An Identification and Dimensionality Robust Test for Instrumental Variables Models

Manu Navjeevan*

University of California, Los Angeles

Revised: August 29, 2023

Abstract

I propose a new identification robust test for the structural parameter in a heteroskedastic linear instrumental variables model with an arbitrary number of instruments. The test is based on a jackknife version of the K-statistic and has asymptotically correct size so long as an auxilary parameter can be consistently estimated. This is possible under approximate sparsity even when the number of instruments is potentially very large. The limiting χ^2 distribution of the test statistic is derived using a novel direct gaussian approximation technique; under weak identification the numerator and denominator of the test statistic may have non-negligible distributions however neither is required to converge individually to a weak limit. To avoid a power decline faced by the jackknife K-statistic, I propose a simple combination with the sup-score statistic of Belloni et. al (2012) based on a thresholding rule.

Keywords: Instrumental Variables, Weak Identification, High Dimensional

JEL Codes: C12, C36, C55

*Email: mnavjeevan@g.ucla.edu. Thanks...

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 identification robust testing procedures all require some conditions on the rate of growth of the number of instruments, d_z, in relation to the sample size, n. To control size under heteroskedasticity, the tests of [1, 23, 24, 31, 32, 39] are shown by [4] to to require $d_z^3/n \to 0$. Recent tests devloped [17, 27, 29] allow the number of instruments to be proportional to sample size, $d_z/n \to \phi \in [0, 1)$, but require that the number of instruments be large, $d_z \to \infty$. And while the tests proposed by [8, 28] require only $\log^M(d_z)/n \to 0$ for some constant $M > 1$, they either rely on sample splitting or have a critical value that is increasing with the number of instruments, lending them suboptimal power properties in overidentified models. These conditions on the growth rate of d_z can be difficult to interpret and the variety of tests available under alternate regimes may make it unclear to the researcher which test should be applied in their particular setting.[1]

I introduce a new identification robust test for heteroskedastic linear IV models with an arbitrary number of instruments. The proposed test statistic is similar in spirit to a jackknife version of the K-statistic and makes use of an auxilary parameter to construct first stage estimates that are uncorrelated with the structural errors under the null hypothesis. So long as this auxiliary parameter can be consistently estimated, the proposed test statistic has a limiting χ^2 distribution with a degrees of freedom equal to the number of structural parameters. The auxiliary parameter is simple to estimate using out of the box methods and consistency is achievable under approximate sparsity even when the number of instruments is potentially much larger than sample size. This approximate sparsity assumption is trivially satisfied where errors are homoskedastic.

The limiting behavior of the test statistic is examined via a direct gaussian approximation method. In local neighborhoods of the null I show that quantiles of the test statistic can be uniformly approximated by quantiles of an analog statistic that replaces each observation in the expression of the test statistic with a gaussian version that has the same mean and covariance matrix as the original. These local neighborhoods of the null are characterized by a local power index which I introduce in Section 3. This direct approach has the benefit of not requiring the conditions of any particular central limit theorem to hold, which is crucial for relaxing assumptions on the growth rate of d_z.

In the case of a single endogenous variable I show that, under an additional regularity condition, a version of the test that could be constructed if the auxiliary parameter was known to the researcher is consistent whenever the local power index diverges. When the local power index is bounded, I examine the limiting power of the test by examining the behavior of the analog statistic. Under the alternative hypothesis the analog statistic has a nearly non-central χ^2 distribution conditional on the first stage estimates. The noncentrality parameter is proportional to the correlation between the true first stage model and the first stage estimates. Unfortunately, the process of partialling out the structural parameter may introduce bias into the first stage estimates under the alternate hypothesis. This issue is pointed out by [2, 5, 31] in the context of the non-jackknife K-statistic. Against certain alternatives this bias can be particularly pronounced and the additional regularity condition needed for the consistency result may fail to hold.

To address this, I propose a simple combination of the Jackknife K-statistic with the sup-score

[1] For example, consider a setting similar to that of [18] where the researcher has a dozen or so instruments and a sample size of a few hundred. The number of instruments cubed is larger than sample size, but asymptotic approximations based on $d_z \to \infty$ seem unlikely to resemble the finite sample distribution.

statistic [8] based on a thresholding rule. Similarly to the Anderson-Rubin statistic [2,5,39], while the sup-score statistic is not effecient in general, it does not suffer from a loss of power against any particular alternative. The combination test decides whether to run the Jackknife K-test or the sup-score test by comparing the value of a conditioning statistic to a predetermined cutoff value. In the approximating gaussian regime, this conditioning statistic is marginally independent of both the Jackknife K-statistic and the sup-score statistic. This allows me to show that the thresholding test controls size under the null without having to require that the conditioning statistic converges in distribution to a stable limit. In simulations I find that taking this cutoff value to be the 50th quantile of the distribution of conditioning statistic leads to optimal power. Using results in [9,14] this quantile can be simulated via a multiplier bootstrap procedure.

The direct gaussian approximation argument is a variation of Lindeberg's interpolation method [26]. A direct application of original method would consist of bounding each one step deviation that results from replacing an observation with its gaussian version in the expression of the test statistic. Using this direct approach, however, is complicated by the fact that under sufficiently weak identification the denominator of the Jackknife K-statistic may be arbitrarily close to zero with positive probability. This allows derivatives with respect to terms in the denominator to be arbitrarily large and makes bounding the one step deviations impractical.

When there is a single endogenous variable, a leading case in empirical applications, this issue can be successfully sidestepped by taking advantage of the particular form of the Jackknife K-statistic in this setting. With multiple endogenous variables a more involved argument is needed that requires stronger moment conditions. Both cases require a particular anticoncentration condition on the denominator of the Jackknife K-statistic that I establish using new bounds on densities of gaussian quadratic forms from [19]. An interesting feature of the direct gaussian approximation argument is that while the numerator and denominator the Jackknife K-statistic may both have non trivial distributions under weak identification, neither needs to converge on its own to a weak limit in order to characterize the limiting behavior of the Jackknife K-statistic.

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

Notation. For any $n \in \mathbb{N}$ let $[n]$ denote the set $\{1, \dots, n\}$. I work with a sequence of probability measures P_n on the data $\{(y_i, x_i, z_i) : i \in [n]\}$ where (y_i, x_i, z_i) are variables introduced below. Let $\mathbb{E}_n[f_i] = n^{-1} \sum_{i=1}^n f_i$ denote the empirical expectation and $\bar{\mathbb{E}}[f] = \mathbb{E}[\mathbb{E}_n[f_i]]$ denote the average expectation operator.

1.1 Prior Literature

When the first stage F-statistic is fall, standard asymptotic approximations may fail to accurately describe the finite sample behavior of IV estimates. This was first pointed out by [10,34] who consider the finite sample behavior of the 2SLS in alternate settings where the instrumental variable is only weakly correlated with the endogenous variable. In a seminal paper, [39] capture this phenomena in an asymptotic framework by considering a sequence of first stage models that shrink to zero with the sample size. Under this framework, standard IV estimates are no longer consistent and inference procedures based on these statistics fail to control size.

Due to this, there has been a large interest in developing tests for the structural parameter that control size regardless of identification strength. In their original weak IV paper, [39] propose the use of the Anderson-Rubin statistic which is identification robust. Noting that the Anderson-Rubin test is inefficient in overidentified models, [23, 24, 31] propose the use of the (non-jackknife) K-statistic, which has a limiting null distribution that does not depend on the number of instruments. The conditional likelihood ratio statistic of [30,32] can be viewed as a combination of the K-statistic and Anderson Rubin statistic based on a conditioning statistic that is independent of them both under the null. [2] characterize the power envelope in a homoskedastic weakly identified IV model and show that the test based on the conditional likelihood statistic has nearly optimal power in this setting. When errors are heteroskedastic [5] proposes alternate combinations of the K-statistic and AR-statistic based on a minimax regret criterion. To determine whether the Wald Test or identification robust tests should be used, [40] propose pretesting for the strength of identification based on the first stage F-statistic. Their recommendation for using the standard Wald Test whenever the first stage F-statistic exceeds 10 is based on a model with a single endogenous variable and homoskedastic errors; [25] point out this recommendation is not applicable in heteroskedastic models and update the recommended F-statistic cutoff.

The initial results in [2,23,24,31,32,39] treat the number of instruments as fixed. [21,35] study the behavior of the Anderson-Rubin, K, and conditional likelihood ratio statistics when the number of instruments is growing with sample size. They show these tests are valid under arbitrarily weak identification and heteroskedasticity whenever $d_z^3/n \to 0$, as discussed briefly above. Recent papers [17,27,29] have taken advantage of a new central limit theorem for quadratic forms developed in [12] and proposed weak identification robust tests that are valid even when the number of instruments is proportional to sample size; $d_z/n \to \phi \in [0, 1)$. Following the many instruments asymptotic framework first introduced by [7], the analyses in these papers rely on the number of instruments diverging. Indeed, the convergence rate of the proposed test statistics under the null is at the square root of the number of instruments. A benefit of the approaches in these papers compared to the approach in this paper is that they do not require any nuisance parameters to be estimated. To pretest for weak identification in this asymptotic framework, [29] propose a new $\tilde{F}$ statistic and suggest using identification robust procedures when $\tilde{F} < 4.14$.

Limited identification robust testing procedures exist for the high dimensional case, $d_z \gg n$. To my knowledge, the only two options available are the sup-score test of [8] and the split sample optimal instrument AR test reviewed in [28]. The sup-score test makes use of gaussian approximations for maxima of high dimensional vectors developed in [13] but suffers from the same problem as the Anderson-Rubin test in that its critical value is increasing with the number of instruments. The spilt sample optimal instrument AR test uses one half of the sample to estimate an optimal instrument and the other half to test the null hypothesis. Identification robust tsting procedures in the high dimensional case are of particular interest due to the lack of clarity on how to pretest for weak identification in these settings. In particular, F-statistics resulting from first stage LASSO or Post-LASSO procedures are unlikely to be interpretable.

I contribute to these literatures by proposing a new test for the structural parameter that can work in any of the settings discussed above.

2 Model and Setup

Consider a linear instrumental variables model

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

where the researcher observes the outcome $y_i \in \mathbb{R}$, the endogenous variable $x_i \in \mathbb{R}^{d_x}$, and a set of instruments $z_i \in \mathbb{R}^{d_z}$, but not the structural error $\varepsilon_i \in \mathbb{R}$ nor the first stage errors $v_i \in \mathbb{R}^{d_x}$. The structural error is assumed to be conditional mean independent of the instruments, $\mathbb{E}[\varepsilon_i|z_i] = 0$. I denote $\Pi_i \in \mathbb{R}^{d_x}$ as $\Pi_i = \mathbb{E}[x_i|z_i]$ and make no assumptions about the functional form of Π_i so the instruments are allowed to affect the endogenous variable in a nonlinear fashion.

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

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

As $\Omega(z_i)$ is otherwise left unrestricted, the errors are allowed to be heteroskedastic. There are no controls in the model as they are assumed to be low-dimensional and thus, without loss of generality, partialled out. All results in this paper hold conditionally on a realization of the instruments $z := (z_1', \ldots, z_n') \in \mathbb{R}^{n \times d_z}$ so from this point forth they are treated as fixed.

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

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

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

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

Each element of the nuisance parameter $\rho(z_i)$, $\rho_\ell(z_i)$ for $\ell = 1, \ldots, d_x$, can be interpreted as the (conditional) slope coeffecient from a simple linear regression of $x_{\ell i}$ on $\epsilon_i(\beta_0)$. As such, if $\rho_\ell(\cdot)$ falls in some function class Φ it can be estimated directly [15] under H_0 by solving empirical analogs of:[1]

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

I will largely work under the assumption that $\rho(z_i)$ has an approximately sparse representation in some (growing) basis $b(z_i) := (b_1(z_i), \ldots, b_{d_b}(z_i))' \in \mathbb{R}^{d_b}$, that is $\rho_\ell(z_i) = b(z_i)'\gamma_\ell + \xi_{\ell i}$ where $\xi_{\ell i}$ represents an approximation error that tends to zero with the sample size and γ_ℓ is sparse in the

[1] Under H_1, $\rho_\ell(z_i)$ can be estimated directly by solving emprical 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.

sense that many of its coeffecients are zero. This allows for nesting the low dimensional case, where the number of instruments is fixed, and the high dimensional case, where the number of instruments is potentially much larger than sample size, under a unified estimation procedure. Under homoskedasticity, $\rho_\ell(z_i)$ is a constant function and thus has an approximately spare representation in any basis that contains an intercept. The parameter γ can be estimated via LASSO

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

or via post-LASSO, refitting an unpenalized version of (2.2) using only the basis terms associated with non-zero coeffecients in the first stage LASSO. The estimating procedure in (2.2) is a simple ℓ_1-penalized regression of $x_{\ell i}$ against $\epsilon_i(\beta_0)b(z_i)$. It can be easily implemented using out of the box software available on most platforms. Under standard conditions, this leads to a consistent estimate of $\rho_\ell(z_i)$ so long under the sparsity condition $s^2 \log^M(d_b n)/n \to 0$ where s is the number of non-zero elements of γ_ℓ and M is a positive constant that depends on the moment bounds imposed. The estimation procedure is discussed in more detail in Section 3.3.

2.1 Test Statistic

With this setup, I introduce the test-statistic. The test statistic is based on an arbitrary jackknife-linear estimate of the first stage,

$$\widehat{\Pi}_{\ell i} = \sum_{j \neq i} h_{ij}\hat{r}_{\ell j}, \quad \ell = 1, \ldots, d_x$$

for some hat matrix $H = [h_{ij}] \in \mathbb{R}^{n \times n}$ that has zeros on the diagonal, $h_{ii} = 0$ for all $i = 1, \ldots, n$, and which may only depend on the instruments z.[2] Formally, the only structure I require on the hat matrix H is a balanced design condition described in Section 3. However, for reasons explained in Section 4 it may be optimal to introduce some regularization when estimating the first stage models $\widehat{\Pi}_{\ell i}$ so I suggest using the deleted diagonal ridge-regression hat matrix $H(\lambda^\star)$:

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

where following recommendations in [22,43] the penalty parameter $\lambda^\star$ is set so that the effective degrees of freedom of the resulting hat matrix is no more than a fraction of sample size:

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

The ridge hat matrix has the benefit of being well defined even when the number of instruments is larger than the sample size. The eigenvalues of the ridge-hat matrix are always less weakly than one, suggesting that the matrix stays well behaved when the number of instruments is large, in the sense that $\sum_{j \neq i} h_{ij}^2 \leq 1$. I stress though, that the researcher may use any other hat matrix that they believe will lead to plausible first stage estimates. Other possible choices of hat matrix include the jackknife OLS hat matrix [6], the deleted diagonal projection matrix introduced in [12] and succesfully used in [17,27,29], or hat matrices based on selecting instruments via some prelminary unsupervised technique such as PCA. Remark 3.1 discusses how the balanced design condition

[2] This hat matrix can differ for each $\ell = 1, \ldots, d_x$, but the notation becomes messy. This extension is explored in the Appendix.

may be verified for arbitrary choices of hat matrices.

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

$$\begin{aligned} \varepsilon(\beta_0) &= (\varepsilon_1(\beta_0), \dots, \varepsilon_n(\beta_0))' \in \mathbb{R}^n \\ \widehat{\Pi} &= (\widehat{\Pi}_1', \dots, \widehat{\Pi}_n')' \in \mathbb{R}^{n \times d_x} \\ \widehat{\Pi}_\varepsilon &= (\widehat{\Pi}_{\epsilon 1}', \dots, \widehat{\Pi}_{\epsilon n}')' \in \mathbb{R}^{n \times d_x} \end{aligned} \tag{2.3}$$

The Jackknife K-statistic can then be defined

$$JK(\beta_0) = \epsilon(\beta_0)\widehat{\Pi}(\widehat{\Pi}_\epsilon'\widehat{\Pi}_\epsilon)^{-1}\widehat{\Pi}'\epsilon(\beta_0) \times \mathbf{1}\{\lambda_{\min}(\widehat{\Pi}_\epsilon'\widehat{\Pi}_\epsilon) > 0\} \tag{2.4}$$

Under appropriate conditions on the hat matrix H and moment bounds, I will show that the limiting distribution of $JK(\beta_0)$ under H_0 is $\chi^2_{d_x}$. For exposition, I will largely focus on the case where $d_x = 1$ in which case the form of the test statistic simplifies. The extension to $d_x > 1$ is not immediate but is possible under strengthened moment conditions and is explored in Section 5.

3 Limiting Behavior with a Single Endogenous Variable

The limiting behavior of the test statistic is analyzed via a direct gaussian approximation technique. When there is a single endogenous variable this approach can be considerably simplified. In this section I detail the approach and take advantage of the simplified analysis to characterize the limiting behavior of the test statistic under local alternatives to H_0. This direct approach has the advantage of not relying on any assumptions on the dimensionality of the instruments nor strength of identification.

The goal will be to show that quantiles of $JK(\beta_0)$ can be approximated by the corresponding quantiles of the gaussian statistic,

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

where $\tilde{\Pi}_i = \sum_{j \neq i} h_{ij}\tilde{r}_j$ and for any $i \in [n]$ the random vectors $(\tilde{\epsilon}_i(\beta_0), \tilde{r}_i)'$ are generated independently of each other and the data following a gaussian distribution with the same mean and covariance as $(\epsilon_i(\beta_0), r_i)'$. Since uncorrelated gaussian random variables are independent, under H_0, the vector $(\tilde{\epsilon}_1(\beta_0), \dots, \tilde{\epsilon}_n(\beta_0))'$ is mean zero and independent of $(\tilde{r}_1, \dots, \tilde{r}_n)'$. The null distribution of $JK_G(\beta_0)$ conditional on any realization of $(\tilde{r}_1, \dots, \tilde{r}_n)'$ is then χ^2_1 and so its unconditional null distribution is also χ^2_1.

3.1 Interpolation Approach

Error arising from estimation of $\rho(z_i)$ prevents immediate comparison of the distribution of $JK(\beta_0)$ to the distribution of $JK_G(\beta_0)$. So, to begin, consider the distribution of an infeasible statistic, $JK_I(\beta_0)$, which could be constructed if $\rho(z_i)$ was known to the researcher:

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

where $\widehat{\Pi}_i^I = \sum_{j \neq i} h_{ij} r_j$. To show that the distribution of $JK_I(\beta_0)$ can be approximated by the distribution of $JK_G(\beta_0)$, I adapt Lindeberg's interpolation method, first introduced by [26] in an elegant proof of the central limit theorem. Applying the interpolation method directly on the statistics $JK_I(\beta_0)$ and $JK_G(\beta_0)$, however, is not tractable, as it requires bounding expectations of derivatives with respect to terms in the denominator. When identification is weak, the denominators of $JK_I(\beta_0)$ and $JK_G(\beta_0)$ may both be arbitrarily close to zero with postive probability. Without imposing irregular conditions, derivatives with respect to terms in the denominators thus may not generally have finite expectations.

Instead, I consider a different approach. For a scaling factor s_n, introduced below, define the scaled numerators and denominators

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

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

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

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

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

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

where $L_2(\varphi) := \sup_x |\varphi''(x)|$ and $L_3(\varphi) := \sup_x |\varphi'''(x)|$. By taking $\varphi(\cdot)$ to be a sequence of functions approximating the indicator function, $\mathbf{1}\{x \leq 0\}$, the result in (3.2) can be used to show that quantiles of the infeasible statistic $JK_I(\beta_0)$ can be (uniformly) approximated by quantiles of the gaussian statistic $JK_G(\beta_0)$. Figure 3.1 provides an illustration of this approach.

The Lindeberg interpolation argument on the decomposed test statistics makes use of the fact that the numerator and denominator of the gaussian test statistic are functions of quadratic forms in the random vectors $\epsilon(\beta_0) := (\epsilon_1(\beta_0), \ldots, \epsilon_n(\beta_0))'$ and $r := (r_1, \ldots, r_n)'$. See [37] for another example of the Lindeberg interpolation method applied to approximate the distribution of quadratic forms.

The remainder of the argument relies on showing a particular anticoncentration bound on $\tilde{D}$. I show that that this can be established under either weak or strong identification. This allows for the limiting null distribution of the test statistic under various identification regimes to be derived via a unifying argument.[1] Additionally, even though $(N, D, \tilde{N}, \tilde{D})$ may all have non-negligible distributions when identification is weak, the interpolation argument does not require any of these to individually converge in distribution or probability anywhere stable. This allows for a wide range of possible hat matrices H to be used when constructing the first stage estimates,

[1] This contrasts to the analysis in [23] which considers the behavior of the K-statistic under three alternate regimes.

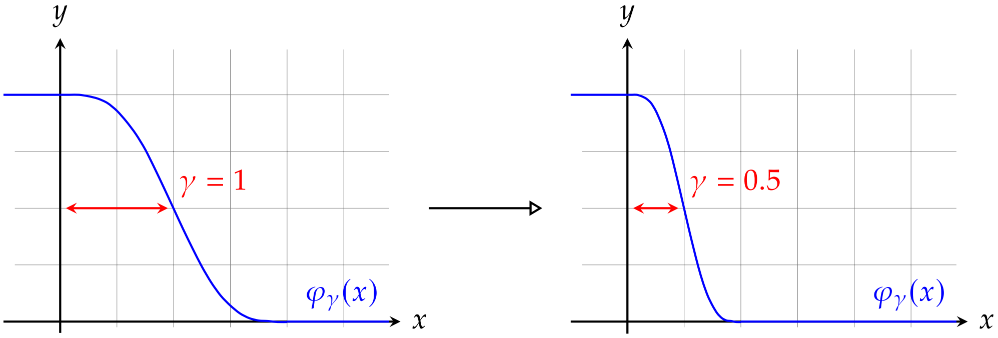

Figure 3.1: The functions $\varphi_\gamma(x)$ approximate the indicator $\mathbf{1}\{x \geq 0\}$ from above. The quality of approximation increases as the parameter γ tends to zero, but the second and third derivatives increase on the order of γ^{-2} and γ^{-3}, respectively.

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

3.2 Limiting Behavior without Estimation Error

The interpolation argument relies on certain assumptions, which are detailed below. Define $\eta_i := (\beta - \beta_0)v_i + \epsilon_i$ and $\zeta_i := v_i - \rho(z_i)\eta_i$, noting $\eta_i = \epsilon_i(\beta_0) - \mathbb{E}[\epsilon_i(\beta_0)]$ and $\zeta_i = r_i - \mathbb{E}[r_i]$. Under H_0, $\eta_i = \epsilon_i$ while $\zeta_i = v_i - \rho(z_i)\epsilon_i$. For each $i \in [n]$, define the infeasible first stage estimator that could be constructed if $\rho(z_i)$ was known to the researcher, $\widehat{\Pi}_i^I := \sum_{j \neq i} h_{ij} r_j$. Assumptions 3.1 and 3.2 allow for characterization of the null distribution of $JK(\beta_0)$. In what comes below $c > 1$ can be considered an arbitrary constant that may be updated upon each use but that does not depend on sample size n.

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

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

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

Assumption 3.2(i) requires that the average second moment of the infeasible first stage estimators is on the same order as the maximum first stage estimator second moment. This is imposed mainly to rule out hat matrices that are all zeroes or nearly all zeros so that the effective number of observations used to test the null is growing with the sample size. Assumption 3.2(ii) requires that the maximum leverage of any observation is bounded. When H is symmetric it is automatically satisfied under Assumption 3.1(i) and the definition of s_n. Remark 3.1 below discusses how both of these may be verified in practice. The scaling factor s_n captures both the "size" of elements in the hat matrix H as well as the strength of identification. If elements of the hat matrix are on the same order as a constant one would expect $s_n = O(n^{-1})$ under strong identification ($\Pi_i \propto 1$) while $s_n = O(n^{-1/2})$ under weak identification ($\Pi_i \lesssim n^{-1/2}$).

Assumption 3.2(iii) can be viewed as a technical requirement that there is more than one "effective" instrument in the hat matrix. In the case of a standard projection matrix (no deleted diagonal), Assumption 3.2(iii) is satisfied whenever $\text{rank}(z(z'z)^{-1}z) > 1$. This condition can be easily verified in practice by examining the eigenvalues of HH'.

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

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

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

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. Assumption 3.3, below, formally defines the local neighborhoods considered.

Assumption 3.3 (Local Identification). *(i) The local power index P is bounded, $P \leq c$.*

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

(ii) $\max_i \mathbb{E}[(s_n \sum_{j\neq i} h_{ji}\epsilon_j(\beta_0))^2] \leq c$.

Under H_0, Assumption 3.3 is trivally satisfied since $(\beta - \beta_0) = 0$ and $\sum_{j\neq i} s_n^2 h_{ji}^2 \leq c$. The local power index is the second moment of the scaled numerator, N and is a measure of the association between the true first stage Π_i and the first stage estimates $\widehat{\Pi}_i$. In Section 4 I discuss how the strength of this association is related to the power of the test under local alternatives. Proposition 3.1 below shows that when Assumption 3.3(ii) holds, $P \to \infty$ implies that the test based on the infeasible statistic $JK_I(\beta_0)$ is consistent.

Assumption 3.3(ii) is an additional technical condition that requires that the maximum value of $\mathbb{E}[(\sum_{j\neq i} h_{ji}\epsilon_j(\beta_0))^2]$ is on the same or lesser order then the maximum value of $\mathbb{E}[(\sum_{j\neq i} h_{ij} r_j)^2]$. Using the moment bounds in Assumption 3.1 and Assumption 3.2(ii) one can verify that Assumption 3.3(ii) is equivalent to the existence of constants $C_1, C_2 > 0$ such that

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

for all $i \in [n]$. It is always satisfied whenever $\mathbb{E}[\epsilon_i(\beta_0)] = \Pi_i(\beta - \beta_0)$ is in a $\sqrt{n}$-neighborhood of zero in the sense that $|\Pi_i(\beta - \beta_0)| \leq C/\sqrt{n}$ for all $i \in [n]$ and some constant C. In general, Assumption 3.3(ii) can be roughly interpreted as requiring the local neighborhoods of H_0 considered to be those in which the means of $(\epsilon_1(\beta_0), \ldots, \epsilon_n(\beta_0))$ are on the same or lesser order than that the means of $(r_1, \ldots, r_n)$.

Under Assumptions 3.1–3.3 the behavior of the infeasible statistic $JK_I(\beta_0)$ can be approximated by the behavior of the gaussian statistic $JK_G(\beta_0)$. Again, this result does not require any stable limiting distribution for $JK_G(\beta_0)$.

Theorem 3.1 (Infeasible Local Power)**.** *Suppose that Assumptions 3.1–3.3 hold. Then*

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

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

Proposition 3.1 (Consistency)**.** *Suppose that Assumptions 3.1, 3.2, and 3.3(ii) hold. Then if $P \to \infty$ the test based on $JK_I(\beta_0)$ is consistent; i.e for any fixed $a \in \mathbb{R}$, $\Pr(JK_I(\beta_0) \le a) \to 0$.*

The dependence of the consistency result on Assumption 3.3(ii) is a non-trivial restriction due to the bias taken on in constructing r_i. In particular, against certain alternatives it is possible that $\mathbb{E}[\widehat{\Pi}_i^I] = 0$ for all $i \in [n]$ even under strong identification. This is an extreme case, however. In general bias in $\mathbb{E}[r_i]$ does not imply a violation of Assumption 3.3(ii), which requires only that the *size* of $\mathbb{E}[r_i]$ is of a weakly greater order than that of $\mathbb{E}[\epsilon_i(\beta_0)]$.

Moreover, Proposition 3.1 does not necessarily rule out that a test based on $JK_I(\beta_0)$ is consistent when $P \to \infty$ but Assumption 3.3(ii) fails to hold. The proof of Proposition 3.1 relies on showing that, when $P \to \infty$ and Assumption 3.3(ii) holds, $\mathbb{E}[|N|] \to \infty$ while $\mathrm{Var}(|N|)$ and $\mathbb{E}[D]$ are bounded. These facts can be combined to show that $\Pr(N^2 - aD \le 0) \to 0$ for any fixed $a \in \mathbb{R}$. When Assumption 3.3(ii) fails, $P \to \infty$ may imply that $\mathrm{Var}(|N|) \to \infty$ as well, making the limiting behavior of the test difficult to analyze. There is reason to believe that this issue can be overcome; [3] show that the K-statistic of [23] is consistent against fixed alternatives under strong identification. However, a full consistency result is not pursued here and left to future work.

Regardless of whether the test is consistent, bias taken on in constructing r_i has consequences on the power of the test in finite samples, particularly when the mean of r_i is of a lesser order than that of $\epsilon_i(\beta_0)$. This is discussed in more detail in Section 4. To rectify this deficiency in tests based on the Jackknife K-statistic, I suggest a thresholding test that decides whether or not to use the Jackknife K-statistic or the sup-score [8] statistic based on the value of conditioning statistic. This conditioning statistic in turn is based on a test-statistic for the null hypothesis that $\mathbb{E}[\widehat{\Pi}_i^I] = 0$ for all $i \in [n]$.

3.3 Controlling Estimation Error

The final step is to show that the difference between the infeasible and feasible statistics is negligible. I begin with a technical lemma that states that the difference between $JK(\beta_0)$ and $JK_I(\beta_0)$ can be treated as negligible so long as the difference between the scaled numerators and the scaled denominators can be treated as negligible. Define these differences

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

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

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

While Lemma 3.1 is a simple statement, it is not obvious. In particular, showing that the difference between the infeasible and feasible statistics is negligible requires showing that $1/(D + \Delta_D)$ is bounded in probability, where D represents the scaled denominator of $JK_I(\beta_0)$. Under a typical analysis, this would be done by arguing that D converges in distribution to a stable limit and then applying continuous mapping theorem.[2] This approach is not applicable here as neither the numerator nor the denominator need converge in distribution individually.

Instead, I directly show that $1/(D + \Delta_D)$ is bounded in probability by showing $\Pr(D \le \delta_n) \to 0$ for any sequence $\delta_n \to 0$. This is done by first showing that quantiles of D can be approximated by quantiles of $\tilde{D}$, the scaled denominator of $JK_G(\beta_0)$. If the variance of $\tilde{D}$ is bounded away from zero, its density can also be bounded using new bounds on gaussian quadratic form densities from [19], which yields the result. Otherwise, if $\mathrm{Var}(\tilde{D}) \to 0$, the result holds by an application of Chebyshev's inequality and $\mathbb{E}[D] > c^{-1}$ from Assumption 3.2(i). This particular anticoncentration bound for $\tilde{D}$ is also important in the proof of Theorem 3.1 to establish anticoncentration for the decomposed gaussian test statistic.

Lemma 3.1 allows the researcher to use alternate choices of estimators of $\rho(z_i)$, so long as they can verify that $(\Delta_N, \Delta_D)' \to_p 0$. Below, I verify that this condition can be satisfied for the ℓ_1-penalized estimation procedure proposed in (2.2). This requires a strengthened moment condition on η_i. Given a random variable X and $v > 0$ the Orlicz (quasi)-norm is defined

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

Random variables with a finite Orlicz norm for some $v \in (0,1] \cup \{2\}$ are termed α-sub-exponential random variables [20,38]. This class encompasses a wide range of potential distributions including all bounded and sub-gaussian random random variables (with $v = 2$), all sub-exponential random variables such as Poisson or non-central χ^2 random variables (with $v = 1$), as well as random variables with fatter tails such as Weibull distributed random variables with shape parameter $v \in (0,1]$.

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

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

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

Assumption 3.4(i) strengthens the moment condition on η_i to require that η_i is in the class of α-sub-exponential random variables. While this condition is more restrictive than the moment condition in Assumption 3.1, as discussed above, it still allows for a wide range of potential distributions. Assumption 3.4(ii) is a standard condition in ℓ_1-penalized estimation and can be enforced in practice

[2] This is the approach taken by [23,24]

by normalizing the basis terms to be between zero and one.[3] Assumption 3.4(iii) is a bound on the rate of decay of the approximation error, similar to the approximate sparsity condition of [8].

Assumption 3.4(iv) is a high level condition on the rate of consistency of the parameter estimate $\hat{\gamma}$ in the ℓ_1 norm. This can be verified under sparsity for both the LASSO estimator in (2.2) or Post-LASSO procedures based on refitting an unpenalized version of (2.2) only using the basis terms selected in a LASSO first stage. See [8, 16, 41, 42] 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 fucntion of sample size. Following analysis in [41] we can see that, under appropriate choice of penalty parameter, this may be satisfied so long as $s^2 \log^{2(v+1)/v}(d_b n)/n \to 0$ where the sparsity index s denotes the number of non-zero elements of γ.

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

Under Assumptions 3.1–3.4 I establish that the difference between the infeasible and feasible statistics can be treated as negligible when using the estimation procedure proposed in (2.2).

Lemma 3.2. *Suppose that Assumptions 3.1–3.4 hold. Then* $(\Delta_N, \Delta_D)' \to_p 0$.

Theorem 3.1 and Lemma 3.2 are combined for the main result, local approximation of the distribution of the feasible test statistic $JK(\beta_0)$ by the distribution of the gaussian statistic $JK_G(\beta_0)$.An immediate corollary of this result is that the limiting null distribution of $JK(\beta_0)$ is χ^2_1.

Theorem 3.2 (Local Power)**.** *Suppose that Assumptions 3.1–3.4 hold. Then*

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

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

If the limiting $JK_G(\beta_0)$ had a stable distribution, Theorem 3.2 would follow immediately from Theorem 3.1, Lemmas 3.1 and 3.2, and an application of Slutsky's Lemma. However, under H_1, there is nothing preventing the distribution of $JK_G(\beta_0)$ changing with the sample size. Instead I establish Theorem 3.2 directly using the fact that both $JK(\beta_0)$ and $JK_G(\beta_0)$ are bounded in probability and that $JK_G(\beta_0)$ has a density that is bounded uniformly over n.

Despite the fact that $JK_G(\beta_0)$ may not have a stable limiting distribution, examining its behavior is much more tractable than examining the behavior of $JK(\beta_0)$. Thus, Theorem 3.2 still allows for analysis of the local power properties of the proposed test.

4 Improving Power against Certain Alternatives

Using the characterization of the limiting behavior of the test statistic derived in Section 3, I analyze the local power properties of the test. Unfortunately, against certain alternatives the test statistic

[3] This is a standard normalization and in practice typically leads to better performance of the LASSO estimator.

may have trivial power, a deficiency shared with the K-statistics of [23, 24]. To combat this, I propose a simple combination with the sup-score statistic of [8] based on a thresholding rule.

4.1 Local Power Properties

Under local alternatives to H_0, as defined in Assumption 3.3, Theorems 3.1 and 3.2 suggest that the limiting behavior of $JK(\beta_0)$ can be analyzed by examining the behavior of $JK_G(\beta_0)$. Conditional on the vector $\tilde{r} = (\tilde{r}_1, \ldots, \tilde{r}_n)$ the gaussian statistic $JK_G(\beta_0)$ is distributed nearly non-central χ^2_1 with non-centrality parameter $\mu(\tilde{r})$, $JK_G(\beta_0)|\tilde{r} \sim A^2(\tilde{r}) \cdot \chi^2_1(\mu(\tilde{r}))$ where

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

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

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

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

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 [11]. 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 use of ℓ_2-regularization when 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 [31], this is not be an issue as long as there is a fixed constant $C \neq 0$ such that $\mathbb{E}[r_i] = C\Pi_i$ for all $i \in [n]$. However, in general this will introduce bias into the first stage estimates $\widehat{\Pi}_i$ under H_1. This particularly pronounced when $\rho(z_i)$ is a constant $(\beta-\beta_0) = 1/\rho(z_i)$. In this case, $\mathbb{E}[r_i]$, and thus $\mathbb{E}[\tilde{\Pi}_i]$, will equal zero for each $i \in [n]$ and the $JK(\beta_0)$ statistic will select a direction completely at random to direct power into.[1]

4.2 A Simple Combination Test

To combat this loss of power for tests based on the K-statistic, a common strategy is to combine the K-statistic with the AR-statistic based on a conditioning statistic. While the AR-statistic does not have optimal power on its own, it has the benefit of directing power equally in all directions avoiding the pitfalls of the K-statistic which lacks power in certain directions. Prominent examples of such tests are the conditional likelihood ratio test of [32], the GMM-M test of [24], and the minimax regret tests of [5]. These combinations make use of the fact that the AR-statistic is asymptotically independent of both the K-statistic and the conditioning statistic.

[1] [2, 5] point out this deficiency in the context of the K-statistic.

Unfortunately, the asymptotic validity of these tests under heteroskedasticity is based on the assumption that $d_z^3/n \to 0$ which may be difficult to verify in practice. Instead to combat the power deficiency of tests based on jackknife K-statistic, I consider a simple combination with the sup-score statistic of [8], which is similar in spirit to AR-statistic but controls size even when d_z grows near exponentially as a function of sample size. The sup-score statistic is

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

And the size θ test based on the sup-score statistic rejects whenever $S(\beta_0) > \tilde{c}_{1-\theta}^S$ where for $e_1, \dots, e_n \overset{\text{iid}}{\sim} N(0,1)$ generated independently of the data and any $\theta \in (0,1)$, $c_{1-\theta}^S$ is the simulated multiplier bootstrap critical value:

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

Similarly to tests based on the AR-statistic, tests based on the sup-score statistic may have suboptimal power properties in overidentified models since the critical value is increasing the number of instruments. However, the sup-score statistic does retain the benefit of directing power evenly in all directions avoiding the pitfall of tests based on $JK(\beta_0)$ against certain alternatives.

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

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

Depending on the value of the conditioning statistic, the thresholding test decides whether or not to run the test based on $JK(\beta_0)$ or one based on $S(\beta_0)$.

$$T(\beta_0; \tau) = \begin{cases} \mathbf{1}\{JK(\beta_0) > \chi^2_{1;1-\alpha}\} & \text{if } C \geq \tau \\ \mathbf{1}\{S(\beta_0) > c^S_{1-\alpha}\} & \text{if } C < \tau \end{cases} \tag{4.4}$$

for some cutoff τ which I take in practice to be the 50th quantile of the distribution of D.

As mentioned by [5] in the context of the standard K-statistic, this attempt to rectify the power deficiency via this particular conditioning statistic is not perfect. In particular, under heteroskedasticity, the means of the partialled out endogenous variables, $\mathbb{E}[r_i]$, may not be scaled versions of the true first stages. However, so long as $\mathbb{E}[r_i] \neq 0$, one can still expect $\mathbb{E}[\widehat{\Pi}_i^I] = \sum_{j\neq i} h_{ij}\Pi_i + (\beta - \beta_0)\sum_{j\neq i} h_{ij}\rho(z_i)\Pi_i$ to be related to the true fist stage Π_i and for the test to have non-trivial power. Moreover, in light of the dependence of the consistency result in Proposition 3.1 on Assumption 3.3(ii), in the case where $\mathbb{E}[\widehat{\Pi}_i] = 0$ for all $i \in [n]$ it may be particularly important to avoid using the Jackknife K-statistic to test H_0.

To show that the thresholding test controls size, I compare the rejection probability to that of a gaussian analog. In addition to $JK_G(\beta_0)$ defined in (3.1) define the gaussian analogs of $S(\beta_0)$ and the conditioning statistic C:

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

where as in Section 3, $(\tilde{\epsilon}_i(\beta_0), \tilde{r}_i)'$ are independent and generated independent of the data gaussian with the same mean and covariance matrix as $(\epsilon_i(\beta_0), r_i)$. Since $\mathrm{Cov}(\tilde{\epsilon}_i(\beta_0), \tilde{r}_i) = 0$ under H_0, the statistics C_G and $S_G(\beta_0)$ are independent under the null. Similarly the null distribution of $JK_G(\beta_0)$ is the same conditional on any realization of $(\tilde{r}_1, \ldots, r_n)$, it is also independent of C_G under the null. The gaussian analog thresholding test decides whether to run a test based on $S_G(\beta_0)$ or $JK_G(\beta_0)$ depending on the value of C_G as in (4.4).

Notice that $JK_G(\beta_0)$ and $S_G(\beta_0)$ are only marginally independent of the conditioning statistic C_G under the null. This limits the ways in which the test statistics can be combined using the conditioning statistic while still controlling size. This marginal independence in the gaussian limit is enough, however, for the asymptotic validity of the thresholding test, $T(\beta_0; \tau)$. To establish that the behavior of the pairs $(C, JK(\beta_0))$ and $(C, S(\beta_0))$ can be approximated by the behavior of $(C_G, JK_G(\beta_0))$ and $(C_G, S_G(\beta_0))$, respectively, I rely on the following assumption:

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

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

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

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

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

In particular, since $(JK_G(\beta_0) \perp C_G)$ *and* $(S_G(\beta_0) \perp C_G)$ *under* H_0 *the test based on* $T(\beta_0; \tau)$ *has asymptotic size* α *for any choice of cutoff* τ*.*

Theorem 4.1 establishes the asymptotic validity of the thresholding test $T(\beta_0; \tau)$ for any choice of cutoff τ. In practice, it seems reasonable to choose τ to be some intermediate quantile of the distribution of C_G under the assumption that $\mathbb{E}[\widehat{\Pi}_i^I] = 0$ for all $i \in [n]$. Following [9, 14] this can be approximated via a multiplier bootstrap procedure. Let $e_1, \ldots, e_n \overset{\text{iid}}{\sim} N(0,1)$ be generated independently of the data and for any $\theta \in (0,1)$, define the conditional quantile

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

Proposition 4.1 establishes the validity of the multiplier bootstrap to approximate quantiles of the conditioning statistic. It follows directly from results in [9] after verifying that the conditions needed for error taken on from estimation of $\rho(z_i)$ can treated as negligible under Assumption 3.4.

Proposition 4.1 ([9, Theorem 2.2])**.** *Suppose that Assumptions 3.1, 3.4 and 4.1 hold and that* $\mathbb{E}[\Pi^I_i] = 0$ *for all* $i \in [n]$*. Then there are sequences* $\delta_n \searrow 0$ *and* $\beta_n \searrow 0$ *such that with probability* $1-\delta_n$

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

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

5 Analysis with Multiple Endogeneous Variables

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

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

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

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

As in Section 3 notice that, since uncorrelated random variables are independent, under H_0 the vector $\tilde{\epsilon}(\beta_0)$ is mean zero and independent of $(\tilde{\Pi}, \tilde{\Pi}_\epsilon)$. Conditional on any realization of $(\tilde{\Pi}, \tilde{\Pi}_\epsilon)$ the $JK_G(\beta_0)$ statistic then follows a $\chi^2_{d_x}$ distribution and thus its unconditional distribution is also $\chi^2_{d_x}$.

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

5.1 Modified Interpolation Approach

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

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

for $\widehat{\Pi}^I$ and $\widehat{\Pi}^I_\epsilon$ defined the same way as $\widehat{\Pi}$ and $\widehat{\Pi}_\epsilon$ (2.3), respectively, but using the true values $(r_1, \ldots, r_n)'$ in place of their estimates $(\hat{r}_1, \ldots, \hat{r}_n)'$.

When there are multiple endogenous variables, $d_x > 1$, I cannot take advantage of the simplified form of the test statistic to establish this approximation as in Section 3. Instead I deal directly with the test statistics themselves. As in Figure 3.1 consider functions $\varphi_\gamma(\cdot) \in C_b^3(\mathbb{R})$ that approximate the indicators $\mathbf{1}\{\cdot \leq a\}$, where $a \in \mathbb{R}$ is arbitrary and γ is a scaling factor inversely proportional to the quality of the approximation but positively proportional to the derivatives of φ_γ. Broadly speaking, the goal is to show, for a sequence γ_n tending to zero, that

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

A standard interpolation argument would then attempt to show (5.1) by one-by-one replacing each pair, $(\epsilon_i(\beta_0), r_i)'$, in the expression of $\varphi_{\gamma_n}(JK_I(\beta_0))$ with its gaussian analog $(\tilde{\epsilon}_i(\beta_0), \tilde{r}_i)'$ and bounding the size of each of these deviations. As mentioned in Section 3, the problem arises as the derivative of the test statistic, $JK_I(\beta_0)$, with respect to terms in the denominator matrix, $\widehat{\Pi}_\epsilon'\widehat{\Pi}_\epsilon$, may be as large as the inverse of the minimum eigenvalue of the denominator maximum. When identification is sufficiently weak, the denominator matrix will have a non-negligible distribution and the inverse of its minimum eigenvalue will not be bounded and may not have finite moments under regular conditions.

To get around this problem, I modify the argument by considering a dynamic choice of approximation parameter γ_n. This dynamic choice of approximation parameter inversely scales with the determinant of the denominator matrix and thus, since the determinant is the product of the eigenvalues, inversely scales with the minimum eigenvalue.[1] Geometrically, this approach can be thought of as "stretching out" the function $\varphi_{\gamma_n}(\cdot)$ in directions where the minimum eigenvalue of the denominator matrix is close to zero. Since the overall derivatives of $\varphi_{\gamma_n}(JK_I(\beta_0))$ with respect to $(\epsilon_i(\beta_0), r_i)'$ depends on the product of derivatives with respect to the test statistic and derivatives of $\varphi_{\gamma_n}(\cdot)$, which scale inversely with the approximation parameter, this adjustment of the approximation parameter allows for control of the overall derivative. Details of this approach can be found in Appendix D.

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

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

Assumption 5.2 (Balanced Design). *(i) For any $\ell = 1, \ldots, d_x$ let $s_{\ell,n}^{-2} = \max_{1 \leq i \leq n} \mathbb{E}[(\widehat{\Pi}^I_{\ell i})^2]$ then the minimum eigenvalue of the following matrix is bounded away from zero*

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

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

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

$\lambda_k(HH')$ represents the k^{th} largest eigenvalue of the matrix HH'.

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

Assumption 5.2(i) is a natural extension of Assumption 3.2(i) to the setting where $d_x > 1$. It requires that the average second moment of any linear combination of the first stage estimates is proportional to the maximum second moment of the same linear combination. Assumption 5.2(ii,iii) are the same condition as Assumption 3.2(ii,iii) and can again be implicitly thought of as requiring that there are more than two effective instruments in the hat matrix. Assumption 5.2 thus reduces to Assumption 3.2 when $d_x = 1$.

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

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

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

Theorem 5.1 (Infeasible Local Power). *Suppose that Assumptions 5.1–5.3 hold. Then*

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

5.2 Controlling Estimation Error

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

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

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

So long as these scaled differences tend to zero, Lemma 5.1 shows that the difference between the feasible and infeasible test statistics converges to zero:

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

As with Lemma 3.1, while Lemma 5.1 is a simple statement, it is not immediate. In particular, establishing Lemma 5.1 requires showing that $\lambda_{\max}(D^{-1})$ is bounded in probability, where D represents a scaled version of the denominator matrix. This requires some work as the scaled denominator matrix is not required to converge in distribution to a stable limit. Instead I directly show that $\lambda_{\max}(D^{-1})$ is bounded in probability by showing that $\Pr(\lambda_{\min}(D) \le \delta_n) \to 0$ for any sequence $\delta_n \to 0$.

To do this I first demonstrate that it is sufficient to show that $\Pr(a'Da \leq \delta_n) \to 0$ for any $\delta_n \to 0$ and fixed $a \in \mathcal{S}^{d_x-1} = \{v \in \mathbb{R}^{d_x-1} : \|v\| = 1\}$. I then establish the claim for an arbitrary choice of a. As in Lemma 3.1 this is done by comparing the scaled quadratic form of the denominator matrix to a gaussian analog and then establishing the corresponding result for the gaussian analog. This corresponding result is again also useful for establishing the validity of the interpolation approach with a dynamic choice of approximation parameter.

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

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

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

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

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

Theorem 5.2 (Local Power). *Suppose that Assumptions 5.1–5.4 hold. Then*

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

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

Just as in Theorem 3.1, the result in Theorem 5.2 does not require any stable distribution for the limiting statistic $JK_G(\beta_0)$ under H_1.

5.3 IMPROVING POWER AGAINST CERTAIN ALTERNATIVES

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

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

The conditioning statistic C attempts to detect whether, for *some* $\ell \in [d_x]$, $\mathbb{E}[\widehat{\Pi}_i] = 0$ for all $i \in [n]$. Based on the value of the conditioning statistic the researcher can decide whether to run a test based on $JK(\beta_0)$ or a test based on the sup-score statistic $S(\beta_0)$.

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

Theorem 5.3 establishes the validity of the test based on the the thresholding statistic for any choice of cutoff τ. As with Theorem 4.1, this is done by first establishing that quantiles of $(JK(\beta_0), C)$ and $(S(\beta_0), C)$ can jointly be approximated by gaussian analogs and then using marginal independence of the testing and conditioning statistics under the null; $(JK(\beta_0) \perp C)$ and $(S\beta_0) \perp C)$ under H_0.

Theorem 5.3. *Suppose that Assumptions 4.1(ii,iii), 5.1, 5.2, and 5.4 hold. Then the test based on $T(\beta_0); \tau)$ has asymptotic size α for any choice of cutoff τ.*

Finally, I show that the quantiles of the generalized conditioning statistic C can be approximated by a multiplier bootstrap procedure under the assumption that $\mathbb{E}[\widehat{\Pi}_{\ell i}] = 0$ for all $\ell = 1, \ldots, d_x$ and $i = 1, \ldots, n$. Let $e_1, \ldots, e_n$ be generated iid standard normal independent of the data and for any $\theta \in (0, 1)$, define the conditional bootstrap quantile:

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

Proposition 5.1. *Suppose that Assumptions 4.1(ii,iii), 5.1, 5.2, and 5.4 hold. Then if $\mathbb{E}[\widehat{\Pi}_{\ell i}] = 0$ for all $\ell \in [d_x]$ and $i \in [n]$ there exist constant $\delta_n \searrow 0$ and $\beta_n \searrow 0$ such that*

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

with probability $1 - \delta_n$.

In practice, I recommend taking the cutoff τ to be the 50th quantile of the distribution of C under the assumption that $\mathbb{E}[\widehat{\Pi}^I_{\ell i}] = 0$ for all $\ell \in [d_x]$ and $i \in [n]$.

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. So long as an auxiliary parameter can be consistently estimated, I show that the test is robust to both 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 to run a test via the Jackknife K-statistic or one via the sup-score statistic based on the value of a conditioning statistic. While this combination does not fully address the power decline, I show that it works well in simulation study and leave further refinements to future work.

References

[1] T. W. Anderson and Herman Rubin, *Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations*, The Annals of Mathematical Statistics **20** (1949), no. 1, 46 –63.

[2] Donald W. K. Andrews, Marcelo J. Moreira, and James H. Stock, *Optimal two-sided invariant similar tests for instrumental variables regression*, Econometrica **74** (2006), no. 3, 715–752.

[3] Donald W.K. Andrews, Marcelo Moreira, and James H Stock, *Optimal invariant similar tests for instrumental variables regression* **299** (2004August).

[4] Donald W.K. Andrews and James H. Stock, *Testing with many weak instruments*, Journal of Econometrics **138** (2007), no. 1, 24–46. 50th Anniversary Econometric Institute.

[5] Isaiah Andrews, *Conditional linear combination tests for weakly identified models*, Econometrica **84** (2016), no. 6, 2155–2182, available at https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA12407.

[6] J. D. Angrist, G. W. Imbens, and A. B. Krueger, *Jackknife instrumental variables estimation*, Journal of Applied Econometrics **14** (1999), no. 1, 57–67.

[7] Paul A. Bekker, *Alternative approximations to the distributions of instrumental variable estimators*, Econometrica **62** (1994), no. 3, 657–681.

[8] A. Belloni, D. Chen, V. Chernozhukov, and C. Hansen, *Sparse models and methods for optimal instruments with an application to eminent domain*, Econometrica **80** (2012), no. 6, 2369–2429, available at https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA9626.

[9] Alexandre Belloni, Victor Chernozhukov, Denis Chetverikov, Christian Hansen, and Kengo Kato, *High-dimensional econometrics and regularized gmm*, 2018.

[10] John Bound, David A. Jaeger, and Regina M. Baker, *Problems with instrumental variables estimation when the correlation between the instruments and the endogeneous explanatory variable is weak*, Journal of the American Statistical Association **90** (1995), no. 430, 443–450.

[11] Gary Chamberlain, *Asymptotic efficiency in estimation with conditional moment restrictions*, Journal of Econometrics **34** (1987), no. 3, 305–334.

[12] John C Chao, Norman R Swanson, Jerry A Hausman, Whitney K Newey, and Tiemen Woutersen, *Asymptotic distribution of jive in a heteroskedastic iv regression with many instruments*, Econometric Theory **28** (2012), no. 1, 42–86.

[13] Victor Chernozhukov, Denis Chetverikov, and Kengo Kato, *Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors*, The Annals of Statistics **41** (2013), no. 6, 2786 –2819.

[14] ______, *Central limit theorems and bootstrap in high dimensions*, The Annals of Probability **45** (2017), no. 4, 2309–2352.

[15] Victor Chernozhukov, Whitney K. Newey, and Rahul Singh, *Automatic debiased machine learning of causal and structural effects*, Econometrica **90** (2022), no. 3, 967–1027, available at https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA18515.

[16] Denis Chetverikov and Jesper Riis-Vestergaard Sørensen, *Analytic and bootstrap-after-cross-validation methods for selecting penalty parameters of high-dimensional m-estimators*, ArXiv **NA** (2021), 1–50, available at 2104.04716.

[17] Federico Crudu, Giovanni Mellace, and Zsolt Sándor, *Inference in instrumental variable models with heteroskedasticity and many instruments*, Econometric Theory **37** (2021), no. 2, 281–310.

[18] Ellora Derenoncourt, *Can you move to opportunity? evidence from the great migration*, American Economic Review **112** (2022February), no. 2, 369–408.

[19] Friedrich Götze, Alexey Naumov, Vladimir Spokoiny, and Vladimir Ulyanov, *Large ball probabilities, Gaussian comparison and anti-concentration*, Bernoulli **25** (2019), no. 4A, 2538 –2563.

[20] Friedrich Gotze, Holger Sambale, and Arthur Sinulis, *Concentration inequalities for polynomials in alpha-sub-exponential random variables*, Electronic Journal of Probability **26** (2021), no. none, 1 –22.

[21] Chirok Han and Peter C. B. Phillips, *Gmm with many moment conditions*, Econometrica **74** (2006), no. 1, 147–192, available at https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1468-0262.2006.00652.x.

[22] Frank E. Harrell, *Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis*, Spring Series in Statistics, Springer Cham, 2015.

[23] Frank Kleibergen, *Pivotal statistics for testing structural parameters in instrumental variables regression*, Econometrica **70** (200202), 1781–1803.

[24] ______, *Testing parameters in gmm without assuming that they are identified*, Econometrica **73** (2005), no. 4, 1103–1123, available at https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1468-0262.2005.00610.x.

[25] David S. Lee, Justin McCrary, Marcelo J. Moreira, and Jack Porter, *Valid t-ratio inference for iv*, American Economic Review **112** (2022October), no. 10, 3260–90.

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

[27] Yukitoshi Matsushita and Taisuke Otsu, *A jackknife lagrange multiplier test with many weak instruments*, Econometric Theory (2022), 1–24.

[28] Anna Mikusheva, *Many weak instruments in time series econometrics*, Working Paper (2023).

[29] Anna Mikusheva and Liyang Sun, *Inference with many weak instruments*, The Review of Economic Studies **89** (202112), no. 5, 2663–2686, available at https://academic.oup.com/restud/article-pdf/89/5/2663/45764026/rdab097_supplementary_data.pdf.

[30] Marcelo Moreira, *Tests with correct size when instruments can be arbitrarily weak*, Journal of Econometrics **152** (200910), 131–140.

[31] Marcelo J Moreira, *Tests with correct size when instruments can be arbitrarily weak*, Citeseer, 2001.

[32] Marcelo J. Moreira, *A conditional likelihood ratio test for structural models*, Econometrica **71** (2003), no. 4, 1027–1048, available at https://onlinelibrary.wiley.com/doi/pdf/10.1111/1468-0262.00438.

[33] Fedor Nazarov, *On the maximal perimeter of a convex set in r^n with respect to a gaussian measure* (Vitali D. Milman and Gideon Schechtman, eds.), Springer Berlin Heidelberg, Berlin, Heidelberg, 2003.

[34] Charles R. Nelson and Richard Startz, *Some further results on the exact small sample properties of the instrumental variable estimator*, Econometrica **58** (1990), no. 4, 967–976.

[35] Whitney K. Newey and Frank Windmeijer, *Generalized method of moments with many weak moment conditions*, Econometrica **77** (2009), no. 3, 687–719, available at https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA6224.

[36] K. B. Petersen and M. S. Pedersen, *The matrix cookbook*, Technical University of Denmark, 2012. Version 20121115.

[37] Demian Pouzo, *Bootstrap consistency for quadratic forms of sample averages with increasing dimension*, Electronic Journal of Statistics **9** (2015), no. 2, 3046 –3097.

[38] Holger Sambale, *Some notes on concentration for α-subexponential random variables*, 2022.

[39] Douglas Staiger and James H. Stock, *Instrumental variables regression with weak instruments*, Econometrica **65** (1997), no. 3, 557–586.

[40] James Stock and Motohiro Yogo, *Testing for weak instruments in linear iv regression* (Donald W.K. Andrews, ed.), Cambridge University Press, Cambridge University Press, New York, 2005.

[41] Zhiqiang Tan, *Regularized calibrated estimation of propensity scores with model misspecification and high-dimensional data*, ArXiv **NA** (2017), 1–60, available at 1710.08074.

[42] Sara van der Greer, *Estimation and testing under sparsity*, Lecture Notes in Mathematics, Springer, New York, NY, 2016.

[43] Wessel N. van Wieringen, *Lecture notes on ridge regression*, 2023.

A Proofs of Results in Section 3

A.1 Proof of Theorem 3.1

Before proceeding, we will introduce some notations. Let $\tilde{H} = s_n H$ and $\tilde{h}_{ij} = s_n h_{ij}$, where s_n is as in Assumption 3.2. Define

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

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

where $(\tilde{\epsilon}_i(\beta_0), \tilde{r}_i)$ are jointly gaussian with the same 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 $K(\beta_0) = \frac{N^2}{D} \mathbf{1}_{\{D>0\}}$ and $\tilde{K}(\beta_0) = \frac{\tilde{N}^2}{\tilde{D}}$. 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 that $\Pr(\tilde{D} > 0) = 1$, for any $a \geq 0$ we can rewrite the events

$$\{JK(\beta_0) \leq a\} = \{N^2 - aD \leq 0\} \text{ and } \{J\tilde{K}(\beta_0) \leq a\} \overset{\text{a.s}}{=} \{\tilde{N}^2 - a\tilde{D} \leq 0\} \tag{A.1}$$

with this in mind define

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

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

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

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

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

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

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

where $\dot{\epsilon}_\ell(\beta_0)$ 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 r_ℓ 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 since we will only be considering one deviation at a time we will supress the dependence of these variables on i in order to simplify notation.

Next, define the one step deviations

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

We can then write the difference $\mathbb{E}[\varphi(K^a) - \varphi(\tilde{K}^a)]$ as a telescoping sum

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

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

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

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 any i the mean and covariance matrix of $(\epsilon_i(\beta_0), r_i)$ is the same as the mean and covariance matrix of $(\tilde{\epsilon}_i(\beta_0), \tilde{r}_i)$, (b) $\mathbb{E}[\epsilon_i^2(\beta_0)] = \kappa_i^2(\beta_0)$, and (c) random variables are independent across indices, we have that

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

Using this we can simplify the prior display

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

$$+ \underbrace{4n^{-3/2}\mathbb{E}[\varphi''(K_{-i})N_{-i}(\Delta_{1i}\Delta_{2i}^{a} - \tilde{\Delta}_{1i}\tilde{\Delta}_{2i}^{a})]}_{\mathbf{E}_i} + \underbrace{2n^{-2}\mathbb{E}[\varphi''(K_{-i})(\Delta_{1i}^2\Delta_{2i} - \tilde{\Delta}_{1i}^2\tilde{\Delta}_{2i})}_{\mathbf{F}_i}$$
$$+ R_i + \tilde{R}_i$$

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

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

Applications of Lemmas E.1 and E.2, Cauchy-Schwarz, and generalized Hölder inequality,[1] will allow us to bound for a fixed constant M that only depends on c,

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

and

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

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

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

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

Proof of Lemma A.2. By Assumption 3.1, we know that $\kappa_i^2(\beta_0) \in [c^{-1}, c]$ for all $i = 1, \dots, n$ so that $\tilde{D} \geq \frac{c^{-1}}{n}\sum_{i=1}^{n}(\sum_{j\neq i}\tilde{h}_{ij}r_j)^2$. Then

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

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

$$\Pr(\tilde{r}'\bar{H}\tilde{r} \leq \delta_n) = \Pr(\tilde{r}'\bar{H}\tilde{r} - \mathbb{E}[\tilde{r}'\bar{H}\tilde{r}] \leq \delta_n - \mathbb{E}[\tilde{r}'\bar{H}\tilde{r}])$$

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

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

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

Second Bound. For the second bound, we will directly use bounds on the density of gaussian quadratic forms from [19]. The vector $r'\bar{H}^{1/2}$ is gaussian with covariance matrix $\Sigma_r = \bar{H}^{1/2}\boldsymbol{R}\bar{H}^{1/2}$ where $\boldsymbol{R} = \operatorname{diag}(\operatorname{Var}(r_1), \ldots, \operatorname{Var}(r_n))$. Let $\Lambda_1 = \sum_{k=1}^n \lambda_k^2(\Sigma_r)$ and $\Lambda_2 = \sum_{k=2}^n \lambda_k^2(\Sigma_r)$. By Assumption 3.2 and Lemma F.5, Λ_2/Λ_1 is bounded away from zero. Using Theorem G.4 we can then bound for some constant $C > 0$

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

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

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

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

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

where the first equality uses the symmetric square root of $\bar{H}$, the first inequality comes from Courant-Fischer min-max 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 p.s.d, $\lambda_1(M) \leq \sqrt{\operatorname{trace}(M^2)}$ 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

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

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

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

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

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

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

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

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

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

By Lemma A.1 we can write for some constant M that only depends on Δ:

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

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

$$\begin{aligned}
&\leq \Pr(K^a \leq 0) + \underbrace{\Pr(a \leq \tilde{N}^2/\tilde{D} \leq a + \gamma_n^{1/2})}_{\mathbf{A}} \\
&\quad + \underbrace{\Pr(\tilde{D} \leq \gamma_n^{1/2})}_{\mathbf{B}} + \frac{M}{\sqrt{n}}(\gamma_n^{-2} + \gamma_n^{-3})
\end{aligned}$$

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

$$\Pr(\tilde{K}(\beta_0) \leq a) + o(1) \leq \Pr(K(\beta_0) \leq a) \leq \Pr(\tilde{K}(\beta_0) \leq a) + o(1)$$

where the $o(1)$ is uniform over all $a \leq \Delta$. Noting that the numerator $\tilde{K}(\beta_0)$ is $O_p(1)$ under Assumption 3.3 while the inverse of the denominator of $\tilde{K}(\beta_0)$ is $O_p(1)$ by Lemma A.2, we can apply Lemma A.6 to obtain the approximation uniformly over all $a \in \mathbb{R}$. □

Lemma A.4. *Let X_n and Y_n be two sequences of random variables and let $W_n = X_n/Y_n$. Then for any $c \in \mathbb{R}$ and any $\delta > 0$:*

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

and

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

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

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

The second statement of the lemma follows symmetrically. □

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

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

Then $X_n = O_p(1)$.

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

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

□

Lemma A.6. *Suppose that X_n and Y_n are sequences of (real-valued) random variables such that $Y_n = O_p(1)$ and for any $\Delta \in \mathbb{R}$*

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

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

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

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

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

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

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

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

$$\leq |\Pr(X_n \leq M_{\epsilon/3}) - \Pr(Y_n \leq M_{\epsilon/3})| + \epsilon/3 + \epsilon/3 \qquad \text{(A.11)}$$

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

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

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

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

A.2 Proof of Proposition 3.1

Proof of Proposition 3.1. As at the top of Appendix A.1 define

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

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

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

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

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

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

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

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

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

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

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

$(\mathbb{E}[|X_n|])^2$, so $\mathbb{E}[X_n^2] \to \infty$ and $\text{Var}(|X_n|) = O(1)$ implies that $\mathbb{E}[|X_n|] \to \infty$. Then,

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

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

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

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

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

□

A.3 Proof of Lemma 3.1

Proof of Lemma 3.1. For N and D defined at the top of Appendix A.1 define $\widehat{N} = N + \Delta_N$ and $\widehat{D} = D + \Delta_D$. We can then write $JK(\beta_0) = \widehat{N}^2/\widehat{D}$ and rewrite

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Using Lemma E.1 we can bound

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

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

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

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

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

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

Proof of Lemma A.11. Let $\tilde{\varphi}(\cdot) : \mathbb{R} \to \mathbb{R}$ be three times continuously differentiable with bounded derivatives up to the third order such that $\tilde{\varphi}(x)$ is 1 if $x \leq 0$, $\tilde{\varphi}(x)$ is decreasing if $x \in (0, 1)$, and $\tilde{\varphi}(x)$ is zero if $x \geq 1$. Consider a second sequence $\gamma_n \searrow 0$ slowly enough such that $(\gamma_n^{-2} + \gamma_n^{-3})/\sqrt{n} \to 0$. Take $\varphi_n(x) = \tilde{\varphi}(\frac{x-\delta_n}{\gamma_n})$. By Lemma A.10 and since $\tilde{\varphi}(\cdot)$ has bounded derivatives up to the third order, there is a fixed constant $M_1 > 0$ that depends only on c such that

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

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

A.4 Proof of Lemma 3.2

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

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

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

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

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

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

$$\max_{1 \leq j \leq d_b} |\epsilon(\beta_0)'\boldsymbol{H}B_j\epsilon(\beta_0) - \mathbb{E}[\epsilon(\beta_0)'\boldsymbol{H}B_j\epsilon(\beta_0)]| = O_p(\log^{2/a}(d_b)) \tag{A.12}$$

Since $\max_{1\le j\le d_b} |\mathbb{E}[\epsilon(\beta_0)'\boldsymbol{H}B_j\epsilon(\beta_0)]| \le c$ under Assumption 3.4(v), (A.12) gives that

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

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

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

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

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

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

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

To do so write

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

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

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

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

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

A.5 Proof of Theorem 3.2

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

Lemma A.12. *Let X_n, Y_n, and Z_n be sequences of random variables such that $|X_n - Y_n| \to_p 0$, the distribution of Z_n is absolutely continuous with respect to Lebesgue measure and the density functions of Z_n*

are uniformly bounded and $\sup_{a\in\mathbb{R}} |\Pr(Y_n \leq a) - \Pr(Z_n \leq a)| \to 0$. *Then* $\sup_{a\in\mathbb{R}} |\Pr(X_n \leq a) - \Pr(Z_n \leq a)| \to 0$.

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

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

so that

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

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

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

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

We can apply a symmetric argument to show that $\sup_{a\in\mathbb{R}} \Pr(Z_n \leq a) - \Pr(X_n \leq a) \leq o(1)$ which completes the claim of the lemma. □

B Proofs of Results in Section 4

The first statement of Theorem 4.1 relies on a joint interpolation argument of the $JK(\beta_0)$ test statistic and the infeasible conditioning statistic C_I, which could be constructed if $\rho(z_i)$ was known to the researcher.

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

This joint interpolation argument is rather involved however, and deferred to Appendix D. The interpolation argument for the conditioning statistic very closely follows the results in [13]. The results of Section 4 rely on showing that the difference between C and C_I can be treated as negligible. This in turn reduces to verifying Assumption G.2, which is done in Lemma B.1, below.

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

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

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

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

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

Notice that we can bound

$$\begin{aligned}\max_{1\le i\le n}\frac{1}{n}\sum_{j\neq i}^{n}(\hat{r}_i - r_i)^2 &= \max_{1\le i\le n}\left|(\widehat{\gamma}-\gamma)'n^{-1}\sum_{j\neq i}\epsilon_j^2(\beta_0)b(z_i)b(z_j)'(\widehat{\gamma}-\gamma)\right| \\ &\quad + \max_{1\le i\le n}|n^{-1}\sum_{j\neq i}\dot{h}_{ij}^2\xi_j^2| \\ &\le \max_{\substack{1\le i\le n\\ 1\le j,k\le d_b}}\underbrace{\left|n^{-1}\sum_{j\neq i}\epsilon_j^2(\beta_0)b_j(z_j)b_k(z_j)\right|}_{\mathbf{A}_{ijk}}\|\widehat{\gamma}-\gamma\|_1^2 \\ &\quad + n^{-1/2}\max_{1\le i\le n}(n^{-1}\sum_{j\neq i}\dot{h}_{ij}^4)^{1/2}(\sum_{j\neq i}\xi_j^4)^{1/2}\end{aligned}$$

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

B.1 Proof of Theorem 4.1

The result with $JK(\beta_0)$ and C replaced with their infeasible analogs $JK_I(\beta_0)$ and C_I follows from the argument in Appendix D. After verifying that $|JK(\beta_0) - JK(\beta_0)| \to_p 0$ via Lemma 3.2 and that Assumption G.2 is satisfied via Lemma B.1 follow the same steps as in the proof of [9, Theorem 2.1] to see that approximation result holds for the feasible $JK(\beta_0)$ and C.

B.2 Proof of Proposition 4.1

Assumption G.1(i,ii) is satisfied under Assumption 3.1 by the definition of $\dot{h}_{ij} = h_{ij}/(n^{-1}\sum_{j\neq i}h_{ij}^2)^{1/2}$ and the variance of each r_j being bounded from zero and fourth moments being bounded from above. Assumption G.1(iii) is satisfied with $B_n = \log^{1/v}(n)$ by Assumption 4.1(i,iii) and Lemma F.2. Finally Assumption G.2 is satisfied by applying Lemma B.1. Apply Theorem G.6 to conclude.

C Proofs of Results in Section 5

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

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

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

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

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

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

C.1 Proof of Theorem 5.1

Theorem 5.1 follows immediately from the joint interpolation argument established in Appendix D.

D Joint Gaussian Approximation of $JK(\beta_0)$ and C

The main results of Sections 4 and 5 rely on a joint interpolation of the conditioning and testing statistics as well as a joint interpolation of the conditioning and testing statistics. The joint interpolation of $JK(\beta_0)$ and the conditioning statistic C is given in Appendix D.2 after introducing some notation in Appendix D.1. The joint gaussian approximation of $S(\beta_0)$ and C follows immediately from results in [9, 14]. The result is presented below for the general form of the $JK(\beta_0)$ under H_0 however the proof strategy is very similar when using the decomposed form of $JK(\beta_0)$ when $d_x = 1$ and under local alternatives as defined in Assumption 3.3. This proof is available on request.

D.1 Notation

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

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

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

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 3.1 while the definitions of $\dot{\epsilon}_j(\beta_0), \dot{r}_{\ell j}$, and $\ddot{\epsilon}_j(\beta_0)$ depend on i this dependence is suppressed to conslidate notation and since we only consider one step deviations at a time.

Also define the one step deviations

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

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

$$\Delta_{Di} = \underbrace{\left[(\Delta^a_{Di})_{\ell m}\right]_{\substack{1\le\ell\le d\\1\le m\le d}}}_{\Delta^a_{Di}} + \underbrace{\left[(\Delta^b_{Di})_{\ell m}\right]_{\substack{1\le\ell\le d\\1\le m\le d}}}_{\Delta^b_{Di}}$$

$$\tilde\Delta_{Di} = \underbrace{\left[(\tilde\Delta^a_{Di})_{\ell m}\right]_{\substack{1\le\ell\le d\\1\le m\le d}}}_{\tilde\Delta^a_{Di}} + \underbrace{\left[(\tilde\Delta^b_{Di})_{\ell m}\right]_{\substack{1\le\ell\le d\\1\le m\le d}}}_{\tilde\Delta^b_{Di}}$$

where

$$\begin{aligned}
(\Delta^a_{Di})_{\ell m} &= \epsilon_i^2(\beta_0)\Big(\sum_{j\neq i}\tilde h_{\ell,ij} r_{\ell j}\Big)\Big(\sum_{j\neq i}\tilde h_{\ell,ij}\dot r_{\ell j}\Big)\Big(\sum_{j\neq i} h_{m,ij} r_{m,ij}\Big)^2 + r_{\ell i} r_{ki}\sum_{j\neq i}\tilde h_{\ell,ij}\tilde h_{m,ij}\ddot\epsilon_j^2(\beta_0)\\
(\tilde\Delta^a_{Di})_{\ell m} &= \tilde\epsilon_i^2(\beta_0)\Big(\sum_{j\neq i}\tilde h_{\ell,ij} r_{\ell j}\Big)\Big(\sum_{j\neq i}\tilde h_{\ell,ij}\dot r_{\ell j}\Big)\Big(\sum_{j\neq i} h_{m,ij} r_{m,ij}\Big)^2 + \tilde r_{\ell i}\tilde r_{ki}\sum_{j\neq i}\tilde h_{\ell,ij}\tilde h_{m,ij}\ddot\epsilon_j^2(\beta_0)\\
(\Delta^b_{Di})_{\ell m} &= r_{\ell i}\sum_{j\neq i}\ddot\epsilon_j^2(\beta_0)\sum_{k\neq i,j}\tilde h_{\ell,ji}\tilde h_{m,jk}\dot r_{mk} + r_{ki}\sum_{j\neq i}\ddot\epsilon_j^2(\beta_0)\sum_{k\neq i,j}\tilde h_{\ell,ji}\tilde h_{m,jk}\dot r_{\ell k}\\
(\tilde\Delta^b_{Di})_{\ell m} &= \tilde r_{\ell i}\sum_{j\neq i}\ddot\epsilon_j^2(\beta_0)\sum_{k\neq i,j}\tilde h_{\ell,ji}\tilde h_{m,jk}\dot r_{mk} + \tilde r_{ki}\sum_{j\neq i}\ddot\epsilon_j^2(\beta_0)\sum_{k\neq i,j}\tilde h_{\ell,ji}\tilde h_{m,jk}\dot r_{\ell k}
\end{aligned}$$

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

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

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

$$\begin{aligned}
U &= U_{-1} + \Delta_{U1}/\sqrt{n} & \tilde U &= U_{-n} + \tilde\Delta_{Un}/\sqrt{n}\\
D &= D_{-1} + \Delta_{D1}/n & \tilde D &= D_{-n} + \Delta_{Dn}/n
\end{aligned}$$

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 $\omega_{\ell i} = n^{-1}\sum_{j\neq i}|h_{\ell,ij}|$. Also define the one-step deviations:

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

And the leave-one-out vector

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

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

Function Definitions. As in [13] 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

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

Appendix E.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 $x \leq 0$, $\varphi'(x) < 0$ for $x \in (0, 1)$, and $\varphi(x) = 0$ for $x \geq 1$. For any $\gamma > 0$ and $a = (a_1, a_2)' \in \mathbb{R}^2$ define the function $\tilde{\varphi}(\cdot, \cdot, \cdot) : \mathbb{R}^{d_x} \times \mathrm{vec}(\mathbb{R}^{d_x \times d_x}) \times \mathbb{R}^{2nd_x} \to \mathbb{R}$ via

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

where

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

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

D.2 Main Argument

Lemma D.1 (Joint Lindeberg Interpolation). *Suppose that Assumption 5.1 holds. Then there is a fixed constant M*

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

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

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

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

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

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

First Order Terms. First order terms can be expressed

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

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

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

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

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 E.12 and E.15. Under Assumption 5.1 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 E.1. Putting these together with generalized Holder's inequality will yield a finite constants M_1 and M_2 such that $|\mathbf{A}_{lmn}| \leq M_1 \log^{M_2}(n)(\gamma^{-1} + \gamma^{-2})$, $\mathbf{B}_{\ell mno} \leq M_1 \log^{M_2}(n)(\gamma^{-1} + \gamma^{-2})$, and $|\mathbf{C}_{\ell mno}| \leq$

$M_1 \log^{M_2}(n) n^{1/2}(\gamma^{-1} + \gamma^{-2})$. To bound $\mathbf{D}_{\ell mn}$ terms notice that

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

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

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

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

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

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

Putting these all together and summing over the remaining indices gives

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

Remainder Terms. The first remainder term can be expressed

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

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

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

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

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

Lemma D.2 (Denominator Anticoncentration). *Suppose that Assumption 5.1 holds. Then for any sequence $\delta_n \to 0$ we have that* $\Pr(\lambda_{\min}(\tilde{D}) \le \tilde{\delta}_n) \to 0$.

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

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

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

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

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

Lemma D.3 (Gaussian Approximation).

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

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

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

Let $\gamma \longrightarrow 0$ at a rate such that $\frac{\log^{M_2}(n)}{\sqrt{n}}\gamma^{-3} \longrightarrow 0$.

□

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

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

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

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

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

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

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

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

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

Define the events

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

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

E Relevant Moment Bounds

E.1 Moment Bounds for Section 3

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

Lemma E.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 Assumptions 3.1 and 3.2 there is a constant $M > 0$ such that for any $k = 1, \dots, 6$:*

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

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

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

Proof. First, since

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

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

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

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

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

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

Lemma E.2. *Let N and N_{-i} be defined as in Appendix A.1. Under Assumptions 3.1–3.3 there is a fixed constant M such that for all $i = 1, \dots, n$ and any $k = 1, \dots, 6$,*

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

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 $N = N_1 + N_2 + N_3$:

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

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

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

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

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

An absolute bound on the higher moments of N_2 then follows from an application of Lemma E.4 with $X_i = r_i \sum_{j\neq i} h_{ji}\gamma_j/\sqrt{n}$. 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 E.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 E.7.

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

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

Proof. Recall that

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

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

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

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

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

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

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

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

□

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

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

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

Proof. First write

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

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

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

where the implied constant in the second line is 32 by an application of Lemma F.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 E.5. *Let $h = (h_1, \dots, h_n) \in \mathbb{R}^n$ be such that $\sum_{i=1}^n h_i^2 \leq b$. Suppose that $X_1, \dots, X_n$ are such that $\mathbb{E}[|X_i|^k] \leq M$ for all $k = 1, 2, 3$. Then*

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

Proof. We can expand out

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

□

Lemma E.6. *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*

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

Proof. We can expand out

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

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

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

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

Lemma E.7. *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*

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

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

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

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

E.2 Useful Properties of Smooth Max

Lemma E.8 ([13, Lemma A.2])**.** *For every $1 \leq j, k, l \leq p$,*

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

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

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

Moreover,

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

Lemma E.9 ([13, 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 E.10 ([13, Lemma A.4])**.** *Let $\varphi(\cdot) : \mathbb{R} \to \mathbb{R}$ be such that $\varphi \in C_b^3(\mathbb{R})$ and define $m : \mathbb{R}^p \to \mathbb{R}$, $z \mapsto \varphi(F_\beta(z))$. The derivatives (up to the third order) of m are given*

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

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

Lemma E.11 ([13, 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* $1 \le j, k, l \le p$,

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

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

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

Moreover,

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

E.3 Moment Bounds for Sections 4 and 5

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

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

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

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

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 G.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 F.2 provides the moment bound for arbitrary moments. To bound moments of N_ℓ^2 we use the fact that $\max_i \left| \sum_{j \ne i} \tilde{h}_{ij} \mathbb{E}[r_{\ell j}] \right|$ is bounded by assumption and apply Burkholder-Davis-Gundy (Theorem G.3) after adding and subtracting $\mathbb{E}[\epsilon_i(\beta_0)]$.

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

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

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

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

First Order:

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

Second Order:

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

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

Third Order:

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

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

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

Proof. The derivative of an element of the the inverse of a matrix $\boldsymbol{X}$ can be expressed [36]

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

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

$$\begin{aligned}\frac{\partial\,(\det(D))^p}{\partial D_{lm}} &= p(\det(D))^{p-1}(D^{-1})_{lm}\\ \frac{\partial^2\,(\det(D))^p}{\partial D_{lm}\partial D_{pq}} &= \frac{p!}{(p-2)!}(\det(D))^{p-2}(D^{-1})_{pq}(D^{-1})_{lm}\end{aligned}$$

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

Proof. We can express the derivative of the detrminant [36],

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

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

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

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

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

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

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 E.13 and E.14 as well as the inequality (E.5) below. For the moment bounds, we may extensive use of following bounds on elements of D^{-1} for a positive-definite D^{-1}:

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

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

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

□

F Technical Lemmas

F.1 Probability Lemmas

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

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

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

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

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

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

Lemma F.2. *Suppose that* $X_1, \dots, X_n$ *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. □

F.2 Matrix Lemmas

Lemma F.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 F.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 F.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

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

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

Lemma F.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*

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

F.3 Miscellaneous Lemmas

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

Proof. Replace $a_i \leq Ub_i$ for the upper bound and $a_i \geq Lb_i$ for the lower bound. □

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

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

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

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

G Assorted Results from Literature

G.1 Concentration Inequalities and Tail Bounds

Theorem G.1 ([20, Theorem 1.2]). *Let $X_1, \dots, X_n$ be independent random variables satisfying $\|X_i\|_{\Psi_a} \leq M$ for some $a \in (0,1] \cup \{2\}$ and let $f : \mathbb{R}^n \to \mathbb{R}$ be a polynomial of total degree $D \in \mathbb{N}$. Then for all $t > 0$;*

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

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

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

or equivalently

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

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

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

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

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

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

G.2 Anticoncentration Bounds

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

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

Theorem G.4 ([19, Theorem 2.6]). *Let ξ be a gaussian element with zero mean and covariance Σ_ξ. Then it holds for any $\boldsymbol{a} \in \mathbb{R}^n$ that*

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

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

We use the following anticoncentration lemma from [33] noted in [14].

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

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

where C is a constant only depending on b.

G.3 Gaussian Comparasions and Approximations

We also use the following gaussian approximation results from [9, 14]. 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 $\widehat{X}_1, \dots, \widehat{X}_n \in \mathbb{R}^p$.

Define the sums

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

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

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

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

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

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

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

as well as the following bounds on the estiamtion error

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

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

Theorem G.5 ([9, Theorem 2.1])**.** *Suppose that Assumptions G.1 and G.2 hold. Then there is a constant C which depends only on b such that*

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

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

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

Theorem G.6 ([9, Theorem 2.2])**.** *Suppose that Assumptions G.1 and G.2 hold. Then there is a constant C which depends only on b such that*

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

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