doi.org/10.3150/18-BEJ1062

Gotze, F., Sambale, H., Sinulis, A., 2021. Concentration inequalities for polynomials in alpha-sub-exponential random variables. Electron. J. Probab. 26 (none), 1 – 22. https://doi.org/10.1214/21-EJP606

van der Greer, S., 2016. Estimation and Testing under Sparsity. Lecture Notes in Mathematics, Springer, New York, NY. https://doi.org/10.1007/978-3-319-32774-7

Horn, R.A., Johnson, C.R., 2012. Matrix Analysis. Cambridge University Press. https://books.google.com/books?id=O7sgAwAAQBAJ

Kleibergen, F., 2002. Pivotal statistics for testing structural parameters in instrumental variables regression. Econometrica 70, 1781–1803. https://doi.org/10.1111/1468-0262.00353

Kleibergen, F., 2005. Testing parameters in GMM without assuming that they are identified. Econometrica 73 (4), 1103–1123. https://doi.org/10.1111/j.1468-0262.2005.00610.x
Kleibergen, F., Zhan, Z., 2025. Double robust inference for continuous updating GMM. Quant. Econ. 16 (1), 295–327.
Lim, D., Wang, W., Zhang, Y., 2022. A conditional linear combination test with many weak instruments. arXiv:2207.11137
Lim, D., Wang, W., Zhang, Y., 2024. A dimension-agnostic bootstrap anderson-rubin test for instrumental variable regressions. arXiv:2412.01603
Lindeberg, J.W., 1922. Eine neue herleitung des exponentialgesetzes in der wahrscheinlichkeitsrechnung. Math. Z. 15, 211–225.
Matsushita, Y., Otsu, T., 2022. A jackknife lagrange multiplier test with many weak instruments. Econometr. Theory, First View. https://doi.org/10.1017/S0266466622000433
Mikusheva, A., 2023. Many weak instruments in time series econometrics. Working paper.
Mikusheva, A., Sun, L., 2021. Inference with many weak instruments. Rev. Econ. Stud. 89 (5), 2663–2686. https://doi.org/10.1093/restud/rdab097
Mikusheva, A., Sun, L., 2024. Weak identification with many instruments. Econom. J. 27 (2), C1–C28.
Moreira, M., 2009. Tests with correct size when instruments can be arbitrarily weak. J. Econom. 152, 131–140. https://doi.org/10.1016/j.jeconom.2009.01.012
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. https://doi.org/10.1111/1468-0262.00438
Nazarov, F., 2003. On the maximal perimeter of a convex set in r^n with respect to a gaussian measure. In: Milman, V.D., Schechtman, G. (Eds.), Geometric Aspects of Functional Analysis: Israel Seminar 2001-2002. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 169–187. https://doi.org/10.1007/978-3-540-36428-3-15
Newey, W.K., Windmeijer, F., 2009. Generalized method of moments with many weak moment conditions. Econometrica 77 (3), 687–719. https://doi.org/10.3982/ECTA6224
Nyquist, H., 1988. Applications of the jackknife procedure in ridge regression. Comput. Stat. Data Anal. 6 (2), 177–183.
Paravisini, D., Rappoport, V., Schnabl, P., Wolfenzon, D., 2014. Dissecting the effect of credit supply on trade: evidence from matched credit-export data. Rev. Econ. Stud. 82 (1), 333–359. https://doi.org/10.1093/restud/rdu028
Peña, V.H., Lai, T.L., Shao, Q.M., 2008. Self-Normalized Processes: Limit Theory and Statistical Applications. Probability and Its Applications, Springer Berlin Heidelberg. https://books.google.com/books?id=NzyzvtxZdioC
Petersen, K.B., Pedersen, M.S., 2012. The matrix cookbook. Version 20121115. http://www2.compute.dtu.dk/pubdb/pubs/3274-full.html
Pouzo, D., 2015. Bootstrap consistency for quadratic forms of sample averages with increasing dimension. Electron. J. Stat. 9 (2), 3046 – 3097. https://doi.org/10.1214/15-EJS1090
Sambale, H., 2022. Some notes on concentration for α-subexponential random variables. arXiv:2002.10761
Schmidt-Hieber, J., 2020. Nonparametric regression using deep neural networks with reLU activation function. Ann. Stat. 48, 1875–1897. https://doi.org/10.1214/19-AOS1875
Staiger, D., Stock, J.H., 1997. Instrumental variables regression with weak instruments. Econometrica 65 (3), 557–586. http://www.jstor.org/stable/2171753
Tan, Z., 2017. Regularized calibrated estimation of propensity scores with model misspecification and high-dimensional data. ArXiv NA, 1–60. arXiv:1710.08074
van Wieringen, W.N., 2023. Lecture notes on ridge regression. arXiv:1509.09169
Yap, L., 2025. Inference with many weak instruments and heterogeneity. Working paper.