

# A Robust Test for Weak Instruments with Multiple Endogenous Regressors\*

DANIEL J. LEWIS

University College London

KAREL MERTENS<sup>†</sup>

Federal Reserve Bank of Dallas

CEPR

December 9, 2022

## Abstract

We generalize the popular bias-based test of Stock and Yogo (2005) for instrument strength in two-stage least-squares models with multiple endogenous regressors to be robust to heteroskedasticity and autocorrelation. Equivalently, we extend the robust test of Montiel Olea and Pflueger (2013) for a single endogenous regressor to the general case with multiple endogenous regressors. We describe a simple procedure for applied researchers to conduct our generalized first-stage test of instrument strength, and provide fast Matlab code for its implementation. In simulations, our test controls size and is powerful. We demonstrate our testing procedures by considering the estimation of the state-dependent effects of fiscal policy as in Ramey and Zubairy (2018).

*JEL classification:* C26, C36.

*Keywords:* Instrumental Variables, Weak Instruments Test, Multiple Endogenous Regressors, Heteroskedasticity, Serial Correlation.

---

\*The views in this paper are those of the authors and do not necessarily reflect the views of the Federal Reserve Banks of Dallas or the Federal Reserve System. We are grateful to José Luis Montiel Olea and Carolin Pflueger for helpful discussions and for sharing their computer code, to Nikolay Gospodinov for a great discussion, and to Isaiah Andrews, Lutz Kilian, Sarah Zubairy, and seminar participants at the Federal Reserve System Econometrics Meeting for helpful comments. All errors are ours.

<sup>†</sup>*Contact:* mertens.karel@gmail.com, tel: +(214) 922-6000

Empirical researchers using instrumental variables (IV) estimation frequently report first-stage  $F$ -statistics – or in the multiple endogenous regressors case, Cragg and Donald (1993) statistics – to assess instrument relevance. These statistics are typically compared to critical values based on the bias of two-stage least-squares (2SLS) relative to the bias of OLS, which are tabulated by Stock and Yogo (2005). However, this testing procedure requires the assumption of conditionally homoskedastic and serially uncorrelated errors, and is generally invalid if that assumption does not hold. As discussed in a recent survey by Andrews et al. (2019), in practice this often means that researchers assume conditionally homoskedastic and serially uncorrelated errors for the purpose of a first-stage test, but make different assumptions when conducting inference for the second-stage parameters of interest.

In an important paper, Montiel Olea and Pflueger (2013) introduce a new statistic – the ‘effective  $F$ -statistic’ – that accounts for heteroskedasticity and autocorrelation in the model errors. Their test is based on a Nagar approximation of the 2SLS bias relative to a worst-case benchmark, and the limiting distribution and associated critical values depend on the application-specific structure of the robust asymptotic covariance matrix of the reduced-form and first-stage parameters. A limitation, however, is that this test only applies to models with a single endogenous regressor. Andrews et al. (2019) point to the lack of a heteroskedasticity and autocorrelation robust (HAR) test for models with multiple endogenous regressors as an important remaining gap in the IV pre-testing literature. The contribution of this paper is to fill that gap by extending the Montiel Olea and Pflueger (2013) test to allow for an arbitrary number of endogenous regressors, and in so doing generalize the Stock and Yogo (2005) bias-based test to be heteroskedasticity and autocorrelation robust. We restrict our attention to the 2SLS estimator, since it is by far the most commonly used IV estimator in empirical practice.

Like Stock and Yogo (2005), we use a weighted quadratic loss in the asymptotic bias of the 2SLS estimates under a local-to-zero assumption as the criterion for the test. The bias criterion is relative to the same worst-case benchmark as in Montiel Olea and Pflueger (2013), and also nests the criterion of Stock and Yogo (2005) in conditionally homoskedastic and serially uncorrelated models. The null hypothesis is that the minimum eigenvalue of the concentration matrix is in the set of values for which the bias is greater than some fraction  $\tau$  of the worst-case benchmark. The latter is the 2SLS bias that arises for instruments that have no explanatory power for the endogenous regressors, and structural errors that are perfect linear combinations of the second-stage regressors in small samples. As in Montiel Olea and Pflueger

(2013), we characterize the boundary of the weak instrument set using a Nagar approximation to the bias. We show analytically that this ‘Nagar bias’ has a sharp upper bound that is inversely proportional to the minimum eigenvalue of the concentration matrix, but depends otherwise only on the covariance matrix of the reduced-form and first-stage parameters, which can be consistently estimated using HAR or cluster-robust methods. The minimum eigenvalue of the concentration matrix is therefore the object of the test, as it is sufficient to characterize the sharp upper bound on the Nagar bias. We propose a generalized test statistic that is an extension of the Cragg-Donald statistic adopted by Stock and Yogo (2005), and also nests the effective  $F$ -statistic for models with a single endogenous regressor as a special case. Our test statistic, which is the minimum eigenvalue of a matrix consisting of traces of partitions of a noncentral Wishart random matrix, has an intractable limiting distribution. Just as Stock and Yogo (2005), we therefore make use of a bounding limiting distribution. Specifically, we show that the second and third cumulants of our generalized test statistic are bounded by expressions that depend only on the robust covariance matrix of the reduced-form and first-stage parameters and the minimum eigenvalue of the concentration matrix. The right tail of a bounding limiting distribution for our test statistic can typically be approximated simply by matching the bounding cumulants using an Imhof (1961) distribution.<sup>1</sup> We provide Matlab code that exploits the specialized numerical algorithm of Wen and Yin (2013) to compute the sharp upper bound on the Nagar bias efficiently, leading to trivial computation times in typical applications. For unusually large-dimensional models, we also provide a simplified – but more conservative – procedure that avoids numerical optimization entirely.

We evaluate the performance of our test in simulations for models with various dimensions using millions of randomly generated DGPs. When the number of instruments is at least two more than the number of endogenous regressors, the sharp upper bound on the Nagar approximation to the bias is a highly effective upper bound on the Monte Carlo bias in the 2SLS estimates. This is not the case, however, in models with degree of overidentification less than two, and we therefore propose a more conservative bound in those models that is effective in bounding the Monte Carlo bias. In the simulations, we find that the size of our first-stage test is very well-controlled. Despite the fact that the critical values are based on a bounding limiting distribution for the test statistic, we also find that the power of our test rises sharply not too far

---

<sup>1</sup>The Matlab code we provide always ensures that a correct bounding distribution is selected even when this is not the case.

below the chosen bias tolerance level, for worst-case bias levels that are well above zero. Despite being conservative, our test is therefore useful in practice for discriminating between weak and strong instruments.

As an empirical application of our test, we consider the Ramey and Zubairy (2018) estimates of state-dependent government spending multipliers. Starting from existing empirical specifications for estimating spending multipliers, the authors introduce a second endogenous regressor by interacting government spending with an indicator for the state of the business cycle or for the monetary policy regime. This application is an example of how multiple endogenous regressors often arise in practice, as it is common for researchers to explore specifications with interactions between an endogenous regressor and other variables. We find that our robust test regularly leads to different conclusions regarding instrument strength than the Stock and Yogo (2005) test. When results of the Montiel Olea and Pflueger (2013) test in the separate regime subsamples are in conflict, the outcome of our generalized test for the full sample is dominated by the more weakly identified regime. Multiple endogenous regressors of course arise in many other contexts as well, including in time series, cross-sectional, and panel data models. Our generalized test should therefore be useful to empirical researchers across a broad range of applications.

Our contribution to the literature is to generalize the bias-based tests of Stock and Yogo (2005) and Montiel Olea and Pflueger (2013). In conditionally homoskedastic and serially uncorrelated models, our first-stage test statistic is the same as that in Stock and Yogo (2005), and numerical differences with the critical values of the Stock and Yogo (2005) tables arise only because our test uses the Nagar approximation instead of Monte Carlo integration to evaluate the bias. In models with a single endogenous regressor, our test-statistic and critical values are equivalent to Montiel Olea and Pflueger (2013), except in models with two instruments, as the potential poor quality of the Nagar approximation leads us to prefer a more conservative bound on the bias.

The test proposed in this paper differs from several other existing approaches to evaluating instrument strength. Stock and Yogo (2005) also propose a size-based test for conditionally homoskedastic and serially uncorrelated models. While developing a generalized size-based test is beyond the scope of this paper, we explore the relationship between our test statistic and the size distortion in second stage  $t$ -statistic inference in simulations. Researchers sometimes report results for the robust version of the Kleibergen and Paap (2006) test of under-identification. Our test is a test of weak instruments, not of under-identification, and therefore directly addresses concerns

about the bias of the 2SLS estimator. Relatedly, Sanderson and Windmeijer (2016) derive a bias-based test for models with multiple regressors by considering cases in which the matrix of first-stage coefficients has a rank deficiency local to unity, as opposed to a rank local to zero as in the asymptotic embedding common in the literature. As far as we are aware, the Sanderson and Windmeijer (2016) approach has only been developed for conditionally homoskedastic and serially uncorrelated models. Andrews (2018) offers a distinctly different diagnostic for identification strength, which is broadly applicable to GMM problems, whereby both robust and non-robust confidence sets are computed for the parameters of interest. The relationship between the sets, linked to the size distortion of standard inference, determines whether identification is strong enough to proceed with non-robust methods. Finally, a recent paper by Carrasco and Doukali (2021) introduces a first-stage test that is robust to heteroskedasticity and many instruments. However, as the Montiel Olea and Pflueger (2013) test, it only applies to models with a single endogenous regressor.

When conducted in conjunction with hypothesis tests for identified parameters, pretests for weak instruments form part of a multiple hypothesis testing problem, which can contribute to size distortions in inference on estimated second-stage parameters (Andrews 2018; Keane and Neal 2022; Lee et al. 2022). In applying our proposed test, we encourage empirical researchers to be mindful of this issue, and also consider robust inference procedures for second-stage parameters, particularly in cases of marginal rejections. Given the popularity of first-stage tests in applied work, we view a test that allows researchers to make consistent assumptions in both estimation stages as an important improvement over the common practice of assuming away heteroskedasticity and autocorrelation in the first stage, but not in the second.

## 1 Model and Summary of Testing Procedure

In this section, we provide a non-technical summary of the model assumptions and testing procedures. We also discuss in more detail how our test compares to those of Stock and Yogo (2005) and Montiel Olea and Pflueger (2013).

**Model** Our proposed test applies to linear instrumental variables models with  $N$  endogenous regressors and  $K \geq N$  instruments,

$$(1) \quad y = Y\beta + u,$$

$$(2) \quad Y = Z\Pi + v,$$

where  $\beta \in \mathbb{R}^N$  contains the parameters of interest and  $\Pi \in \mathbb{R}^{K \times N}$  contains the first-stage parameters. The econometrician observes  $y \in \mathbb{R}^T$ ,  $Y \in \mathbb{R}^{T \times N}$ , and  $Z \in \mathbb{R}^{T \times K}$ , where  $T$  is the sample size. Without loss of generality, we assume that there are no additional exogenous regressors, and that  $Z'Z/T = I_K$ . In the presence of additional exogenous regressors, it suffices to first project  $y$ ,  $Y$ , and  $Z$  on these regressors and replace all variables with the resulting projection errors. In addition, the formulas in this paper assume the user has subsequently normalized  $Z$  such that its sample average is zero and its sample covariance is the identity matrix.

Consider the reduced-form of (1),  $y = Z\Pi\beta + w$ , where  $w = v\beta + u$ , and assume that  $T^{-\frac{1}{2}}[Z'w \text{vec}(Z'v)']' \xrightarrow{d} \mathcal{N}(0, \mathbf{W})$ , where  $\mathbf{W}$  is the asymptotic covariance of the reduced-form and first-stage OLS coefficients, and  $\text{vec}$  is the vectorization operator. The weak instruments test of Stock and Yogo (2005) requires  $\mathbf{W}$  to be of the Kronecker form  $\Sigma_{wv} \otimes I_K$ , where  $[w \ v]'[w \ v]/T \xrightarrow{p} \Sigma_{wv}$ . The Kronecker structure arises generally only in conditionally homoskedastic and serially uncorrelated models. The purpose of this paper is to let  $\mathbf{W}$  be any positive definite matrix, thereby allowing for arbitrary distributional assumptions about the model errors. Montiel Olea and Pflueger (2013) also relax the Kronecker form assumption, but only consider models with  $N = 1$ . We defer a detailed discussion of all our assumptions to Section 2, but they are otherwise entirely analogous to Stock and Yogo (2005) and Montiel Olea and Pflueger (2013). They include the assumption that the first-stage relationship in (2) is ‘local to zero’, and that the instruments are therefore weak under the null hypothesis.

**Testing Procedure** Our generalized test for weak instruments assesses the null hypothesis that the largest possible bias associated with a given set of instruments exceeds a certain tolerance level,  $\tau$ , relative to a worst-case benchmark for the 2SLS bias over all possible instruments. This benchmark is defined exactly as in Montiel Olea and Pflueger (2013), and considers a scenario in which all first-stage parameters are zero,  $\Pi = 0$  and the errors  $u$  are a perfect linear combination of the second-stage regressors  $\hat{Y}$ . In conditionally homoskedastic and serially correlated models, the bias criterion has the familiar interpretation from Stock and Yogo (2005) as the 2SLS bias relative to the OLS bias. In general, however, the bias does not have this interpretation.

In practice, our test for weak instruments consists of the following steps:

1. Replace all variables with the residuals after projection on any additional exogenous regressors. Standardize the instruments such that the sample average is zero and  $Z'Z/T = I_K$ .

2. Regress  $y$  on  $Z$  (reduced-form) and  $Y$  on  $Z$  (first-stage) and obtain the residuals  $\hat{w}_t$  and  $\hat{v}_t$ , respectively. Compute the robust covariance matrix of choice (e.g., heteroskedasticity-robust, heteroskedasticity-and-autocorrelation robust, or clustered),  $\hat{\mathbf{W}}$ , for the vector  $(\hat{w}_t, \hat{v}_t)' \otimes Z_t$ , or, equivalently, the reduced-form and first-stage parameters.
3. Compute the test statistic,  $g_{\min}$ ,

$$(3) \quad g_{\min} = T^{-1} \text{mineval}\{\hat{\Phi}^{-\frac{1}{2}} Y' Z Z' Y \hat{\Phi}^{-\frac{1}{2}}\},$$

where  $\hat{\Phi} = (I_N \otimes \text{vec}(I_K))' (\hat{\mathbf{W}}_2 \otimes I_K) (I_N \otimes \text{vec}(I_K))$ ,  $\hat{\mathbf{W}}_2$  is the estimated covariance matrix of the first-stage coefficients (the lower  $NK \times NK$  diagonal block of  $\hat{\mathbf{W}}$ ), and  $\text{mineval}\{\cdot\}$  is the smallest eigenvalue.

4. Compare  $g_{\min}$  to the application-specific critical value for a user-supplied relative bias threshold,  $\tau$ , (e.g., 0.10) and a desired significance level,  $\alpha$  (e.g. 0.05). If  $g_{\min}$  exceeds the critical value, reject the null hypothesis of weak instruments.

We provide a Matlab file, `gweakivtest.m`, to implement steps 1-4 using the estimation data as input, with several standard options for obtaining  $\hat{\mathbf{W}}$ . The Matlab function `gweakivtest_critical_values.m` calculates the critical value in step 4 for a given  $\hat{\mathbf{W}}$ ,  $\tau$  and  $\alpha$ . The function can also be applied with a different user-supplied estimator of  $\hat{\mathbf{W}}$  if preferred, allowing the researcher to conduct steps 1-3 themselves.

The calculation of the critical value involves a numerical optimization step to obtain the sharp upper bound of the Nagar bias. In the vast majority of applications, the computation time will be trivial. However, as in Montiel Olea and Pflueger (2013), we also provide a simplified conservative version of the test for very large-dimensional models. The simplified version of the test follows the same steps but replaces the sharp upper bound for the Nagar bias with an easier-to-compute non-sharp bound. This version is therefore computationally much faster in models with a very large number of endogenous variables and instruments.

**Comparison with Existing Tests** In conditionally homoskedastic and serially uncorrelated models, the test statistic  $g_{\min}$  in (3) reduces to the Cragg and Donald (1993) statistic of the Stock and Yogo (2005) test, and in models with  $N = 1$  it reduces to the effective  $F$ -statistic of Montiel Olea and Pflueger (2013). Figure 1 illustrates the critical values for  $\alpha = 0.05$  and  $\tau = 0.10$  generated by our code in those special cases, as well as in more general models with  $N > 1$  and an arbitrary covariance matrix  $\mathbf{W}$ . The panels in the first

column of Figure 1 show critical values for models with  $N = 1, 2$  and  $3$  endogenous regressors for the homoskedastic and serially uncorrelated model. The figure reports the critical values based on the sharp bound on the Nagar bias, as well as those from the more conservative simplified procedure. Each panel in the left column also plots the critical values from the Stock and Yogo (2005) tables for comparison, which are available for  $K > N + 1$ . The top left panel showing the  $N = 1$  case also reports the critical values from the Montiel Olea and Pflueger (2013) test for  $K \geq 1$ , as well as the analytical critical values derived for  $K > 1$  by Skeels and Windmeijer (2018).

In the conditionally homoskedastic and serially uncorrelated model, the critical values only depend on the number of endogenous regressors  $N$  and the number of instruments  $K$ . In the general model, the critical values depend additionally on the covariance  $\mathbf{W}$ . The critical values are therefore different for each regression. To nevertheless give a sense of the critical values that arise in practice, the right column in Figure 1 shows the average robust critical values across 500 different general covariance matrices drawn randomly from a central Wishart distribution with an identity covariance matrix. The top panel for the model with  $N = 1$  additionally reports the robust Montiel Olea and Pflueger (2013) critical values.

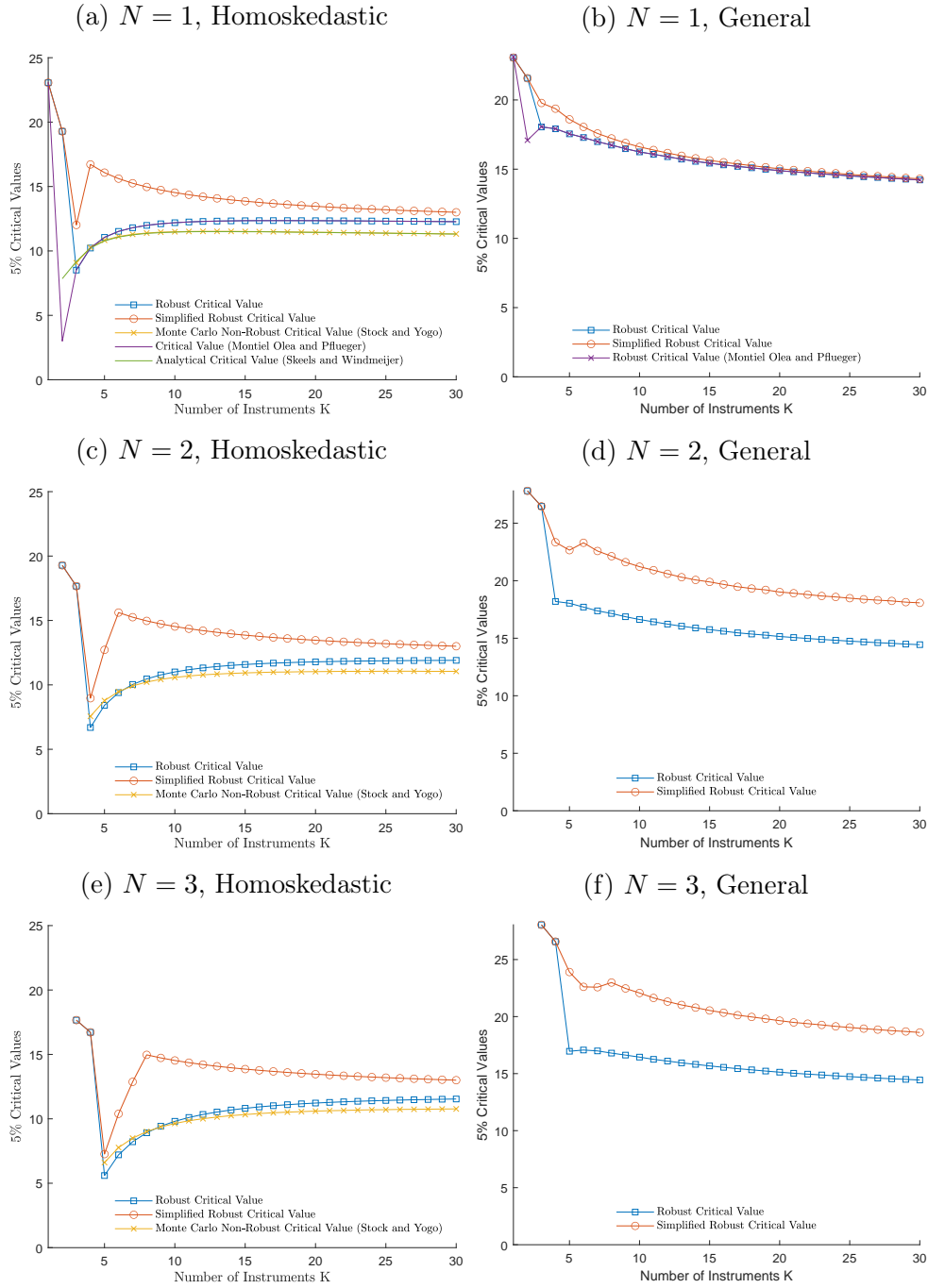
The main takeaways from Figure 1 are as follows. For models with a single endogenous regressor ( $N = 1$ ) – either homoskedastic or with general  $\mathbf{W}$  – there are only negligible numerical differences between our critical values and those of Montiel Olea and Pflueger (2013).<sup>2</sup> The only exception is when  $K = 2$ , since we conduct the test with a more conservative bound on the bias whenever  $K \leq N + 1$ . We discuss the need for a more conservative bound in more detail later on. However, that the sharp upper bound on the Nagar bias is generally not adequate can already be seen in the conditionally homoskedastic and serially uncorrelated model with  $N = 1$  in the top left panel of Figure 1. This panel shows that the Montiel Olea and Pflueger (2013) critical value for  $K = 2$  is far below the analytical critical value obtained by Skeels and Windmeijer (2018). In the  $N = 1, K = 1$  model, our alternative upper bound coincides with the sharp upper bound on the Nagar bias, such that in practice the only difference with Montiel Olea and Pflueger (2013) occurs in the  $K = 2$  case. Across the randomly generated  $\mathbf{W}$ 's, the critical values for the models with general  $\mathbf{W}$  are higher on average than the conventional rule-of-thumb value of 10. Note, however, that these higher critical values must be applied

---

<sup>2</sup>When  $K \neq 2$ , these differences arise only because we use the Imhof (1961) approximation, matching the first three cumulants of the target distribution, instead of the Patnaik (1949) approximation, matching the first two cumulants. We consistently found these differences to be very small.



Figure 1: Comparison of Critical Values



*Notes:* The left column reports critical values for  $\alpha = 0.05$  and  $\tau = 0.10$  assuming conditional homoskedasticity and no serial correlation for various numbers of endogenous regressors ( $N$ ) and instruments ( $K$ ). The right column repeats the exercise under arbitrary heteroskedasticity and/or autocorrelation. Critical values depend on  $\mathbf{W}$  and therefore vary for each application. The figures show averages over 500 draws of  $\mathbf{W}$  from a central Wishart distribution with identity covariance matrix for illustrative purposes only. For comparison, we plot applicable critical values from Montiel Olea and Pflueger (2013), Stock and Yogo (2005), and Skeels and Windmeijer (2018).

to  $g_{\min}$  (or equivalently the effective  $F$ -statistic when  $N = 1$ ), and not to the regular  $F$ -statistic (for any  $K$ ) or robust  $F$ -statistic (unless  $K = 1$ ).

In homoskedastic models our critical values are close to those of Stock and Yogo (2005), which remain roughly around the rule-of-thumb value of 10 regardless of  $N$  and  $K$ . The numerical differences between the critical values in this case arise because our test is based on a Nagar approximation to the bias rather than on Monte Carlo integration. Despite these numerical differences, in homoskedastic models our test is a close equivalent of the Stock and Yogo (2005) test.

For models with  $N > 1$  and general  $\mathbf{W}$ 's, the average critical values are higher than the conventional rule-of-thumb threshold for homoskedastic model of roughly 10, just as in the general model with  $N = 1$ . We emphasize again, however, that these critical values must be applied to the  $g_{\min}$  statistic for non-homoskedastic models in (3), and not to the Cragg and Donald (1993) statistic. We also reiterate that the critical values depend on  $\mathbf{W}$ , and will therefore be different in each application. A widely applicable rule-of-thumb critical value such as in homoskedastic models is therefore not available in the general setting.

For illustrative purposes, Figure 1 also plots the more conservative critical values from our simplified procedure. We generally do not recommend using these values, as the tighter values will typically be sufficiently easy to compute. Figure 1 suggests that they may nevertheless be of use in applications where  $K$  or  $N$  are much larger.

In the special case of a just-identified model with a single endogenous regressor ( $N = K = 1$ ), our test statistic  $g_{\min}$  is also equal to the robust  $F$ -statistic. As noted in Andrews et al. (2019), an alternative valid approach when  $N = K = 1$  is to evaluate the robust  $F$ -statistic against the critical values of the size-based test of Stock and Yogo (2005) (16.38 for 5% distortion, 8.96 for a 10% distortion). Unfortunately, this option is not available for  $N = K > 1$ .

## 2 Testing the Null Hypothesis of Weak Instruments

This section presents the key results underlying our proposed test. Before doing so, we establish some specific notation:  $\|U\|_2$  is the spectral norm of  $U$  (the positive square root of the maximum eigenvalue of  $UU'$ , also the  $\ell_2$ -norm if  $U$  is a vector),  $\mathbb{P}^n$  is the set of positive definite  $n \times n$  matrices,  $\mathbb{O}^{n \times m}$  is the set of  $n \times m$  orthogonal real matrices  $U$  such that  $UU' = I_n$ ,  $\mathcal{K}_{n,m}$  denotes the  $n \times m$  commutation matrix such that  $\mathcal{K}_{n,m} \text{vec}(U) = \text{vec}(U')$

where  $U \in \mathbb{R}^{n \times m}$ . We also define the special matrix  $R_{n,m} = I_n \otimes \text{vec}(I_m)$ . The dimension of  $R_{n,m}$  is  $nm^2 \times n$ . For  $U \in \mathbb{R}^{nm \times nm}$ , the  $(i, j)$ -th element of  $V = R'_{n,m}(U \otimes I_m)R_{n,m} \in \mathbb{R}^{n \times n}$  is  $\text{Tr}(U_{ij})$  where  $U_{ij} \in \mathbb{R}^{m \times m}$  is  $(i, j)$ -th block of  $U$  and  $\text{Tr}(\cdot)$  is the trace. For  $U \in \mathbb{R}^{nm \times m}$ , the  $i$ -th element of  $V = R'_{n,m} \text{vec}(U') \in \mathbb{R}^n$  is equal to  $\text{Tr}(U_i)$  where  $U_i \in \mathbb{R}^{m \times m}$  is the  $i$ -th row block of  $U$ . Note that  $R'_{n,m}R_{n,m} = mI_n$ .

## 2.1 Weak IV Asymptotic Representation of the 2SLS Estimator

The 2SLS estimator for the model in (1)–(2) is

$$(4) \quad \hat{\beta}_{2SLS} = (Y'P_Z Y)^{-1} Y' P_Z y,$$

where  $P_Z = ZZ'/T$  is the projection matrix given the normalization of the instruments. Following Staiger and Stock (1997) and the subsequent literature, we model weak instruments by assuming the first-stage relationship is local-to-zero,

**Assumption 1.**  $\Pi = C/\sqrt{T}$  where  $C \in \mathbb{R}^{K \times N}$  is a fixed full rank matrix.

The following assumptions enable us to characterize the weak instrument asymptotic distributions of the 2SLS estimator.

**Assumption 2.** The following limits hold as  $T \rightarrow \infty$ :

$$(2.a) \quad u'u/T \xrightarrow{p} \sigma_u^2 \in \mathbb{R}_+, \quad v'u/T \xrightarrow{p} \sigma_{vu} \in \mathbb{R}^N, \quad v'v/T \xrightarrow{p} \Sigma_v \in \mathbb{P}^N,$$

$$(2.b) \quad T^{-\frac{1}{2}} \begin{bmatrix} Z'w \\ \text{vec}(Z'v) \end{bmatrix} \xrightarrow{d} \mathcal{N}(0, \mathbf{W}),$$

$$(2.c) \quad \text{and } \hat{\mathbf{W}} \xrightarrow{p} \mathbf{W},$$

where  $\mathbf{W} = \begin{bmatrix} \mathbf{W}_1 & \mathbf{W}_{12} \\ \mathbf{W}'_{12} & \mathbf{W}_2 \end{bmatrix} \in \mathbb{P}^{(N+1)K}$ .

The 2SLS estimator in (4) can be written as

$$(5) \quad \hat{\beta}_{2SLS} = (R'_{N,K}(s_{ZY}s'_{ZY} \otimes I_K)R_{N,K})^{-1} R'_{N,K} \text{vec}(s_{ZY}s'_{ZY}),$$

where  $s_{ZY} = T^{-\frac{1}{2}}Z'y$  and  $s_{ZY} = T^{-\frac{1}{2}}\text{vec}(Z'Y)$ . This more complicated expression reformulates the 2SLS estimator in terms of random vectors whose asymptotic distributions are given in Assumption 2. Specifically, define the random variables  $\eta_1, \eta_2$  ( $K \times 1$  and  $NK \times 1$  respectively) as

$$(6) \quad \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mathbf{0}_K \\ \text{vec}(C') \end{pmatrix}, \mathbf{S}\right)$$

where  $\mathbf{S} \in \mathbb{P}^{(N+1)K}$ , partitioned as  $\mathbf{W}$  with

$$(7) \quad \begin{aligned} \mathbf{S}_1 &= \mathbf{W}_1 + (\beta' \otimes I_K)\mathbf{W}_2(\beta \otimes I_K) - (\beta' \otimes I_K)\mathbf{W}'_{12} - \mathbf{W}_{12}(\beta \otimes I_K), \\ \mathbf{S}_{12} &= \mathbf{W}_{12} - (\beta' \otimes I_K)\mathbf{W}_2, \quad \mathbf{S}_2 = \mathbf{W}_2, \end{aligned}$$

such that  $\mathbf{S}$  is the covariance of  $T^{-\frac{1}{2}}[Z'u \text{ vec}(Z'v)']'$  as  $T \rightarrow \infty$ .

Proposition 1 characterizes the distribution of the random variable  $\beta_{2SLS}^* = \hat{\beta}_{2SLS} - \beta$ .

**Proposition 1.** *Under Assumptions 1 and 2,  $s_{ZY} \xrightarrow{d} \eta_2$  and  $s_{Zy} \xrightarrow{d} (\beta' \otimes I_K)\eta_2 + \eta_1$ , and thus*

$$\hat{\beta}_{2SLS} - \beta \xrightarrow{d} \beta_{2SLS}^* = (R'_{N,K}(\eta_2\eta'_2 \otimes I_K)R_{N,K})^{-1} R'_{N,K} \text{vec}(\eta_1\eta'_2).$$

*Proof.* The proposition follows directly from the stated assumptions, the expression for  $\hat{\beta}_{2SLS}$  in (5), and the continuous mapping theorem.  $\square$

Since  $\beta_{2SLS}^*$  converges to a quotient of quadratic forms in normal random variables,  $\hat{\beta}_{2SLS}$  is not a consistent estimator of  $\beta$  under the local-to-zero Assumption 1. The asymptotic bias of the 2SLS estimator is the expected value  $E[\beta_{2SLS}^*]$ , which has no known analytical form except for the special case in which  $\mathbf{W}$  has the Kronecker form and  $N = 1$  and  $K > 1$ , see Skeels and Windmeijer (2018). Note that when  $K = N$  – and depending on assumptions also when  $K = N + 1$  – the expected value  $E[\beta_{2SLS}^*]$  generally does not exist.<sup>3</sup>

Finally, we define the concentration matrix for models with general  $\mathbf{W}$  and  $N \geq 1$ . Under Assumptions 1 & 2,  $Y'P_Z Y/T \xrightarrow{p} C'C + R'_{N,K}(\mathbf{S}_2 \otimes I_K)R'_{N,K}$  which naturally leads to the following definition.

**Definition 1.** *The concentration matrix for (2) is*

$$\Lambda = \Phi^{-\frac{1}{2}}C'C\Phi^{-\frac{1}{2}}$$

where  $\Phi = R'_{N,K}(\mathbf{S}_2 \otimes I_K)R_{N,K}$ .

This concentration matrix nests that of previous approaches as follows: When  $\mathbf{W}$  has the Kronecker form,  $\mathbf{S}_2 = \Sigma_v \otimes I_K$  and  $\Phi = K\Sigma_v$  such that  $\Lambda = K^{-1}\Sigma_v^{-\frac{1}{2}}C'C\Sigma_v^{-\frac{1}{2}}$  which is the concentration matrix in Stock and Yogo (2005). For general  $\mathbf{W}$  and  $N = 1$ ,  $\Lambda = \|C\|_2^2 \text{Tr}(\mathbf{W}_2)^{-1}$ , which is the concentration parameter in Montiel Olea and Pflueger (2013).

---

<sup>3</sup>See, for example, Basmann (1961), Mariano (1972), Kinal (1980), Phillips (1980), or Skeels and Windmeijer (2018)

## 2.2 Definition of Weak Instruments

As in Montiel Olea and Pflueger (2013) and Stock and Yogo (2005), our definition of weak instruments is based on a bias criterion. Specifically, we consider instruments weak when a weighted  $\ell_2$  norm of the asymptotic bias  $E[\beta_{2SLS}^*]$  is large relative to a worst-case benchmark.

**Definition 2.** *The bias criterion is*

$$(8) \quad B = \text{Tr}(\mathbf{S}_1)^{-\frac{1}{2}} \|E[\beta_{2SLS}^*]' \Phi^{\frac{1}{2}}\|_2.$$

As in Stock and Yogo (2005), the  $\ell_2$  norm in the bias criterion aggregates the  $N$  elements of the bias through a quadratic loss function, such that  $B$  is weakly positive and penalises larger biases more heavily. The criterion applies a weighting matrix,  $\Phi$ , to put the elements of  $E[\beta_{2SLS}^*]$  on a comparable scale. The weighting matrix  $\Phi$  is the asymptotic covariance of the first-stage parameter estimates and effectively standardizes the regressors in the second stage, so that they have unit standard deviation and are orthogonal. The bias criterion also scales by  $\text{Tr}(\mathbf{S}_1)$ , which is the probability limit of  $T^{-1}u'P_Zu$ . This scaling expresses  $B$  as a ratio, relative to the same worst-case bias as in Montiel Olea and Pflueger (2013). The intuition for the worst-case bias is given by the ad-hoc approximation of  $E[\beta_{2SLS}^*]$  in terms of a ratio of expectations as in Staiger and Stock (1997):

$$(9) \quad E[\beta_{2SLS}^*] \approx \frac{\text{vec}(\mathbf{S}_{12})' R_{N,K} \Phi^{-\frac{1}{2}}}{\text{Tr}(\mathbf{S}_1)^{\frac{1}{2}}} (I_N + \Lambda)^{-1} \Phi^{-\frac{1}{2}} \text{Tr}(\mathbf{S}_1)^{\frac{1}{2}}$$

Using this approximation, the bias criterion in Definition 2 reaches a maximum of unity when the error term  $u$  is a perfect linear combination of the second stage regressors,  $\hat{Y}$  such that the first term in (9) is an  $N \times 1$  unit vector, and when the instruments are completely uninformative and the concentration matrix  $\Lambda$  is zero. The bias criterion in Definition 2 nests that in Montiel Olea and Pflueger (2013) as a special case.<sup>4</sup> In the conditionally homoskedastic and serially uncorrelated model, the bias criterion becomes the same as that of Stock and Yogo (2005). Specifically, when  $\mathbf{W}$  has the Kronecker form,  $\Phi = K\Sigma_v$  and  $\text{Tr}(\mathbf{S}_1) = \sigma_u$  which are the weighting matrix and scaling factor used in the criterion of Stock and Yogo (2005). In homoskedastic models, the bias criterion has the interpretation as the 2SLS bias relative to the OLS bias. In models where  $\mathbf{W}$  does not have the Kronecker form,  $B$

---

<sup>4</sup>When  $N = 1$ ,  $B = E[\beta_{2SLS}^*] \sqrt{\text{Tr}(\mathbf{S}_2)} / \sqrt{\text{Tr}(\mathbf{S}_1)}$ , which is identical to the criterion in Montiel Olea and Pflueger (2013) after replacing  $E[\beta_{2SLS}^*]$  with a Nagar (1959) approximation.

does not have the same interpretation.

The weak instrument set is defined using the bias criterion in Definition 2:

**Definition 3.** *The weak instrument set is*

$$\mathbb{B}_\tau(\mathbf{W}) = \{C \in \mathbb{R}^{N \times K}, \beta \in \mathbb{R}^N : B > \tau\}.$$

The weak instrument set is the set of values for  $\beta$  and the first-stage parameters  $C$  such that bias  $B$  exceeds a tolerance level  $\tau$ . This set depends on  $\mathbf{W}$ , which can be consistently estimated, but also on the  $NK$  parameters in  $C$ , and the  $N$  unknown parameters in  $\beta$ .

### 2.3 Characterizing the Boundary of the Weak Instrument Set

A practical generalized test for weak instruments requires a scalar sufficient statistic that is informative for whether  $C$  lies within the weak instrument set for any possible value of  $\beta$ . Moreover, the boundary of the weak instrument set over the remaining parameters must be relatively easy to compute since – unlike in the conditionally homoskedastic and serially uncorrelated model – this boundary depends on  $\mathbf{W}$  and is therefore model-specific. To address both of these challenges, we follow Montiel Olea and Pflueger (2013) and adopt an analytical approximation to the bias based on Nagar (1959).

We start with the following lemma presenting a useful decomposition of the bias criterion in Definition 2.

**Lemma 1.** *Under Assumptions 1 & 2, the bias criterion in Definition 2 can be decomposed as  $B = \|\mathbf{h}\rho\|_2$ , where*

$$\mathbf{h} = KE \left[ (R'_{N,K}(\mathcal{S}(l + \psi)(l + \psi)' \mathcal{S}' \otimes I_K) R_{N,K})^{-1} R'_{N,K}(\mathcal{S}(l + \psi)\psi' \mathcal{S}^{-1} \otimes I_K) \right],$$

$$\rho = (\Phi^{-\frac{1}{2}} \otimes I_{K^2}) \text{vec}(\mathbf{S}_{12}) / \sqrt{\text{Tr}(\mathbf{S}_1)},$$

$$l = \mathbf{S}_2^{-\frac{1}{2}} \text{vec}(C'), \quad \psi = \mathbf{S}_2^{-\frac{1}{2}}(\eta_2 - \text{vec}(C')) \sim \mathcal{N}(0, I_{NK}), \quad \text{and } \mathcal{S} = ((\Phi/K)^{-\frac{1}{2}} \otimes I_K) \mathbf{S}_2^{\frac{1}{2}}.$$

*Proof.* See Appendix A. □

Lemma 1 formulates the bias as the  $\ell_2$  norm of the product of the  $N \times NK^2$  matrix  $\mathbf{h}$  and the  $NK \times 1$  vector  $\rho$ . The matrix  $\mathbf{h}$  is the expected value of a random matrix that is a function of  $\psi$ , a vector of i.i.d standard normal variables. This expected value – when it exists – also depends on location parameters  $C$  and on  $\mathbf{W}_2$ . The vector  $\rho$  depends on  $\mathbf{W}$  and  $\beta$ .

In general, there is no tractable analytical expression for the integral underlying the expectation in  $\mathbf{h}$ , which is required to evaluate the bias. Whereas Stock and Yogo (2005) evaluate this integral using Monte Carlo methods, we adopt a Nagar (1959) approximation to  $\mathbf{h}$ , which we denote by  $\mathbf{h}_n$ . This leads to the following definition.

**Definition 4.** *The Nagar bias is  $B_n = \|\mathbf{h}_n\rho\|_2$ , where  $h_n$  is a Nagar (1959) approximation of  $\mathbf{h}$  around  $\psi = 0$ .*

The Nagar approximation  $\mathbf{h}_n$  is derived analytically in Appendix B. The Nagar bias  $B_n$  still depends on  $C$ ,  $\beta$ , and  $\mathbf{W}$ . Unlike the original bias criterion in Definition 2, the bias criterion under the Nagar approximation exists even when  $K \leq N+1$  and  $E[\beta_{2SLS}^*]$  does not exist. This motivates Montiel Olea and Pflueger (2013) to expand the use of their test to models with  $K \leq N+1$ , for which Stock and Yogo (2005) do not report critical values. We will proceed as in Montiel Olea and Pflueger (2013) and also consider models with  $K \leq N+1$ .

The following theorem provides upper bounds on the Nagar bias for a given minimum eigenvalue of the concentration matrix,  $\lambda_{\min}$ .

**Theorem 1.** *The Nagar bias has the following bounds:*

- (i)  $B_n \leq B_n^*(\mathbf{W}, \lambda_{\min}) = \lambda_{\min}^{-1} \mathcal{B}(\mathbf{W})$ ,  $\mathcal{B}(\mathbf{W}) = K^{-\frac{1}{2}} \sup_{L_0 \in \mathbb{O}^{N \times K}} \{ \|M_1(I_N \otimes L_0 \otimes L_0)M_2\Psi\|_2 \}$ ,
- (ii)  $B_n^*(\mathbf{W}, \lambda_{\min}) \leq \lambda_{\min}^{-1} \mathcal{B}^s(\mathbf{W})$ ,  $\mathcal{B}^s(\mathbf{W}) = \min\{(2(N+1)/K)^{\frac{1}{2}} \|M_2\Psi\|_2, \|\Psi\|_2\}$ ,

where  $\lambda_{\min} = \text{mineval}\{\Lambda\}$ ,  $\Psi = (\mathcal{S}\mathbf{W}_2^{-\frac{1}{2}}[\mathbf{W}_{12} : \mathbf{W}_2]' \otimes I_K)R_{N+1,K}(R'_{N+1,K}(\mathbf{W} \otimes I_K)R_{N+1,K})^{-\frac{1}{2}}$ ,  $M_1 = R'_{N,N}(I_{N^3} + (\mathcal{K}_{N,N} \otimes I_N))$ , and  $M_2 = R_{N,K}R'_{N,K}/(N+1) - I_{NK^2}$ .

*Proof.* See Appendix B. □

Part (i) of the theorem characterizes a sharp upper bound on the Nagar bias,  $B_n^*(\mathbf{W}, \lambda_{\min})$ , that only depends on  $\mathbf{W}$  and  $\lambda_{\min}$ , the minimum eigenvalue of the concentration matrix in Definition 1. The sharp upper bound, which we will refer to as the ‘worst-case Nagar bias’, is the product of  $\lambda_{\min}^{-1}$  and  $\mathcal{B}(\mathbf{W})$ . The latter is equal to the largest possible singular value of a matrix that depends on the  $NK - (N+1)/2$  nuisance parameters in the orthogonal matrix  $L_0$ , which needs to be obtained numerically. Part (ii) of the theorem provides an alternative – but generally non-sharp – upper bound that requires no numerical optimization.

Theorem 1 addresses the two main practical challenges for a test of weak instruments in models with multiple endogenous variables. First, it provides

a straightforward mapping between the worst-case Nagar bias and a single parameter,  $\lambda_{\min}$ :  $B_n^*(\mathbf{W}, \lambda_{\min}) = \lambda_{\min}^{-1} \mathcal{B}(\mathbf{W})$ . The proof of the theorem shows that the Nagar bias under the worst-case scenario for  $\beta$  is non-increasing in all eigenvalues of the concentration matrix, such that the worst-case Nagar bias occurs when all eigenvalues of  $\Lambda$  equal  $\lambda_{\min}$ . The Nagar approximation therefore provides an analytical justification for  $\lambda_{\min}$  as the object of the test, whereas the parallel justification in Stock and Yogo (2005) is based on simulation evidence and the many-weak instrument limit.

Second, the numerical optimization problem underlying  $\mathcal{B}(\mathbf{W})$  is relatively straightforward. The general problem of optimizing  $B_n$  over  $\beta$  and  $C$  subject to  $\min\{\Lambda\} = \lambda_{\min}$  is problematic because of the large dimension and the presence of many local maxima. The proof in Appendix B shows that optimizing over  $\beta$  reduces to a straightforward maximum eigenvalue problem. Moreover, when all eigenvalues of  $\Lambda$  are equal to  $\lambda_{\min}$ , the bias only additionally depends on  $L_0$ . Optimization over  $L_0$  has smaller dimension and, importantly, can exploit numerical algorithms specialized for orthogonal matrices. In the code accompanying this paper, we use the curvilinear search algorithm of Wen and Yin (2013) which leads to trivial computation times even for relatively large  $N$  and  $K$ . When  $N$  and/or  $K$  are so large that the optimization becomes prohibitive, the non-sharp bound in part (ii) of Theorem 1 can be used instead. Finally, because the bounds are inversely proportional to the minimum eigenvalue of the concentration matrix, the threshold value of that minimum eigenvalue for a given bias tolerance level  $\tau$  is given simply by  $\mathcal{B}(\mathbf{W})/\tau$  (or  $\mathcal{B}^s(\mathbf{W})/\tau$ ), and requires no additional root-finding operation.

The computational advantages of working with the Nagar approximation comes at the cost of a potential loss in accuracy relative to using Monte Carlo integration to evaluate  $\mathbf{h}$ , as in Stock and Yogo (2005). Monte Carlo integration, however, does not lead to the same simplifications in the optimization over the nuisance parameters, and requires an additional root-finding operation to find the threshold value of the minimum eigenvalue. Because of the dependence on  $\mathbf{W}$ , this threshold value must also be computed separately for each application, and cannot be tabulated in advance as a function of only  $K$  and  $N$ , unlike when  $\mathbf{W}$  has the Kronecker form. In the Online Appendix, we compare the Nagar bias to the bias computed using Monte Carlo integration across millions of DGPs and for various values of  $N$  and  $K$ . The main finding is that the Nagar bias is generally close to the bias obtained via numerical integration. Importantly, the Nagar approximation is especially accurate in the most relevant range for the bias tolerance in practice, that is in a neighborhood of 0.10.



When the degree of overidentification is less than two,  $K \leq N + 1$ , the accuracy of the Nagar approximation – as judged by its relationship with the Monte Carlo bias – can deteriorate sharply for some models. An extreme example is the conditionally homoskedastic and non-serially correlated model with  $K = N + 1$ . When  $\mathbf{W}$  has the Kronecker form, the sharp upper bound on the Nagar bias simplifies to  $B_n^*(\mathbf{W}, \lambda_{\min}) = \lambda_{\min}^{-1}|K - (N + 1)|/K$ . For  $K = N + 1$ , this means that the Nagar bias is always zero.<sup>5</sup> This problematic feature of the worst-case Nagar bias is the reason why in panel (a) of Figure 1 (the  $N = 1$  case) the Montiel Olea and Pflueger (2013) critical value at  $K = 2$  is sharply lower compared to the analytical value of Skeels and Windmeijer (2018). Given these problems, which the Online Appendix shows can also arise for general covariance  $\mathbf{W}$ , we do not rely on the Nagar approximation whenever the degree of overidentification is one. Instead, when  $K = N + 1$ , we use  $\|\Psi\|_2/\tau$  as the threshold value for the minimum eigenvalue of the concentration matrix. We use the same bound in just-identified models with  $N = K$ , even though the integral underlying the expectation in the definition of  $\mathbf{h}$  in Lemma 1 does not converge in that case. When  $N = K = 1$ , the more conservative bound always coincides with the sharp bound on the Nagar bias, such that the choice for the bound is irrelevant in that case.

## 2.4 Test Statistic and Critical Values

Given a bias tolerance level  $\tau$ , a test of the null hypothesis of weak instruments can be based on a test of whether the minimum eigenvalue of  $\Lambda$  is smaller or equal to a threshold value  $\lambda_{\min}^*(\tau)$ . More formally, the null and alternative hypotheses for the test are

$$(10) \quad H_0 : \lambda_{\min} \in \mathcal{H}(\mathbf{W}) \quad \text{vs.} \quad H_1 : \lambda_{\min} \notin \mathcal{H}(\mathbf{W}),$$

where  $\mathcal{H}(\mathbf{W}) = \{\lambda_{\min} \in \mathbb{R}_+ : \lambda_{\min} \leq \lambda_{\min}^*(\tau)\}$ ,

where  $\lambda_{\min}^*(\tau) = \mathcal{B}(\mathbf{W})/\tau$  (or  $\|\Psi\|_2/\tau$  when  $K \leq N + 1$ , and  $\mathcal{B}^s(\mathbf{W})/\tau$  in the case of the simplified test). The null hypothesis is that the minimum eigenvalue of the concentration matrix is in the set of values for which the selected upper bound on the bias is greater than the tolerance level  $\tau$ . Under the alternative, the minimum eigenvalue is not in that set of values.

The following proposition presents our statistic to test the null hypothesis.

---

<sup>5</sup>The matrix  $M_2$  in the Nagar approximation loses rank when  $K = N + 1$ , which translates to a Nagar bias of zero when  $\mathbf{W}$  has the Kronecker form.

**Proposition 2.** *Define the test statistic*

$$g_{\min} = \text{mineval}\{\hat{\Phi}^{-\frac{1}{2}}(Y'P_ZY)\hat{\Phi}^{-\frac{1}{2}}\},$$

where  $\hat{\Phi} = R'_{N,K}(\hat{\mathbf{W}}_2 \otimes I_K)R_{N,K}$ . Then, under Assumptions 1 and 2,

$$g_{\min} \xrightarrow{d} \text{mineval}\{R'_{N,K}(\zeta \otimes I_K)R_{N,K}/K\},$$

where the  $NK \times NK$  random matrix  $\zeta = \mathcal{S}(l + \psi)(l + \psi)'\mathcal{S}'$  has a noncentral Wishart distribution,  $\zeta \sim \mathcal{W}(1, \Sigma, \Omega)$ , with 1 degree of freedom, covariance matrix  $\Sigma = \mathcal{S}\mathcal{S}' \in \mathbb{P}^{NK}$ , and a matrix of noncentrality parameters  $\Omega = \Sigma^{-1}\mathcal{S}\Omega\mathcal{S}'$ .<sup>6</sup>

*Proof.* The proposition follows from Slutsky's theorem, the continuous mapping theorem, and  $Y'P_ZY \xrightarrow{d} R'_{N,K} \left( \mathbf{S}_2^{\frac{1}{2}}(l + \psi)(l + \psi)'\mathbf{S}_2^{\frac{1}{2}} \otimes I_K \right) R_{N,K}$ , which implies the stated distribution of  $\zeta$ .  $\square$

The test statistic  $g_{\min}$  is a generalization of the Cragg and Donald (1993) statistic, re-scaled appropriately to account for heteroskedasticity and autocorrelation. In the homoskedastic and serially uncorrelated model,  $\hat{\mathbf{W}}_2 = \hat{\Sigma}_v \otimes I_K$ , such that  $\hat{\Phi} = K\hat{\Sigma}_v$ , and  $g_{\min} = K^{-1} \text{mineval}\{(\hat{\Sigma}_v)^{-\frac{1}{2}}Y'P_ZY(\hat{\Sigma}_v)^{-\frac{1}{2}}\}$ , which is the Cragg and Donald (1993) statistic adopted by Stock and Yogo (2005). The test statistic  $g_{\min}$  also nests the effective  $F$ -statistic of Montiel Olea and Pflueger (2013). When  $N = 1$ ,  $\hat{\Phi} = \text{Tr}(\hat{\mathbf{W}}_2)$ , such that  $g_{\min} = Y'P_ZY / \text{Tr}(\hat{\mathbf{W}}_2)$ , which is the effective  $F$ -statistic.

While  $\zeta$  has a noncentral Wishart distribution, critical values for the test statistic  $g_{\min}$  require the distribution of  $\text{mineval}\{R'_{N,K}(\zeta \otimes I_K)R_{N,K}\}$ , which is the minimum eigenvalue of the  $N \times N$  matrix consisting of the traces of the  $K \times K$  partitions of  $\zeta$ . To the best of our knowledge, the distribution of this function of  $\zeta$  is unknown. Moreover, the limiting distribution of  $g_{\min}$  depends in general on all parameters in  $\Sigma$  and  $\Omega$ , not just on the threshold for  $\lambda_{\min}$ .

To address both these challenges, we obtain critical values from a bounding limiting distribution of  $g_{\min}$ . Specifically, we first derive upper bounds for the second and third cumulants that only depend on  $\lambda_{\min}$  and  $\mathbf{W}_2$ . Next, we consider the class of approximating distributions proposed by Imhof (1961), which match the first three cumulants of an unknown target distribution. We select the Imhof distribution with the largest critical value at significance level  $\alpha$  subject to the constraints that the first cumulant matches that of the target distribution, and that the second and third cumulants respect the analytical

---

<sup>6</sup>We adopt the notational convention of Muirhead (1982) for the noncentral Wishart distribution.

upper bounds on the cumulants of the limiting distribution of  $g_{\min}$ . The resulting critical value is guaranteed to be conservative relative to the unknown critical value from the true limiting distribution of  $g_{\min}$ .

The first step in our procedure is to derive upper bounds on the cumulants that are free of nuisance parameters. We first discuss the  $N = 1$  case, where we can rely in part on existing results in the literature. When  $N = 1$ ,  $R'_{N,K}(\zeta \otimes I_K)R_{N,K} = \text{Tr}(\zeta)$  is a scalar. The trace of a noncentral Wishart  $\zeta$  is a linear combination of noncentral  $\chi^2$  variables. While there is no tractable formula for its probability distribution that we are aware of, Mathai (1980) provides an analytical expression for the  $n$ -th order cumulant of  $\text{Tr}(\zeta)$ ,

$$(11) \quad \kappa_n = 2^{n-1}(n-1)! \left( \text{Tr}(\Sigma^n) + n \text{Tr}(\Sigma^n \Omega) \right).$$

The mean is  $\kappa_1 = K(1 + \lambda_{\min})$ , since  $\text{Tr}(\Sigma) = K$  and  $\text{Tr}(\Sigma\Omega) = K \text{Tr}(\Lambda) = K\Lambda = K\lambda_{\min}$  when  $N = 1$ . For  $n > 1$ , the cumulants are bounded by

$$(12) \quad \kappa_n \leq 2^{n-1}(n-1)! \left( \text{Tr}(\Sigma^n) + nK\lambda_{\min} \text{maxeval}\{\Sigma\}^{n-1} \right),$$

which follows from the fact that for  $U, V \in \mathbb{P}$ ,  $|\text{Tr}(UV)| \leq \text{maxeval}\{U\} \text{Tr}(V)$ , see Fact 8.12.29 in Bernstein (2009), and the fact that  $\text{Tr}(\Sigma^n \Omega) \geq 0$ . The bounds in (12) only depend on  $\lambda_{\min}$  and on  $\Sigma$ , which only depends on  $\mathbf{W}_2$  and can therefore be consistently estimated.

The general case with  $N \geq 1$  is more involved, as  $g_{\min}$  is now asymptotically distributed as the minimum eigenvalue of a matrix containing the traces of the  $K \times K$  subpartitions of  $\zeta$ . Analogously to Stock and Yogo (2005), we consider the distribution of  $\gamma' R'_{N,K}(\zeta \otimes I_K)R_{N,K}\gamma \geq \text{mineval}\{R'_{N,K}(\zeta \otimes I_K)R_{N,K}\}$  as a bounding distribution, where  $\gamma$  is the eigenvector associated with the minimum eigenvalue of  $\Lambda = R'_{N,K}(\Sigma\Omega \otimes I_K)R_{N,K}/K$  and  $\gamma'\gamma = 1$ . The following theorem extends the earlier results to  $\gamma' R'_{N,K}(\zeta \otimes I_K)R_{N,K}\gamma$ .

**Theorem 2.** For  $\zeta \sim \mathcal{W}(1, \Sigma, \Omega)$ ,

(i) The  $n$ -th cumulant of  $\gamma' R'_{N,K}(\zeta \otimes I_K)R_{N,K}\gamma$  is

$$\kappa_n = 2^{n-1}(n-1)! \left( \text{Tr} \left( ((\gamma\gamma' \otimes I_K)\Sigma)^n \right) + n \text{Tr} \left( ((\gamma\gamma' \otimes I_K)\Sigma)^n \Omega \right) \right).$$

(ii) The  $n$ -th cumulant  $\kappa_n$  with  $n > 1$  is bounded by

$$\kappa_n \leq 2^{n-1}(n-1)! \left( \text{maxeval}\{R'_{N,K}(\Sigma^n \otimes I_K)R_{N,K}\} + nK\lambda_{\min} \text{maxeval}\{\Sigma\}^{n-1} \right).$$

*Proof.* See Appendix C. □

Part (i) of the theorem generalizes Mathai (1980) to provide analytical expressions for the random scalar  $\gamma'R'_{N,K}(\zeta \otimes I_K)R_{N,K}\gamma$ . Part (ii) provides upper bounds on the cumulants that only depend on  $\lambda_{\min}$  and the consistently estimable  $\Sigma$ , such that they can be used to construct a bounding pivotal distribution. The bounds also nest those derived above for the  $N = 1$  case.

For the mean,  $n = 1$ , the expression in part (i) of the theorem again simplifies to  $\kappa_1 = K(1 + \lambda_{\min})$ , which depends only on  $\lambda_{\min}$ . The mean can therefore be targeted exactly, such that only the second and third cumulant of the approximating bounding distribution potentially differ from those of the true distribution. In the conditionally homoskedastic and serially uncorrelated model,  $\Sigma = \mathcal{S} = I_{NK}$ , and the bounds in (ii) simplify to

$$(13) \quad 2^{n-1}(n-1)!(K + nK\lambda_{\min}).$$

These are the cumulants of a noncentral chi-squared distribution with  $K$  degrees of freedom and noncentrality parameter  $K\lambda_{\min}$ , which is the exact bounding distribution for the Stock and Yogo (2005) test.

Next we define the Imhof (1961) approximating distribution matching the first three cumulants  $\kappa_n$ ,  $n = 1, 2, 3$  of the target distribution.

**Definition 5.** *The Imhof (1961) distribution is*

$$\Pr(\chi_\nu^2 < (x - \kappa_1)4\omega + \nu), \quad \nu = 8\kappa_2\omega^2, \quad \omega = \kappa_2/\kappa_3,$$

where  $\chi_\nu^2$  has a central chi-squared distribution with  $\nu$  degrees of freedom.

The following proposition is of practical use in our testing procedures.

**Proposition 3.** *The Imhof (1961) approximation in Definition 5 is conservative in the right tail when replacing the second and third cumulants by the upper bounds in Theorem 2.*

*Proof.* See Appendix D. □

The proposition implies that, for a sufficiently low significance level  $\alpha$ , it suffices to set the second and third cumulants to the upper bounds in Theorem 2 to find the most conservative critical value among all Imhof distributions with cumulants that respect the bounds. In simulations, we found that for a conventional choice of  $\alpha = 0.05$ , setting the cumulants to their upper bounds virtually always yields the most conservative critical value. Our code, however, always checks whether the Kuhn-Tucker conditions of the associated

maximization problem are satisfied at the upper bounds. If this is not the case, which may happen for example for  $\alpha = 0.10$ , the code solves for the most conservative critical value numerically, which is a relatively straightforward optimization problem.

As in Stock and Yogo (2005), the use of a bounding limiting distribution means the critical values are conservative in the sense that the null hypothesis is incorrectly rejected with probability less than or equal to  $\alpha$ . Figure 1 showed that our critical values are essentially the same as those of Montiel Olea and Pflueger (2013) when  $N = 1$ , except when  $K = 2$  where we choose not to rely on the sharp upper bound on the Nagar bias, as explained earlier. Our critical values are also close to those in the Stock and Yogo (2005) tables in the conditionally homoskedastic and serially uncorrelated model. The numerical differences with the Stock and Yogo (2005) values are almost entirely due to the Nagar approximation, since the differences between the Imhof approximation with the exact cumulants in (13) and the noncentral chi-squared bounding distribution are very small (Solomon and Stephens 1977; Bodenham and Adams 2015). The simulation study below will show that, despite the use of a bounding distribution, our test is not prohibitively conservative.

## 2.5 Summary of `gweakivtest_critical_values.m`

We conclude the discussion of our test with a description of the Matlab function `gweakivtest_critical_values.m`. To compute the critical values, the required inputs are (a) an estimate of the robust covariance of the reduced-form and first-stage coefficients,  $\hat{\mathbf{W}}$ , (b) the significance level for the test  $\alpha$ , and (c) the bias tolerance level  $\tau$ .

1. If  $K > N + 1$ , the code obtains the threshold value  $\lambda_{\min}^*(\tau) = \mathcal{B}(\hat{\mathbf{W}})/\tau$  where  $\mathcal{B}(\hat{\mathbf{W}})$  is calculated numerically using the optimization algorithm of Wen and Yin (2013). The algorithm is not guaranteed to find the global optimum. Unless the user specifies otherwise, our code takes the maximum over the optima found for 1000 starting values generated by  $N$  columns of  $K \times K$  matrices drawn from the Haar distribution, i.e. uniformly sampled from the space of all orthogonal matrices. If the user selects the simplified testing option, the threshold value is set to  $\lambda_{\min}^*(\tau) = \mathcal{B}^s(\hat{\mathbf{W}})$ . If  $K \leq N + 1$ , the threshold value is  $\lambda_{\min}^*(\tau) = \|\hat{\Psi}\|_2$ , where  $\hat{\Psi}$  is defined as in Theorem 1 after replacing  $\mathbf{W}$  with  $\hat{\mathbf{W}}$ .
2. Next, the code calculates the upper bounds on the second and third

cumulants of the bounding limiting distribution of the test statistic  $g_{\min}$ ,

$$(14) \quad \kappa_2^* = 2 \left( \text{maxeval}\{(I_N \otimes \text{vec}(I_K))'((\hat{\Sigma}^2 \otimes I_K)(I_N \otimes \text{vec}(I_K)))\} \right. \\ \left. + 2\lambda_{\min}^*(\tau)K \text{maxeval}\{\hat{\Sigma}\} \right),$$

$$(15) \quad \kappa_3^* = 8 \left( \text{maxeval}\{(I_N \otimes \text{vec}(I_K))'((\hat{\Sigma}^3 \otimes I_K)(I_N \otimes \text{vec}(I_K)))\} \right. \\ \left. + 3\lambda_{\min}^*(\tau)K \text{maxeval}\{\hat{\Sigma}\}^2 \right),$$

where  $\hat{\Sigma}$  is defined as in Proposition 2 after replacing  $\mathbf{W}$  with  $\hat{\mathbf{W}}$ .

3. The code checks whether the Kuhn-Tucker conditions in (D.35) are satisfied at  $\kappa_1 = K(1 + \lambda_{\min}^*(\tau))$ ,  $\bar{\kappa}_2 = \kappa_2^*$  and  $\bar{\kappa}_3 = \kappa_3^*$ . If not, the code solves for values  $\bar{\kappa}_2$  and  $\bar{\kappa}_3$  that do satisfy the conditions in the constrained optimization problem in (D.34) using the bounds as starting values.
4. For significance level  $\alpha$ , the code computes the critical value from the limiting distribution using the Imhof (1961) distribution in Definition 5. This critical value is divided by  $K$  to obtain the critical value that can be compared to the test statistic  $g_{\min}$ .

### 3 Simulation Study

We present simulation results for models with six different combinations for the number of endogenous variables and the number of instruments:  $N = 2$ , with  $K = 2, 3, 4, 6$ , and  $N = 3$ , with  $K = 5, 9$ . For each combination of  $N$  and  $K$ , we consider five million randomly drawn DGPs  $\{\beta, C, \mathbf{W}\}$ . We first draw 10,000  $\mathbf{W}$  matrices from a central Wishart distribution with identity covariance matrix. For each  $\mathbf{W}$ , we draw 10 different pairs of values for  $\beta$  and directions  $C_0$ , defined such that  $C = \sqrt{\lambda_{\min}}C_0$ . For each of the resulting 100,000 draws, we consider a grid of 50 minimum eigenvalues of the concentration matrix,  $\lambda_{\min}$ , over a range from 0.25 to 100. To have good coverage of the region where the Nagar bias is maximized for a given  $\lambda_{\min}$ , half of the draws for  $\beta$  and  $C_0$  are in a neighborhood of the ‘worst-case’ values  $\{\beta^{wc}, C_0^{wc}\}$  where the Nagar bias is at the upper bound for a given  $\mathbf{W}$ . The other half of the draws for  $\beta$  and  $C_0$  are from a wider region of the parameter space.<sup>7</sup>

<sup>7</sup>For  $\beta$ , the draws close to the worst-case Nagar bias are  $\beta = \beta^{wc} + 0.1v$  where  $v \sim \mathcal{N}(0, \mathcal{I}_N)$ , and the other draws comprise  $N$  independent draws from the uniform distribution on  $[-100, 100]$ . For  $C_0$ , we use the reparametrization  $\text{vec}(C_0') = \mathbf{S}_2^{\frac{1}{2}} \mathcal{S}^{-1} \sqrt{K} \text{vec}(L_0' \mathcal{D}_{\Lambda_0}^{\frac{1}{2}} Q_{\Lambda}')$ , where  $L_0$  is an orthogonal matrix as in Theorem 1, and  $\lambda_{\min} \mathcal{D}_{\Lambda_0}$  and  $Q_{\Lambda}$  contain the eigenvalues and -vectors of the concentration matrix  $\Lambda$ . The orthogonal matrix  $Q_{\Lambda}$  is always drawn from the Haar distribution. For the draws close to the worst-case Nagar bias, we set  $L_0 = ((L_0^{wc} + \xi)(L_0^{wc} + \xi)')^{-1} (L_0^{wc} + \xi)$  where the elements of  $\xi$  are drawn independently from a uniform distribution on  $[-0.1, 0.1]$  and  $L_0^{wc}$  is the orthogonal matrix that maximizes  $\mathcal{B}(\mathbf{W})$  in Theorem 1. The other draws of  $L_0$  are from the Haar distribution. Finally, the nonzero diagonal elements of  $\mathcal{D}_{\Lambda_0}$  are generated as  $1 + 0.1v$  with  $v \sim \chi^2(1)$  for the draws close to the

For each of the resulting five million DGPs, we generate 1000 samples for the random vectors  $\eta_1$  and  $\eta_2$  in (6) to draw from the limiting distribution of  $g_{\min}$  and obtain the empirical rejection rate of the first-stage test.

### 3.1 Accuracy of the Nagar Approximation

We first assess whether the worst-case Nagar bias is a valid bound for the Monte Carlo bias,  $\hat{B}$ , for each DGP, i.e. by evaluating the bias criterion in Definition 2 by replacing  $E[\beta_{2SLs}^*]$  with the sample averages of  $\beta_{2SLs}^*$ . If the worst-case Nagar bias in part (i) of Theorem 1 is a numerically accurate upper bound, the Monte Carlo bias for a given  $\beta, C, \mathbf{W}$  should not exceed the worst-case Nagar bias at the values of  $\lambda_{\min}$  and  $\mathbf{W}$  of that DGP. For the models with  $K > N + 1$ , we find that the worst-case Nagar bias is a highly effective upper bound on the Monte Carlo bias  $\hat{B}$  at a conventional bias tolerance level of  $\tau = 0.10$ . The Monte Carlo bias exceeds 0.10 in fewer than 0.001 percent of the DGPs for which the worst-case Nagar bias is smaller than 0.10. In the models with  $N = K = 2$  and  $N = 2, K = 3$ , on the other hand, the Monte Carlo bias exceeds 0.10 in 2.64 percent and 0.50 percent, respectively, of the DGPs with worst-case Nagar bias less than 0.10. These relatively frequent failures of the Nagar approximation in models with  $K \leq N + 1$  is why we adopt the more conservative upper bound in our testing procedure whenever  $K \leq N + 1$ .<sup>8</sup> Using this more conservative bound, the test fails to detect Monte Carlo bias above  $\tau = 0.10$  in only 0.15 percent of the DGPs when  $N = K = 2$ , and less than 0.0001 percent of the DGPs when  $N = 2, K = 3$ .

For a more general perspective on the quality of the Nagar approximation, the Online Appendix reports scatter plots of the Monte Carlo bias  $\hat{B}$  and the Nagar approximation  $B_n$  to the true bias  $B$  as in Definition 4. These plots show that, in the models with  $K > N + 1$ , the Nagar and Monte Carlo bias align very closely except when the bias becomes very large, as  $B_n$  is inversely proportional to  $\lambda_{\min}$  while the Monte Carlo bias is smaller than unity across all DGPs. The relationship between  $\hat{B}$  and  $B_n$ , however, is poorer when  $K \leq N + 1$ , especially for  $K = N = 2$ , and  $\hat{B}$  is no longer bounded by one.

---

worst case, and from a  $\chi^2(1)$  distribution in the other draws. In both cases, the draws are normalized such that the smallest diagonal element of  $\mathcal{D}_{\Lambda_0}$  is unity.

<sup>8</sup>The Online Appendix shows that the failures of the Nagar approximation for  $K \leq N + 1$  become even more dramatic in homoskedastic models. For example, the Nagar bias is always zero if  $K = N + 1$  and  $\mathbf{W}$  takes the Kronecker form, as discussed above.

### 3.2 Size and Power of the Generalized First-Stage Test

Figure 2 presents scatter plots of the empirical rejection rates as a function of bias across the five million DGPs. The red dots plot the rejection rates as a function of the worst-case Nagar bias (when  $K > N + 1$ ), or the more conservative bound (when  $K \leq N + 1$ ), evaluated at the values of  $\lambda_{\min}$  and  $\mathbf{W}$  in each DGP. For illustration, the blue dots also plot the rejection rates against the Monte Carlo bias given the values of  $\{\beta, \sqrt{\lambda_{\min}}C_0, \mathbf{W}\}$  in each DGP (but not taking the worst case over all possible  $\beta$  and  $C_0$ ). The horizontal lines in each panel mark the nominal size of the test,  $\alpha = 0.05$ . The vertical line marks the bias tolerance level of  $\tau = 0.10$ . These values for  $\alpha$  and  $\tau$  are the typical choices in applied use.

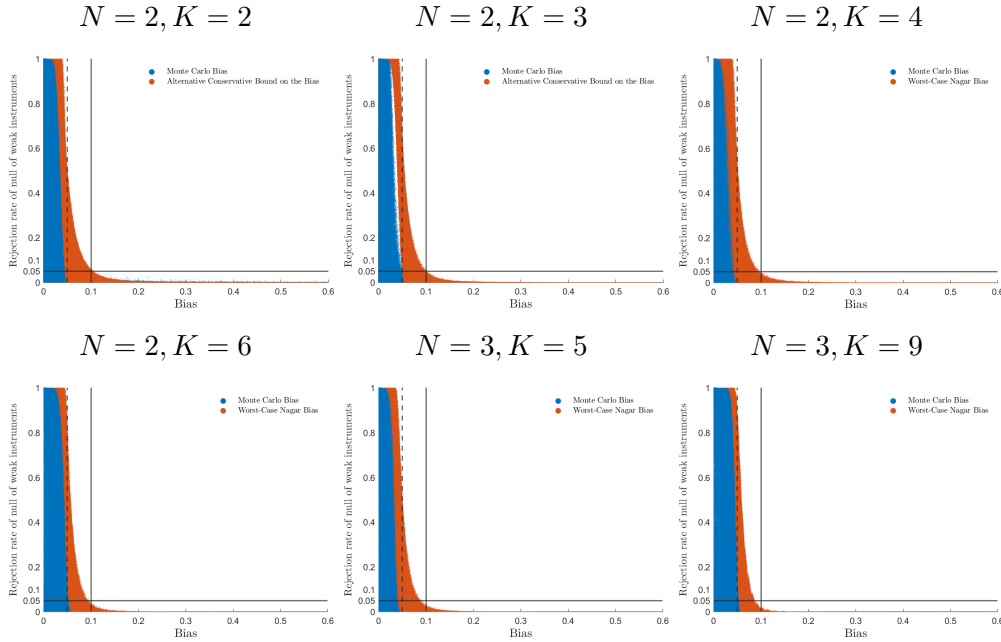
If the test is perfectly sized, the empirical rejection rates should equal the nominal size of  $\alpha = 0.05$  when the worst-case Nagar bias is precisely  $\tau = 0.10$ . Figure 2 shows that the empirical rejection rates never exceed 0.05 at worst-case Nagar bias levels of 0.10 or higher, except marginally so for a handful of DGPs for  $K = N = 2$ . Moreover, at the bias tolerance level of 0.10, the rejection rates are all strictly below 0.05. This is not surprising given our use of a bounding distribution for the test statistic  $g_{\min}$ , which implies that the test is conservative by construction. As there are no DGPs with meaningful positive size distortions, the test controls size well at the nominal level.

Despite the fact that our test is conservative, Figure 2 shows that it nevertheless has meaningful power. The dashed vertical line marks a bias level of 0.05. At that level, the rejection rates in the  $K = N + 2$  models, for example, range from 0 to 60%, while in the  $K > N + 2$  models the rejection rates range from roughly 20% to 90%. At lower – but still strictly positive – bias levels, the empirical rejection rates rise to 100% for all DGPs. The simulations therefore demonstrate that our testing procedure is not prohibitively conservative.

As discussed in the preceding section, the more conservative threshold for the minimum eigenvalue,  $\mathcal{B}^s(\mathcal{W})/\tau = \|\Psi\|_2/\tau$ , is more effective in bounding the Monte Carlo bias in the simulations than the worst-case Nagar bias when  $K \leq N + 1$ . The first panel of Figure 2 shows that rejection rates for  $N = K = 2$  exceed 0.05 in only a handful of DGPs when Monte Carlo bias is greater or equal than 0.10. In the  $N = 2, K = 3$  model, the rejection rates based on the alternative threshold essentially never exceed 0.05 for Monte Carlo bias greater than or equal to 0.10. In the Online Appendix, we further show that when the test is based on the worst-case Nagar bias instead of the more conservative bound in models with  $K \leq N + 1$ , the Monte Carlo bias is



Figure 2: Size and Power of the First-Stage Test



*Notes:* Figure shows rejection rates across 1,000 samples for each of five million DGPs generated as explained in the main text. The red dots shows the rejection rates as a function of the worst-case Nagar bias or the alternative conservative bound on the bias. The blue dots show the rejection rates as a function of the Monte Carlo bias. The vertical full line marks the bias tolerance level  $\tau = 0.10$  in the null hypothesis, the dashed vertical line marks a bias level of 0.05 for reference, and the horizontal full line plots the nominal size  $\alpha = 0.05$ .

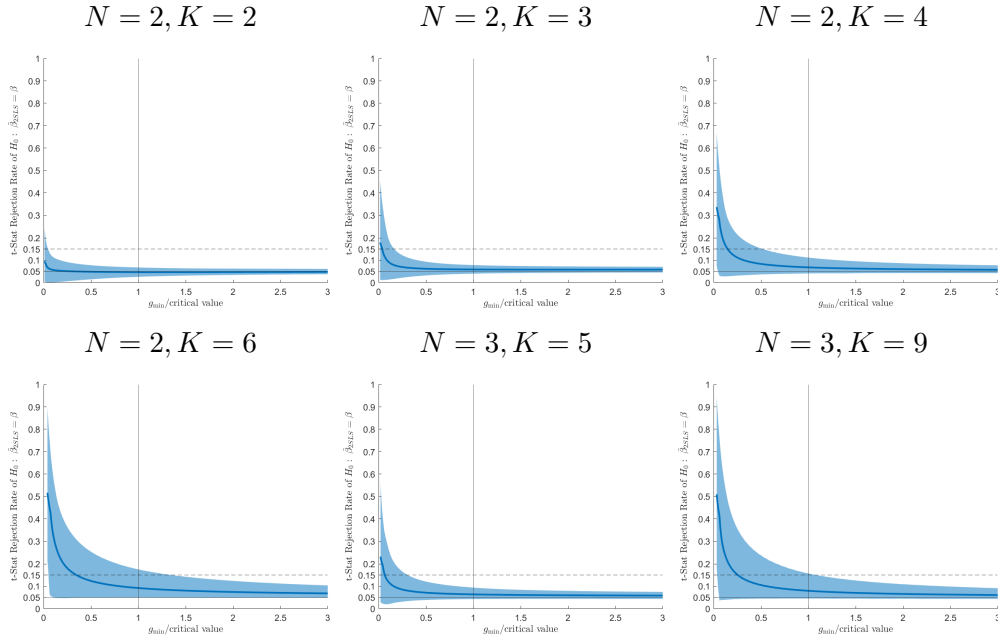
instead not well-controlled.

Overall, these simulation results show that our testing procedures perform as intended, including in models with  $K \leq N + 1$  as long as the more conservative threshold is used in those cases. Ultimately, the theoretical justification of a bias-based test is on a much weaker foundation whenever  $K \leq N + 1$ , since the first moment of the 2SLS bias generally does not exist, and the Monte Carlo bias can exceed one. This is no different than for Stock and Yogo (2005), who do not report critical values for  $K \leq N + 1$ , or Montiel Olea and Pflueger (2013), who proceed with the Nagar bias in these cases, motivated by its existence alone.

### 3.3 Size Distortions of $t$ -Statistic Inference on $\beta$

Alternative testing strategies for weak instruments can be based on controlling size distortions of Wald or  $t$ -statistic inference on  $\beta$ . The generalization of the size-based test of Stock and Yogo (2005) to heteroskedastic and serially correlated models is beyond the scope of this paper. Nevertheless, we explore the relationship between the test statistic  $g_{\min}$  and the distortions of a stan-

Figure 3: Size of  $t$ -statistic inference on  $\beta$



*Notes:* For each specification, we consider five million DGPs as described in the text. For each DGP we take 1000 samples, and for each sample we calculate the first-stage test statistic  $g_{\min}$  and conduct a two-sided  $t$ -test for each element of  $\beta$ . The figure shows the average and 95 percentiles of the  $t$ -test rejection rates as a function of the average  $g_{\min}$  for 100 equally spaced bins.

dard two-sided  $t$ -test in Figure 3. The  $t$ -tests are for the null hypothesis that a given element in  $\hat{\beta}_{2SLS}$  equals the true value. Each panel shows binned averages of the rejection rates across the  $N$   $t$ -tests in the five million DGPs as a function of the average ratio of  $g_{\min}$  to the critical value of the first-stage test. The shaded area plots the 95 percent interval of the rejection rates within each bin. The full horizontal line shows the 0.05 nominal level of the  $t$ -test. For reference, the dashed horizontal line marks the 0.15 level, corresponding to a common tolerance level of 0.10 in size-based tests of weak instruments.

Figure 3 shows that the size distortions generally grow larger as  $g_{\min}$  becomes smaller relative to the critical value. In addition, the size distortions vanish – up to Monte Carlo error – as  $g_{\min}$  grows larger. On average across the DGPs, the  $t$ -tests lead to over-rejection for low values of  $g_{\min}$  relative to the critical value. The size distortions are relatively small in the  $N = 2, K = 2$  model even when  $g_{\min}$  is well below the critical value of the bias-based test. The size distortions become more severe at low relative values of  $g_{\min}$  as the degree of overidentification increases. Overall, these patterns are qualitatively the same as those discussed in Stock and Yogo (2005) for conditionally homoskedastic and serially uncorrelated models. They indicate that size dis-

tortions are well controlled (with a tolerance of 0.10) at values of  $g_{\min}$  well below those required to control bias at  $\tau = 0.10$  when the number of instruments is small. As the number of instruments increases,  $g_{\min}$  eventually needs to exceed the threshold for the bias-based test to control size in  $t$ -statistic inference on  $\beta$ . The relationship shown in Figure 3 naturally suggests an adjustment for  $t$ -based confidence intervals based on  $g_{\min}$  similar to the one suggested recently in Lee et al. (2022) for  $N = 1, K = 1$  models based on the first-stage  $F$ -statistic. We leave the development of such a procedure for future work.

#### 4 Empirical Application

We illustrate our testing procedure in an application to Ramey and Zubairy (2018), who use instrumental variables in local projections to estimate government spending multipliers using military spending news and recursively identified government spending shocks as instruments. The authors' key innovation is to allow the spending multipliers to depend on the state of the business cycle, or alternatively on the monetary policy regime. To this end, the authors interact government spending with an indicator for whether the economy is in a period of slack, or an indicator for whether the policy rate is constrained at the zero lower bound (ZLB). Using their original notation, Ramey and Zubairy (2018) estimate cumulative multipliers for  $h = 0, 1, \dots$  based on

$$(16) \quad \sum_{j=0}^h y_{t+j} = I_{t-1} \left[ \gamma_{A,h} + \phi_{A,h}(L) z_{t-1} + m_{A,h} \sum_{j=0}^h g_{t+j} \right] \\ + (1 - I_{t-1}) \left[ \gamma_{B,h} + \phi_{B,h}(L) z_{t-1} + m_{B,h} \sum_{j=0}^h g_{t+j} \right] + \omega_{t+h},$$

where  $h$  is the horizon in quarters,  $y_t$  is detrended GDP,  $I_{t-1}$  is the regime indicator,  $z_{t-1}$  is a vector of controls,  $\phi_{A,h}(L)$ ,  $\phi_{B,h}(L)$  are polynomials in the lag operator,  $g_t$  is government spending divided by trend GDP, and  $m_{A,h}$ ,  $m_{B,h}$  are the cumulative spending multipliers over  $h$  quarters in the respective states. The use of interaction terms involving endogenous regressors is quite common, and is one example of how multiple endogenous regressors often arise in practical applications.

To assess instrument relevance, Ramey and Zubairy (2018) apply the Montiel Olea and Pflueger (2013) test to the individual subsamples implied by the regime indicators, as for each such subsample there is only a single endogenous

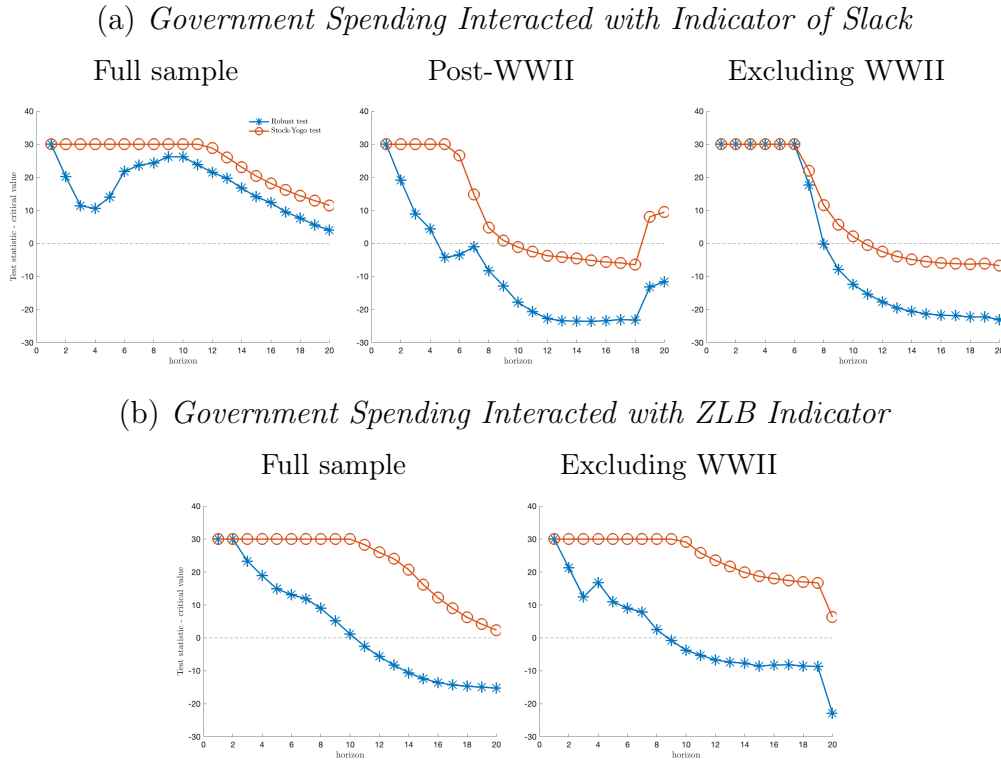
regressor. However, to assess whether the multiplier estimates are statistically different across regimes, government spending ultimately has to be interacted with the indicators in a single specification as in (16), in which there are not one but two endogenous regressors ( $N=2$ ). As autocorrelated errors are an inherent feature of local projections such as (16), Ramey and Zubairy (2018) can unfortunately not rely on the Stock and Yogo (2005) test to assess instrument strength for their regression of interest. Our robust test, in contrast, allows for a direct test of instrument relevance for the specifications in (16). We therefore implement our test as described in Section 1 for  $\tau = 0.10$  and  $\alpha = 0.05$ , and using the same Newey and West (1994) automatic bandwidth HAR estimation procedure as Ramey and Zubairy (2018). The instruments are the military news measure and the recursive spending shocks interacted with the regime indicator as instruments, such that  $K = 4$ . For reference, we replicate the results from Ramey and Zubairy (2018) for the regime subsamples in the Online Appendix.

Figure 4 reports the results for the main specifications and sample periods discussed in Ramey and Zubairy (2018). Panel (a) covers specifications that interact government spending with a measure of slack based on the unemployment rate. The starred blue line plots the difference between our test statistic  $g_{\min}$  and the critical value for  $\tau = 0.10$ ,  $\alpha = 0.05$ , truncated at 30 for readability as in Ramey and Zubairy (2018). Although not reported by Ramey and Zubairy (2018), for illustrative purposes the circled red line plots the difference between the Cragg and Donald (1993) test statistic and the Stock and Yogo (2005) critical value. Panel (b) in Figure 4 shows the corresponding results for specifications where government spending is interacted with an indicator for whether monetary policy is constrained by the ZLB. For the interested reader, we report the test statistics and critical values for our test separately in the Online Appendix.

Similar to the regime-specific results reported by Ramey and Zubairy (2018), our robust test rejects that the instruments are weak at relatively short horizons across all samples. According to our test, issues with instrument relevance start to arise in the specifications with the slack indicator after horizons between 5 to 8 quarters unless the sample includes WWII. In the specifications with the ZLB indicator, instrument relevance becomes a concern for horizons beyond 9 quarters.

Comparison with the results from the Stock and Yogo (2005) test demonstrates the importance of allowing for heteroskedasticity and autocorrelation in the first-stage testing procedures. For three of the five specifications considered in Figure 4, the Stock and Yogo (2005) test leads to a rejection of weak

Figure 4: Test Results for the Ramey and Zubairy (2018) Regression Across Horizons



*Notes:* Panel (a) reports results for specifications with government spending interacted with an indicator for whether the economy was in a state of slack, using combined instruments for different sample periods: 1890-2015, 1947-2015 (post-WWII), and 1890-2015 excluding WWII. The starred blue line plots the difference between the robust test statistic and robust critical values for  $\tau = 0.10$  and  $\alpha = 0.05$  across horizons. As in Ramey and Zubairy (2018), we cap the results at 30 for visibility. The circled red line shows the difference between the Cragg and Donald (1993) statistic and critical values from Stock and Yogo (2005). Panel (b) reports analogous results for specifications with government spending interacted with an indicator for whether monetary policy is constrained by the zero lower bound for different sample periods, 1890-2015 and 1890-2015 excluding WWII.

instruments at all horizons considered. For the remaining two specifications, the Stock and Yogo (2005) test rejects weak instruments for an additional 3 to 4 quarters compared to the robust test.

The comparison of our robust test results to the regime-specific results in Ramey and Zubairy (2018), reported in the Online Appendix, is also informative. Since the point estimates (and thus the bias) in the interacted regression are the same as those obtained from the regressions in the regime subsamples, it is not surprising that when each regime-specific regression appears to be separately strongly identified, the interacted regression generally is too.<sup>9</sup> The more interesting cases occur when one state is strongly identified, but the

<sup>9</sup>In just a few cases where one state is only marginally strongly identified, the interacted specification is weakly identified. We attribute this to the fact that the test is more conservative for  $N > 1$ .

other is weakly identified. Ex ante, it is not obvious whether the interacted model would fall above or below the relevant critical value. In this application, the more weakly identified state appears to dictate the test result. However, this need not be the case in general, depending on the covariance structure of the regressors and instruments.

For models with  $N > 1$ , researchers sometimes report results for the robust Kleibergen and Paap (2006) test of under-identification. In the Online Appendix, we report the results for this test for the Ramey and Zubairy (2018) regression. In general, the Kleibergen and Paap (2006) test statistic decreases with the horizon, as does our robust statistic for the null of weak identification (unlike in our test, the Kleibergen-Paap critical value is constant across horizons). For three of the specifications shown in Figure 4, the Kleibergen and Paap (2006) test rejects the null of under-identification at all horizons. For the remaining two, under-identification cannot be rejected after horizons of around 10 to 15 quarters. These results from a test of under-identification are entirely compatible with our findings based on a test of weak identification, since the tests consider different null hypotheses. Indeed, instruments may be weak enough to induce bias above the tolerance level, even if the null of a rank deficiency in  $\Pi$  can be rejected with high statistical significance.<sup>10</sup> In most applications, the bias of an estimator is of greater interest than knowledge of the rank of  $\Pi$ . In those cases, a bias-based test of weak instruments is therefore more directly useful.

## 5 Concluding Remarks

First-stage tests like those proposed by Stock and Yogo (2005) or, more recently, Montiel Olea and Pflueger (2013), are a widely-used diagnostic tool to assess instrument relevance in empirical applications that involve instrumental variables. When researchers are not comfortable imposing homoskedasticity assumptions for second stage inference, they should also avoid imposing such assumptions in first-stage testing procedures. In this paper, we generalize the testing approach of Montiel Olea and Pflueger (2013) to provide a first-stage test that is valid under heteroskedasticity and autocorrelation regardless of the number of endogenous regressors. The computer code accompanying this paper provides empirical researchers with an easy-to-use bias-based first-stage test under assumptions that match those imposed for second-stage inference. Future work could consider the generalization of the size-based test of Stock

---

<sup>10</sup>An alternative possibility is that the 2SLS estimator is not making optimal use of the information contained in the instruments, see e.g., Windmeijer (2022)

and Yogo (2005). The first-stage test statistic in this paper could also be the foundation for extending the  $t$ -statistic inference approach for  $K = 1, N = 1$  models of Lee et al. (2022). Finally, our generalization of the Nagar approximation to the 2SLS bias should also permit extensions to the methods in Ganics et al. (2021) to construct confidence intervals for the 2SLS bias.

## References

- Alzer, Horst (1997). “On Some Inequalities for the Gamma and Psi Functions”. In: *Mathematics of Computation* 66.217, pp. 373–389.
- Andrews, Isaiah (2018). “Valid Two-Step Identification-Robust Confidence Sets for GMM”. In: *The Review of Economics and Statistics* 100.2, pp. 337–348.
- Andrews, Isaiah, James H. Stock, and Liyang Sun (2019). “Weak Instruments in Instrumental Variables Regression: Theory and Practice”. In: *Annual Review of Economics* 11.1, pp. 727–753.
- Basman, R. L. (1961). “A Note on the Exact Finite Sample Frequency Functions of Generalized Classical Linear Estimators in Two Leading Over-Identified Cases”. In: *Journal of the American Statistical Association* 56.295, pp. 619–636.
- Bernstein, Dennis S. (2009). *Matrix Mathematics, Theory, Facts and Formulas*. Princeton: Princeton University Press.
- Bodenham, Dean and Niall Adams (2015). “A comparison of efficient approximations for a weighted sum of chi-squared random variables”. In: *Statistics and Computing* 26.
- Carrasco, Marine and Mohamed Doukali (2021). “Testing overidentifying restrictions with many instruments and heteroscedasticity using regularised jackknife IV”. In: *The Econometrics Journal* 25.1, pp. 71–97.
- Cragg, John G. and Stephen G. Donald (1993). “Testing Identifiability and Specification in Instrumental Variable Models”. In: *Econometric Theory* 9.2, pp. 222–240.
- Ganics, Gergely, Atsushi Inoue, and Barbara Rossi (2021). “Confidence Intervals for Bias and Size Distortion in IV and Local Projections-IV Models”. In: *Journal of Business & Economic Statistics* 39.1, pp. 307–324.
- Horn, Roger A. and Charles R. Johnson (2013). *Matrix Analysis*. Cambridge: Cambridge University Press.
- Imhof, J. P. (Dec. 1961). “Computing the distribution of quadratic forms in normal variables”. In: *Biometrika* 48.3-4, pp. 419–426.

- Keane, Michael and Timothy Neal (2022). *A Practical Guide to Weak Instruments*. Discussion Papers 2021-05d. School of Economics, The University of New South Wales.
- Kinal, Terrence W. (1980). “The Existence of Moments of k-Class Estimators”. In: *Econometrica* 48.1, pp. 241–249.
- Kleibergen, Frank and Richard Paap (2006). “Generalized reduced rank tests using the singular value decomposition”. In: *Journal of Econometrics* 133.1, pp. 97–126.
- Kollo, Tõnu and Dietrich von Rosen (Dec. 1995). “Approximating by the Wishart distribution”. In: *Annals of the Institute of Statistical Mathematics* 47, pp. 767–783.
- Lee, David S., Justin McCrary, Marcelo J. Moreira, and Jack Porter (2022). “Valid t-Ratio Inference for IV”. In: *American Economic Review* 112.10, pp. 3260–90.
- Magnus, Jan R. and Heinz Neudecker (2019). *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Oxford: John Wiley and Sons Ltd.
- Mariano, Roberto S. (1972). “The Existence of Moments of the Ordinary Least Squares and Two-Stage Least Squares Estimators”. In: *Econometrica* 40.4, pp. 643–652.
- Mathai, Arak (Jan. 1980). “Moments of the trace of a noncentral Wishart matrix”. In: *Communications in Statistics - Theory and Methods* 9, pp. 795–801.
- Montiel Olea, José Luis and Carolin Pflueger (2013). “A Robust Test for Weak Instruments”. In: *Journal of Business & Economic Statistics* 31.3, pp. 358–369.
- Muirhead, Robb J. (1982). *Aspects of Multivariate Statistical Theory*. Hoboken, New Jersey: John Wiley and Sons Ltd.
- Nagar, A. L. (1959). “The Bias and Moment Matrix of the General k-Class Estimators of the Parameters in Simultaneous Equations”. In: *Econometrica* 27.4, pp. 575–595.
- Newey, Whitney K. and Kenneth D. West (1994). “Automatic Lag Selection in Covariance Matrix Estimation”. In: *The Review of Economic Studies* 61.4, pp. 631–653.
- Patnaik, P.B. (1949). “The Non-Central Chi-Square and F-Distributions and Their Applications”. In: *Biometrika* 36, pp. 202–232.
- Phillips, P. C. B. (1980). “The Exact Distribution of Instrumental Variable Estimators in an Equation Containing  $n + 1$  Endogenous Variables”. In: *Econometrica* 48.4, pp. 861–878.



- Ramey, Valerie A. and Sarah Zubairy (2018). “Government Spending Multipliers in Good Times and in Bad: Evidence from US Historical Data”. In: *Journal of Political Economy* 126.2, pp. 850–901.
- Sanderson, Eleanor and Frank Windmeijer (2016). “A weak instrument F-test in linear IV models with multiple endogenous variables”. In: *Journal of Econometrics* 190.2, pp. 212–221.
- Skeels, Christopher L. and Frank Windmeijer (2018). “On the Stock–Yogo Tables”. In: *Econometrics* 6.4.
- Solomon, Herbert and Michael A. Stephens (1977). “Distribution of a Sum of Weighted Chi-Square Variables”. In: *Journal of the American Statistical Association* 72.360, pp. 881–885.
- Staiger, Douglas and James H. Stock (1997). “Instrumental Variables Regression with Weak Instruments”. In: *Econometrica* 65.3, pp. 557–586.
- Stock, James and Motohiro Yogo (2005). “Testing for Weak Instruments in Linear IV Regression”. In: *Identification and Inference for Econometric Models*. Ed. by Donald W.K. Andrews. New York: Cambridge University Press, pp. 80–108.
- Wen, Zaiwen and Wotao Yin (2013). “A Feasible Method for Optimization with Orthogonality Constraints”. In: *Mathematical Programming* 142, pp. 397–434.
- Windmeijer, Frank (2022). *Weak Instruments, First-Stage Heteroskedasticity, the Robust F-Test and a GMM Estimator with the Weight Matrix Based on First-Stage Residuals*. DOI: [10.48550/ARXIV.2208.01967](https://doi.org/10.48550/ARXIV.2208.01967).

## A Proof of Lemma 1

From Assumption 2.b it follows that  $\eta_1 = \mathbf{S}_{12}\mathbf{S}_2^{-1}(\eta_2 - c) + \epsilon$ , where  $\epsilon$  is mean zero and independent of  $\eta_2$  and  $c = \text{vec}(C')$ . Substituting into the expression for  $\beta_{2SLS}^*$  in Proposition 1 and taking expectations,

$$(A.1) \quad E[\beta_{2SLS}^*] = E \left[ (R'_{N,K}(\eta_2\eta_2' \otimes I_K)R_{N,K})^{-1} R'_{N,K} \text{vec}(\mathbf{S}_{12}\mathbf{S}_2^{-1}(\eta_2 - c)\eta_2') \right].$$

Defining  $l = \mathbf{S}_2^{-\frac{1}{2}}c$ ,  $\psi = \mathbf{S}_2^{-\frac{1}{2}}(\eta_2 - c) \sim \mathcal{N}(0, I_{NK})$  and using  $\text{vec}(UV) = (V' \otimes I) \text{vec}(U)$ , the bias in (A.1) can be rewritten as

$$(A.2) \quad E[\beta_{2SLS}^*] = E \left[ \left( R'_{N,K}(\mathbf{S}_2^{\frac{1}{2}}(l + \psi)(l + \psi)'(\mathbf{S}_2^{\frac{1}{2}})' \otimes I_K)R_{N,K} \right)^{-1} \right. \\ \left. \times R'_{N,K} \left( \mathbf{S}_2^{\frac{1}{2}}(l + \psi)\psi'\mathbf{S}_2^{-\frac{1}{2}} \otimes I_K \right) \right] \text{vec}(\mathbf{S}_{12}).$$

Using the definitions  $\Phi = R'_{N,K}(\mathbf{S}_2 \otimes I_K)R_{N,K}$ ,  $\mathcal{S} = ((\Phi/K)^{-\frac{1}{2}} \otimes I_K)\mathbf{S}_2^{\frac{1}{2}}$ , and  $\rho = ((R'_{N,K}(\mathbf{S}_2 \otimes I_K)R_{N,K})^{-\frac{1}{2}} \otimes I_{K^2}) \text{vec}(\mathbf{S}_{12}) / \sqrt{\text{Tr}(\mathbf{S}_1)}$ .<sup>1</sup> The unweighted bias becomes

$$(A.3) \quad E[\beta_{2SLS}^*] = K\Phi^{-\frac{1}{2}}E \left[ (R'_{N,K}(\mathcal{S}(l + \psi)(l + \psi)'\mathcal{S}' \otimes I_K)R_{N,K})^{-1} \right. \\ \left. \times R'_{N,K}(\mathcal{S}(l + \psi)\psi'\mathcal{S}^{-1} \otimes I_K) \right] \rho \sqrt{\text{Tr}(\mathbf{S}_1)}.$$

Using Definition 2,

$$(A.4) \quad B^2 = \frac{E[\beta_{2SLS}^*]'\mathbf{R}'_{N,K}(\mathbf{S}_2 \otimes I_K)R_{N,K}E[\beta_{2SLS}^*]}{\text{Tr}(\mathbf{S}_1)} = \rho'\mathbf{h}'\mathbf{h}\rho,$$

where  $\mathbf{h} = KE \left[ (R'_{N,K}(\mathcal{S}(l + \psi)(l + \psi)'\mathcal{S}' \otimes I_K)R_{N,K})^{-1} R'_{N,K}(\mathcal{S}(l + \psi)\psi'\mathcal{S}^{-1} \otimes I_K) \right]$ . Therefore,  $B = \|\mathbf{h}\rho\|_2$ .

## B Proof of Theorem 1

Define the function  $h : \mathbb{R}^{NK} \mapsto \mathbb{R}^{N \times NK^2}$

$$(B.1) \quad h(\psi) = (R'_{N,K}(\mathcal{S}(l + \psi)(l + \psi)'\mathcal{S}' \otimes I_K)R_{N,K})^{-1} \\ \times R'_{N,K}(\mathcal{S}(l + \psi)\psi'\mathcal{S}^{-1} \otimes I_K).$$

The Nagar approximation of  $\mathbf{h}$ , denoted by  $\mathbf{h}_n$ , is the expectation of the second-order Taylor expansion of  $Kh(\psi)$  evaluated at  $\psi = 0$ , and is given in

<sup>1</sup>The matrix  $\mathcal{S}$  is non-symmetric in general, and  $\text{Tr}(\mathcal{S}) = \text{Tr}(\mathcal{S}\mathcal{S}') = NK$ .

vectorized form by

(B.2)

$$\text{vec}(\mathbf{h}_n) = \frac{K}{2} (\mathbf{1}'_{NK} \otimes I_{(NK)^2}) \left( \sum_{j=1}^{KN} \left( e_j^{NK} (e_j^{NK})' \otimes I_{(NK)^2} \right) (\nabla_2 h(0))_j \right),$$

where  $\mathbf{1}_m$  is the  $m \times 1$  vector of ones,  $e_j^m$  the  $m \times 1$  vector with  $j$ -th element equal to one and zeros in all other rows, and  $\nabla_2 h(\psi)$  is the  $(NK)^3 \times NK$  second matrix derivative of  $h$ . Using the matrix differentiation rules in, for instance, Magnus and Neudecker (2019),  $\nabla_2 h(\psi)$  is given by

$$(B.3) \quad \begin{aligned} \nabla_2 h(\psi) = & (I_{NK} \otimes A_3(\psi)) \nabla A_1(\psi) + (A_1(\psi)' \otimes I_{(NK)^2}) \nabla A_3(\psi) \\ & + (I_{NK} \otimes A_4(\psi)) \nabla A_2(\psi) + (A_2(\psi)' \otimes I_{(NK)^2}) \nabla A_4(\psi), \end{aligned}$$

where

$$\begin{aligned} A_1(\psi) &= -(C_1(\psi) \otimes C_1(\psi)) B_1 ((\mathcal{S} \otimes \mathcal{S}(l + \psi)) + (\mathcal{S}(l + \psi) \otimes \mathcal{S})) \\ A_2(\psi) &= B_2 (((\mathcal{S}^{-1})' \otimes \mathcal{S}(l + \psi)) + ((\mathcal{S}^{-1})' \psi \otimes \mathcal{S})) \\ A_3(\psi) &= (\mathcal{S}(l + \psi) \psi' \mathcal{S}^{-1} \otimes I_K)' R_{N,K} \otimes I_N \\ A_4(\psi) &= I_{NK^2} \otimes C_1(\psi) \\ \nabla A_1(\psi) &= (((\mathcal{S} \otimes \mathcal{S}(l + \psi)) + (\mathcal{S}(l + \psi) \otimes \mathcal{S}))' B_1' \otimes I_{N^2}) C_2(\psi) - (I_{NK} \otimes (C_1(\psi) \otimes C_1(\psi))) B_3 \\ \nabla A_2(\psi) &= (I_{NK} \otimes B_2) ((\text{vec}((\mathcal{S}^{-1})') \otimes I_{NK}) \mathcal{S} + (\mathcal{K}_{NK,NK} \otimes I_{NK}) (I_{NK} \otimes \text{vec}(\mathcal{S})) (\mathcal{S}^{-1})') \\ \nabla A_3(\psi) &= (I_N \otimes \mathcal{K}_{N,NK^2} \otimes I_N) (I_{(NK)^2} \otimes \text{vec}(I_N)) \mathcal{K}_{N,NK^2} A_2(\psi) \\ \nabla A_4(\psi) &= (I_{NK^2} \otimes \mathcal{K}_{N,NK^2} \otimes I_N) (\text{vec}(I_{NK^2}) \otimes I_{N^2}) A_1(\psi) \\ B_1 &= (R'_{N,K} \otimes R'_{N,K}) (I_{NK} \otimes \mathcal{K}_{K,NK} \otimes I_K) (I_{(NK)^2} \otimes \text{vec}(I_K)) \\ B_2 &= (I_{NK^2} \otimes R'_{N,K}) (I_{NK} \otimes \mathcal{K}_{K,NK} \otimes I_K) (I_{(NK)^2} \otimes \text{vec}(I_K)) \\ B_3 &= (I_{NK} \otimes B_1) (\text{vec}(\mathcal{S} \otimes I_{NK}) \mathcal{S} + (\mathcal{K}_{NK,NK} \otimes I_{NK}) (I_{NK} \otimes \text{vec}(\mathcal{S})) \mathcal{S}) \\ C_1(\psi) &= (R'_{N,K} (\mathcal{S}(l + \psi) (l + \psi)' \mathcal{S}' \otimes I_K) R_{N,K})^{-1} \\ C_2(\psi) &= (I_N \otimes \mathcal{K}_{N,N} \otimes I_N) [\text{vec}(A_0(\psi)) \otimes I_{N^2} : I_{N^2} \otimes \text{vec}(A_0(\psi))] [A_1(\psi)' : A_1(\psi)']'. \end{aligned}$$

Writing (B.2) in matrix form and simplifying yields

(B.4)

$$\mathbf{h}_n = \Lambda^{-1} \left( R'_{N,K} - (\text{vec}(\Lambda^{-1}) \otimes I_N)' (I_N \otimes \mathcal{K}_{N,N}) ((I_{N^2} + \mathcal{K}_{N,N}) (I_N \otimes L) \otimes L) \right),$$

where  $L = K^{-\frac{1}{2}} R'_{N,K} (\mathcal{S} l \otimes I_K)$  and  $\Lambda$  is the concentration matrix in Definition 1. To proceed, we reparametrize the functional dependence of the bias on  $l$  through

$$(B.5) \quad l = \mathcal{S}^{-1} \sqrt{K} \text{vec}(L'_0 \mathcal{D}_\Lambda^{\frac{1}{2}} Q'_\Lambda),$$

where  $Q_\Lambda \in \mathbb{O}^{N \times N}$ ,  $\mathcal{D}_\Lambda \in \mathbb{R}^{N \times N}$  contain the eigenstructure of the concentration matrix  $\Lambda$ , and  $L_0 \in \mathbb{O}^{N \times K}$  is an orthogonal matrix. By definition,  $\Lambda = Q_\Lambda \mathcal{D}_\Lambda Q'_\Lambda$ , where  $\mathcal{D}_\Lambda$  is a diagonal matrix containing the eigenvalues  $\lambda_i > 0$ ,  $i = 1, \dots, N$ , and  $Q_\Lambda Q'_\Lambda = I_N$ . The reparametrization in (B.5) reformulates the choice of the  $NK$  parameters of  $l$  as an equivalent choice of the  $N$  free parameters in  $\mathcal{D}_\Lambda$ , the  $N^2 - (N+1)N/2$  free parameters of  $Q_\Lambda$ , and the  $NK - (N+1)N/2$  free parameters of  $L_0$ .

Using the eigenvalue decomposition  $\Lambda = Q_\Lambda \mathcal{D}_\Lambda Q'_\Lambda$ , and the fact that (B.5) implies  $L_0 = Q'_\Lambda \Lambda^{-\frac{1}{2}} L$ , (B.4) can be rewritten as

$$(B.6) \quad \mathbf{h}_n = Q_\Lambda \mathcal{D}_\Lambda^{-\frac{1}{2}} M_1 (\mathcal{D}_\Lambda^{-\frac{1}{2}} Q_\Lambda \otimes L_0 \otimes L_0) M_2,$$

where  $M_1 = R'_{N,N} (I_{N^3} + (\mathcal{K}_{N,N} \otimes I_N))$  and  $M_2 = R_{N,K} R'_{N,K} / (N+1) - I_{NK^2}$ , which provides the Nagar approximation for the expectation in the general bias,  $\mathbf{h}_n$ , with

$$(B.7) \quad B_n(\beta, Q_\Lambda, \mathcal{D}_\Lambda, L_0, \mathbf{W}) = \|\mathbf{h}_n \rho\|_2.$$

Let  $B_n^*(\mathbf{W}, \lambda_{\min}) = \sup_{\beta, Q_\Lambda, \mathcal{D}_\Lambda, L_0} \{B_n(\beta, Q_\Lambda, \mathcal{D}_\Lambda, L_0, \mathbf{W})\}$  denote the sharp upper bound on the Nagar bias over  $\beta \in \mathbb{R}^N$ ,  $Q_\Lambda \in \mathbb{O}^{N \times N}$ ,  $L_0 \in \mathbb{O}^{N \times K}$  and  $\mathcal{D}_\Lambda$  in the set of all diagonal matrices with no diagonal element smaller than  $\lambda_{\min}$ , the smallest eigenvalue of the concentration matrix.

Using the definitions of  $\mathbf{S}_1$ ,  $\mathbf{S}_2$  and  $\mathbf{S}_{12}$  in (7),

$$(B.8) \quad \begin{aligned} \text{Tr}(\mathbf{S}_1) &= \text{Tr}((\tilde{\beta}' \otimes I_K) \mathbf{W} (\tilde{\beta} \otimes I_K)) = \tilde{\beta}' R'_{N+1,K} (\mathbf{W} \otimes I_K) R_{N+1,K} \tilde{\beta}, \\ \text{vec}(\mathbf{S}_{12})' &= \text{vec}((\tilde{\beta}' \otimes I_K) [\mathbf{W}_{12} : \mathbf{W}_2])' \\ &= \text{vec}(\tilde{\beta}' R'_{N+1,K} ([\mathbf{W}_{12} : \mathbf{W}_2] \otimes I_K))', \end{aligned}$$

where  $\tilde{\beta} = [1 : -\beta']'$ . Substituting into the definition of  $\rho$  in Lemma 1 yields

$$(B.9) \quad \rho = K^{-\frac{1}{2}} \Psi x / \sqrt{x'x},$$

where  $x = (R'_{N+1,K} (\mathbf{W} \otimes I_K) R_{N+1,K})^{\frac{1}{2}} \tilde{\beta}$  and  $\Psi = (\mathbf{S} \mathbf{W}_2^{-\frac{1}{2}} [\mathbf{W}_{12} : \mathbf{W}_2]' \otimes I_K) R_{N+1,K} (R'_{N+1,K} (\mathbf{W} \otimes I_K) R_{N+1,K})^{-\frac{1}{2}}$ . Since

$$(B.10) \quad \sup_{\beta \in \mathbb{R}^N} \|\mathbf{h}_n \rho\|_2 = K^{-\frac{1}{2}} \sup_{x \in \mathbb{R}^{N+1}} \frac{\|\mathbf{h}_n \Psi x\|_2}{\|x\|_2} = K^{-\frac{1}{2}} \|\mathbf{h}_n \Psi\|_2,$$

the optimization of the Nagar bias over  $\beta$  amounts to the taking the largest singular value of the matrix  $\mathbf{h}_n \Psi$ .

Next, note that  $\mathbf{h}_n \mathbf{h}'_n = Q_\Lambda \mathcal{D}_h Q'_\Lambda$  where  $\mathcal{D}_h = (K - 2(1 + N))\mathcal{D}_\Lambda^{-2} + \mathcal{D}_\Lambda^{-\frac{1}{2}} M_1 (\mathcal{D}_\Lambda^{-1} \otimes I_{N^2}) M'_1 \mathcal{D}_\Lambda^{-\frac{1}{2}}$  is a diagonal matrix, such that  $Q_\Lambda$  are eigenvectors of  $\mathbf{h}_n \mathbf{h}'_n$  and  $\mathcal{D}_h$  contains the eigenvalues. The  $i$ -th diagonal element of  $\mathcal{D}_h$  is

$$(B.11) \quad \frac{1}{\lambda_i^2} \left( K - N + 1 + \sum_{j \neq i}^N \frac{\lambda_i}{\lambda_j} \right) > 0,$$

where  $\lambda_i > 0$  is the  $i$ -th eigenvalue of  $\Lambda$ . Each eigenvalue of  $\mathbf{h}_n \mathbf{h}'_n$  is decreasing in all eigenvalues of  $\Lambda$ . Making the dependence on  $\mathcal{D}_\Lambda$  explicit by the notation  $\mathbf{h}_n(\mathcal{D}_\Lambda)$  and fixing  $Q_\Lambda$ , it is therefore the case that  $\lambda_{\min}^{-2} \mathbf{h}_n(I_N) \mathbf{h}'_n(I_N) - \mathbf{h}_n(\mathcal{D}_\Lambda) \mathbf{h}'_n(\mathcal{D}_\Lambda)$  is positive semidefinite for all  $\mathcal{D}_\Lambda$  with  $\lambda_{\min}$  as the smallest diagonal element. This in turn implies that  $\lambda_{\min}^{-2} \Psi' \mathbf{h}_n(I_N) \mathbf{h}'_n(I_N) \Psi - \Psi' \mathbf{h}_n(\mathcal{D}_\Lambda) \mathbf{h}'_n(\mathcal{D}_\Lambda) \Psi$  is positive semidefinite, see Proposition 8.1.2 in Bernstein (2009). It follows from Weyl's inequality that  $\lambda_{\min}^{-1} \|\mathbf{h}_n(I_N) \Psi\|_2 \geq \|\mathbf{h}_n(\mathcal{D}_\Lambda) \Psi\|_2$ , see for example Theorem 8.4.9 in Bernstein (2009). Therefore,

$$(B.12) \quad \sup_{\mathcal{D}_\Lambda: \lambda_i \geq \lambda_{\min}} K^{-\frac{1}{2}} \|\mathbf{h}_n(\mathcal{D}_\Lambda) \Psi\|_2 = K^{-\frac{1}{2}} \lambda_{\min}^{-1} \|Q_\Lambda M_1 (Q_\Lambda \otimes L_0 \otimes L_0) M_2 \Psi\|_2,$$

which states that largest bias occurs when all eigenvalues of the concentration matrix are equal to the smallest eigenvalue, and therefore when  $\mathcal{D}_\Lambda = \lambda_{\min} I_N$ . Finally,  $\sup_{L_0 \in \mathbb{O}^{N \times K}} \{\|Q_\Lambda M_1 (Q_\Lambda \otimes L_0 \otimes L_0) M_2 \Psi\|_2\} = \sup_{L_0 \in \mathbb{O}^{N \times K}} \{\|M_1 (I_N \otimes L_0 \otimes L_0) M_2 \Psi\|_2\}$  for any  $Q_\Lambda$ , and therefore the sharp upper bound for the bias does not depend on  $Q_\Lambda$ . This means the sharp upper bound is

$$(B.13) \quad B_n^*(\mathbf{W}, \lambda_{\min}) = \lambda_{\min}^{-1} K^{-\frac{1}{2}} \sup_{L_0 \in \mathbb{O}^{N \times K}} \{\|M_1 (I_N \otimes L_0 \otimes L_0) M_2 \Psi\|_2\},$$

which concludes the proof of part (i) of the theorem.

Turning to part (ii), the upper bound  $B_n^*(\mathbf{W}, \lambda_{\min}) \leq \lambda_{\min}^{-1} (2(N+1)/K)^{\frac{1}{2}} \|M_2 \Psi\|_2$  follows from  $K^{-\frac{1}{2}} \|\mathbf{h}_n(\lambda_{\min} I_N) \Psi\|_2 \leq K^{-\frac{1}{2}} \lambda_{\min}^{-1} \|M_1\|_2 \|(I_N \otimes L_0 \otimes L_0)\|_2 \|M_2 \Psi\|_2$  and the fact that  $\|M_1\|_2 = (2(N+1))^{\frac{1}{2}}$  and  $\|(I_N \otimes L_0 \otimes L_0)\|_2 = 1$ . The inequality follows from Proposition 9.6.1 in Bernstein (2009). Finally, the upper bound,  $B_n^*(\mathbf{W}, \lambda_{\min}) \leq \lambda_{\min}^{-1} \|\Psi\|_2$ , follows from  $K^{-\frac{1}{2}} \|\mathbf{h}_n(\lambda_{\min} I_N) \Psi\|_2 \leq K^{-\frac{1}{2}} \|\mathbf{h}_n(\lambda_{\min} I_N)\|_2 \|\Psi\|_2 = \lambda_{\min}^{-1} \|\Psi\|_2$  since  $\|\mathbf{h}_n(\lambda_{\min} I_N)\|_2 = K^{\frac{1}{2}} \lambda_{\min}^{-1}$ , see (B.11).

## C Proof of Theorem 2.

The Laplace transform of the trace of a noncentral Wishart distribution is given in Mathai (1980), equation (1.6), from which it follows that the cumulant

generating function is

$$(C.14) \quad K_{\text{Tr}(\zeta)}(t) = -\frac{1}{2} \text{Tr}(\Omega) - \frac{K}{2} \ln |I_{NK} - 2\Sigma| + \frac{1}{2} \text{Tr}((I_{NK} - 2\Sigma)^{-1}\Omega).$$

We follow e.g., Muirhead (1982) and Kollo and Rosen (1995) in evaluating the cumulant generating function for a submatrix,

$$(C.15) \quad K_{\text{Tr}(\zeta)}(T_N) = -\frac{1}{2} \text{Tr}(\Omega) - \frac{K}{2} \ln |I_{NK} - 2M(T_N)\Sigma| + \frac{1}{2} \text{Tr}((I_{NK} - 2M(T_N)\Sigma)^{-1}\Omega),$$

where  $T_N$  is a  $N \times N$  matrix and

$$(C.16) \quad M(T_N) = \sum_{i,j=1,\dots,N} t_{ij} M_{ij}, \quad M_{ij} = e_j e_i'$$

where  $e_i$  is the  $i$ -th block of  $K$  columns of the matrix  $I_{NK}$ , so that  $M_{ij}\zeta$  is the matrix containing the  $i$ -th block of  $K$  rows of  $\zeta$  in its  $j$ -th block of  $K$  rows, and zero otherwise. Indexing each selection matrix  $M_{ij}$  to a scalar value  $t_{ij}$  yields the cumulant generating function of the trace of  $M_{ij}\zeta$ , analogously to Mathai (1980), when the remainder of  $T_N$  is set to zero. Note that  $\text{Tr}(M_{ij}\zeta) = \text{Tr}(\zeta_{ij})$ , since the  $j$ -th diagonal block of  $M_{ij}W$  is  $\zeta_{ij}$ , and all other diagonal blocks are zero.

The  $ij$  entry of  $R'_{N,K}(\zeta \otimes I_K)R_{N,K}$  corresponds to the trace of the  $ij$   $K \times K$  block of  $\zeta$ . Thus, the  $n$ -th cumulants of  $R'_{N,K}(\zeta \otimes I_K)R_{N,K}$  are obtained by taking the coefficients on  $\frac{T_N^n}{n!}$  in the Taylor expansion of  $K_{\text{Tr}(\zeta)}(T_N)$ , evaluated at  $T_N = 0$ . Let  $\iota(i)$  denote the index of some  $K \times K$  block of a  $NK \times NK$  matrix. Then  $\kappa_n(\zeta)^{\iota(1),\iota(2),\dots}$  denotes the  $n$ -th cumulant of  $\text{Tr}(\zeta_{\iota(1)})$  with  $\text{Tr}(\zeta_{\iota(2)}) \dots$  (i.e. the covariance for  $n = 2$ , etc.). Taking such partial

derivatives yields the following expressions:

(C.17)

$$\kappa_1^{\iota(1)}(\zeta) = \frac{K}{2} \text{Tr}(2M_{\iota(1)}\Sigma) + \frac{1}{2} \text{Tr}(2M_{\iota(1)}\Sigma\Omega)$$

(C.18)

$$\kappa_2^{\iota(1),\iota(2)}(\zeta) = \frac{K}{2} \text{Tr}(2^2 M_{\iota(2)}\Sigma M_{\iota(1)}\Sigma) + \frac{1}{2} \sum_{p \in \mathcal{P}(\iota(1),\iota(2))} \text{Tr}(2^2 M_{p(1)}\Sigma M_{p(2)}\Sigma\Omega)$$

(C.19)

$$\begin{aligned} \kappa_3^{\iota(1),\iota(2),\iota(3)}(\zeta) &= \frac{K}{2} \sum_{p \in \mathcal{P}(\iota(2),\iota(3))} \text{Tr}(2^3 M_{p(1)}\Sigma M_{p(2)}\Sigma M_{\iota(1)}\Sigma) \\ &\quad + \frac{1}{2} \sum_{p \in \mathcal{P}(\iota(1),\iota(2),\iota(3))} \text{Tr}(2^3 M_{p(1)}\Sigma M_{p(2)}\Sigma M_{p(3)}\Sigma\Omega) \end{aligned}$$

(C.20)

$$\begin{aligned} \kappa_n^{\iota(1),\iota(2),\dots}(\zeta) &= 2^{n-1} \left( K \sum_{p \in \mathcal{P}(\iota(2),\iota(3),\dots)} \text{Tr}(M_{p(1)}\Sigma M_{p(2)}\Sigma \dots M_{p(n-1)}\Sigma M_{\iota(1)}\Sigma) \right. \\ &\quad \left. + \sum_{p \in \mathcal{P}(\iota(1),\iota(2),\iota(3),\dots)} \text{Tr}(M_{p(1)}\Sigma M_{p(2)}\Sigma M_{p(3)}\Sigma \dots M_{p(n)}\Sigma\Omega) \right), \end{aligned}$$

where  $\mathcal{P}(\cdot)$  denotes the set of all permutations of the indices in the argument and  $p(i)$  denotes the  $i$ -th index in a given permutation. Note that for  $N = 1$ , the formulas collapse to those for the trace in Mathai (1980).

We next prove a lemma relating  $\text{Tr}(M_{\iota(1)}\Sigma M_{\iota(2)}\Sigma M_{\iota(3)}\Sigma \dots M_{\iota(n)}\Sigma)$  to  $\text{Tr}(\Sigma_{\iota(1)}\Sigma_{\iota(2)}\Sigma_{\iota(3)} \dots \Sigma_{\iota(n)})$ . For this purpose, denote the row block index of  $\iota(i)$  as  $\iota(i)_1$  and the column block index as  $\iota(i)_2$ , so  $M_{\iota(i)} = M_{\iota(i)_1, \iota(i)_2}$ . Additionally, let  $U_i$  denote the  $i$ -th block of  $K$  rows of the matrix  $U$ , and similarly  $U_i$  for the block of columns.

**Lemma 2.**  $\text{Tr}(M_{\iota(1)}\Sigma M_{\iota(2)}\Sigma \dots M_{\iota(n)}\Sigma) = \text{Tr}(\Sigma_{\iota(1)_1, \iota(2)_2} \Sigma_{\iota(2)_1, \iota(3)_2} \dots \Sigma_{\iota(n)_1, \iota(1)_2})$ .

*Proof.* For a general symmetric matrix  $U$ , the matrix  $M_{\iota(1)}U$  has the  $\iota(1)_1$  block of  $K$  rows of  $U$  transferred to its  $\iota(1)_2$  block of rows, with the remainder zeros. The product of a sequence of such matrices,  $M_{\iota(1)}U \dots M_{\iota(n-1)}U$  contains zeros except for the  $\iota(1)_2$  block of  $K$  rows. Suppose that this block of rows can be written as  $U_{\iota(1)_1, \iota(2)_2} U_{\iota(2)_1, \iota(3)_2} \dots U_{\iota(n-1)_1}$ . Then the product of  $M_{\iota(1)}U \dots M_{\iota(n-1)}U M_{\iota(n)}U$  is also a matrix of zeros except for the  $\iota(1)_2$  block of  $K$  rows, which are equal to  $U_{\iota(1)_1, \iota(2)_2} U_{\iota(2)_1, \iota(3)_2} \dots U_{\iota(n-1)_1, \iota(n)_2} U_{\iota(n)_1}$ . Consider as a base case  $M_{\iota(1)}U M_{\iota(2)}U$ . The product is zeros except for the  $\iota(1)_2$  block of  $K$  rows, which is equal to  $U_{\iota(1)_1, \iota(2)_2} U_{\iota(2)_1}$ . Thus, by induction, it follows that

the product of  $M_{\iota(1)}U \dots UM_{\iota(n)}U$  is a matrix of zeros, except for the  $\iota(1)_2$  block of  $K$  rows, which are equal to  $U_{\iota(1)_1, \iota(2)_2}U_{\iota(2)_1, \iota(3)_2} \dots U_{\iota(n)_1, \cdot}$ , for all  $n$ . It follows immediately that  $\text{Tr}(M_{\iota(1)}U \dots UM_{\iota(n)}U) = \text{Tr}(U_{\iota(1)_1, \iota(2)_2}U_{\iota(2)_1, \iota(3)_2} \dots U_{\iota(n)_1, \iota(1)_2})$ , since the latter argument is the single non-zero diagonal block. Letting  $U = \Sigma$  and applying the preceding result yields the stated lemma.  $\square$

Applying Lemma 2 to the cumulants above yields

(C.21)

$$\kappa_1^{\iota(1)}(\zeta) = \frac{K}{2} \text{Tr}(2\Sigma_{\iota(1)}) + \frac{1}{2} \text{Tr}(2\Sigma_{\iota(1)}\Omega)$$

(C.22)

$$\kappa_2^{\iota(1), \iota(2)}(\zeta) = \frac{K}{2} \text{Tr}(2^2\Sigma_{\iota(2)_1, \iota(1)_2}\Sigma_{\iota(1)_1, \iota(2)_2}) + \frac{1}{2} \sum_{p \in \mathcal{P}(\iota(1), \iota(2))} \text{Tr}(2^2\Sigma_{p(1)_1, p(2)_2}\Sigma_{p(2)_1, p(1)_2}\Omega)$$

(C.23)

$$\begin{aligned} \kappa_3^{\iota(1), \iota(2), \iota(3)}(\zeta) &= \frac{K}{2} \sum_{\iota \in \mathcal{P}(\iota(2), \iota(3))} \text{Tr}(2^3\Sigma_{p(1)_1, p(2)_2}\Sigma_{p(2)_1, \iota(1)_2}\Sigma_{\iota(1)_1, p(1)_2}) \\ &\quad + \frac{1}{2} \sum_{\iota \in \mathcal{P}(\iota(1), \iota(2), \iota(3))} \text{Tr}(2^3\Sigma_{p(1)_1, p(2)_2}\Sigma_{p(2)_1, p(3)_2}\Sigma_{p(3)_1, p(1)_2}\Omega) \end{aligned}$$

(C.24)

$$\begin{aligned} \kappa_n^{\iota(1), \iota(2), \dots}(\zeta) &= 2^{n-1} \left( K \sum_{p \in \mathcal{P}(\iota(2), \iota(3), \dots)} \text{Tr}(\Sigma_{p(1)_1, p(2)_2}\Sigma_{p(2)_1, p(3)_2} \dots \Sigma_{p(n-1)_1, \iota(1)_2}\Sigma_{\iota(1)_1, p(1)_2}) \right. \\ &\quad \left. + \sum_{p \in \mathcal{P}(\iota(1), \iota(2), \dots)} \text{Tr}(\Sigma_{p(1)_1, p(2)_2}\Sigma_{p(2)_1, p(3)_2} \dots \Sigma_{p(n)_1, p(1)_2}\Omega) \right). \end{aligned}$$

We ultimately need the cumulants of  $\gamma'R'_{N,K}(\zeta \otimes I_K)R_{N,K}\gamma$ . Using the preceding expressions for cumulants of  $R'_{N,K}(\zeta \otimes I_K)R_{N,K}$ , we can compute the cumulants of such quadratic forms. By homogeneity, the  $n$ -th cumulant of the quadratic form  $\gamma'U\gamma$  for a random matrix  $U$  is given by

(C.25)

$$\kappa_n(\gamma'U\gamma) = \sum_{\iota(1)_1=1}^N \sum_{\iota(1)_2=1}^N \dots \sum_{\iota(n)_1=1}^N \sum_{\iota(n)_2=1}^N \left( \prod_{j=1}^n \gamma_{\iota(j)_1} \gamma_{\iota(j)_2} \right) \kappa_n^{\iota(1), \dots, \iota(n)}(U),$$

where  $\iota(i)$  denote indices of individual elements of  $U$ . Given the previously derived expressions for the cumulants of the entries of  $R'_{N,K}(\zeta \otimes I_K)R_{N,K}$ , we can now compute the cumulants of  $\gamma'R'_{N,K}(\zeta \otimes I_K)R_{N,K}\gamma$ , noting that the cumulants for the  $ij$  entry are equal to those for the trace of the  $ij$   $K \times K$  block of  $\zeta$ ,  $\text{Tr}(\zeta_{ij})$ .



Plugging in the first summation in  $\kappa_n^{\iota(1), \dots, \iota(n)}(\zeta)$ , equation (C.24),

$$\begin{aligned}
& \text{(C.26)} \\
& \sum_{\iota(1)_1=1}^N \sum_{\iota(1)_2=1}^N \cdots \sum_{\iota(n)_1=1}^N \sum_{\iota(n)_2=1}^N \left( \prod_{j=1}^n \gamma_{\iota(j)_1} \gamma_{\iota(j)_2} \right) \sum_{p \in \mathcal{P}(\iota(2), \iota(3), \dots)} \text{Tr} \left( \Sigma_{p(1)_1, p(2)_2} \Sigma_{p(2)_1, p(3)_2} \cdots \right. \\
& \quad \left. \Sigma_{p(n-1)_1, \iota(1)_2} \Sigma_{\iota(1)_1, p(1)_2} \right) \\
& = \sum_{p \in \mathcal{P}(\iota(2), \iota(3), \dots)} \sum_{\iota'(1)_1=1}^N \sum_{\iota'(1)_2=1}^N \cdots \sum_{\iota'(n)_1=1}^N \sum_{\iota'(n)_2=1}^N \left( \prod_{j=1}^n \gamma_{\iota'(j)_1} \gamma_{\iota'(j)_2} \right) \text{Tr}(\Sigma_{\iota'(1)} \cdots \Sigma_{\iota'(n)}) \\
& = (n-1)! \sum_{\iota'(1)_1=1}^N \sum_{\iota'(1)_2=1}^N \cdots \sum_{\iota'(n)_1=1}^N \sum_{\iota'(n)_2=1}^N \left( \prod_{j=1}^n \gamma_{\iota'(j)_1} \gamma_{\iota'(j)_2} \right) \text{Tr}(\Sigma_{\iota'(1)} \cdots \Sigma_{\iota'(n)}),
\end{aligned}$$

where we used a change of indices to move from the first line to the second (recognizing that each set of permuted indices on the blocks of  $\Sigma$  is just the index for some other block of  $\Sigma$  indexed by  $\iota'(i)$ ) and in moving to the third observed that the summand of the outer summation does not depend on the indices of that summation. By the definition of matrix multiplication and considerable algebra,

$$\text{Tr}(((\gamma\gamma' \otimes I_K)U)^n) = \sum_{\iota(1)_1=1}^N \sum_{\iota(1)_2=1}^N \cdots \sum_{\iota(n)_1=1}^N \sum_{\iota(n)_2=1}^N \left( \prod_{j=1}^n \gamma_{\iota(j)_1} \gamma_{\iota(j)_2} \right) \text{Tr}(U_{\iota(1)} \cdots U_{\iota(n)}).$$

Thus, the expression further simplifies to

$$\text{(C.27)} \quad (n-1)! \text{Tr}(((\gamma\gamma' \otimes I_K)\Sigma)^n).$$

Next, we can apply the same steps to the second summation in the cumulants to obtain

$$\text{(C.28)} \quad n! \text{Tr}(((\gamma\gamma' \otimes I_K)\Sigma)^n \Omega).$$

Combining both terms yields the expression in part (i) of the theorem.

Turning to part (ii), using the fact that for a positive semi-definite matrix  $V$ ,  $|\text{Tr}(UV)| \leq \text{maxeval } U \text{Tr}(V)$ , see Fact 8.12.29 in Bernstein (2009), and the fact that  $\text{Tr}(((\gamma\gamma' \otimes I_K)\Sigma)^n \Omega) \geq 0$ , we have

$$\begin{aligned}
\text{(C.29)} \quad \text{Tr}(((\gamma\gamma' \otimes I_K)\Sigma)^n \Omega) & \leq \text{maxeval}\{((\gamma\gamma' \otimes I_K)\Sigma)^{n-1}\} \text{Tr}((\gamma\gamma' \otimes I_K)\Sigma \Omega) \\
& = K \lambda_{\min} \text{maxeval}\{((\gamma\gamma' \otimes I_K)\Sigma)^{n-1}\},
\end{aligned}$$

where the last step follows from  $\text{Tr}((\gamma\gamma' \otimes I_K)\Sigma\Omega) = \gamma'R'(\Sigma\Omega \otimes I_K)R\gamma = K\lambda_{\min}$ . Next note that

(C.30)

$$\begin{aligned} \text{maxeval}\{((\gamma\gamma' \otimes I_K)\Sigma)^{n-1}\} &= (\text{maxeval}\{(\gamma\gamma' \otimes I_K)\Sigma\})^{n-1} \\ &= \left(\text{maxeval}\{\Sigma^{\frac{1}{2}}(\gamma\gamma' \otimes I_K)\Sigma^{\frac{1}{2}}\}\right)^{n-1} \\ &\leq (\text{maxeval}\{\Sigma\} \text{maxeval}\{(\gamma\gamma' \otimes I_K)\})^{n-1} \\ &= \text{maxeval}\{\Sigma\}^{n-1}, \end{aligned}$$

where the inequality follows from Ostrowski's theorem, see for example Theorem 4.5.9 in Horn and Johnson (2013), and the last step is due to the fact that the matrix  $\gamma\gamma'$  has only one non-zero eigenvalue that is equal to one. We therefore have the inequality

$$(C.31) \quad \text{Tr}(((\gamma\gamma' \otimes I_K)\Sigma)^n) \leq K\lambda_{\min} \text{maxeval}\{\Sigma\}^{n-1}.$$

Using the Lieb-Thirring inequality for positive semi-definite matrices, see Bernstein (2009) Fact 8.12.17,

$$\text{Tr}(((\gamma\gamma' \otimes I_K)\Sigma)^n) \leq \text{Tr}((\gamma\gamma' \otimes I_K)^n \Sigma^n) = \text{Tr}(((\gamma\gamma')^n \otimes I_K)\Sigma^n) = \text{Tr}((\gamma\gamma' \otimes I_K)\Sigma^n),$$

where the last equality results from the fact that the matrix  $\gamma\gamma'$  has only one non-zero eigenvalue that is equal to one. Since  $\text{Tr}((\gamma\gamma' \otimes I_K)\Sigma^n) = \gamma'R'(\Sigma^n \otimes I_K)R\gamma$  with  $\gamma'\gamma = 1$ , we have

$$(C.32) \quad \text{Tr}(((\gamma\gamma' \otimes I_K)\Sigma)^n) \leq \text{maxeval}\{R'(\Sigma^n \otimes I_K)R\}.$$

Applying the two inequalities (C.31) and (C.32) leads to the bounds in part (ii) of the theorem.

## D Proof of Proposition 3.

The distribution under the Imhof (1961) approximation in Definition 5 is

(D.33)

$$\begin{aligned} \Pr(\chi_\nu^2 < (x - \kappa_1)4\omega + \nu) &= \int_{\kappa_1 - \nu(4\omega)^{-1}}^x \phi(z) dz, \text{ where} \\ \nu &= 8\kappa_2\omega^2 \quad ; \quad \omega = \kappa_2/\kappa_3 \quad ; \quad \phi(z) = \left(1 + \frac{z - \kappa_1}{2\kappa_2\omega}\right)^{\nu/2-1} e^{-\frac{\nu}{2}\left(1 + \frac{z - \kappa_1}{2\kappa_2\omega}\right)} \frac{(\nu/2)^{\nu/2-1}\omega}{2^{\nu/2-2}\Gamma(\nu/2)}. \end{aligned}$$

The pdf  $\phi(z)$  has a mode at  $z^m = \kappa_1 - (2\omega)^{-1}$  if  $\nu \geq 2$ , and at zero otherwise. The critical value associated with the upper  $\alpha$ -percentile,  $x(\alpha)$ , is implicitly defined by  $\alpha = \int_{x(\alpha)}^{\infty} \phi(z) dz$ . To find the largest possible critical value among all possible distributions, we solve the following optimization problem:

$$(D.34) \quad \max_{\kappa_1, \kappa_2, \kappa_3} x(\alpha) \quad \text{s.t.} \quad \kappa_n \leq \bar{\kappa}_n \quad \text{for } n = 1, 2, 3 .$$

Consider the Kuhn-Tucker conditions

$$(D.35) \quad \int_{x(\alpha)}^{\infty} \frac{\partial \phi(z)}{\partial \kappa_n} dz = \mu_n,$$

together with  $\mu_n \geq 0$ ,  $n = 1, 2, 3$ , the constraints and the complementary slackness conditions, where  $\mu_n$  are the multipliers times  $\phi(x(\alpha)) > 0$ . The Kuhn-Tucker conditions follow from the implicit function theorem and Leibniz's rule:  $1 = -\phi(x(\alpha)) \frac{\partial x(\alpha)}{\partial y} + \int_{x(\alpha)}^{\infty} \frac{\partial \phi(z)}{\partial y} dz \Rightarrow \frac{\partial x(\alpha)}{\partial y} = \int_{x(\alpha)}^{\infty} \frac{\partial \phi(z)}{\partial y} dz / \phi(x(\alpha))$  with  $\phi(x(\alpha)) > 0$  for  $\alpha \in (0, 1)$ .

The partial derivatives are

$$(D.36) \quad \frac{\partial \phi(z)}{\partial \kappa_1} = \frac{1 + (z - \kappa_1)2\omega}{2\kappa_2\omega} \left( 1 + \frac{z - \kappa_1}{2\kappa_2\omega} \right)^{-1} \phi(z),$$

$$(D.37) \quad \frac{\partial \phi(z)}{\partial \kappa_2} = \frac{\phi(z)}{\kappa_2} G_1((z - \kappa_1)4\omega + \nu),$$

$$(D.38) \quad \frac{\partial \phi(z)}{\partial \kappa_3} = \frac{\phi(z)}{\kappa_3} G_2((z - \kappa_1)4\omega + \nu),$$

where

$$(D.39) \quad G_1(y) = -\frac{1}{2}(y - 2\nu(\nu - 2)/y + \nu) + 3/2(\ln(y/2) - \psi(\nu/2))\nu,$$

$$(D.40) \quad G_2(y) = \frac{1}{2}(y - \nu(\nu - 2)/y) - (\ln(y/2) - \psi(\nu/2))\nu,$$

and  $\psi(x) = \Gamma'(x)/\Gamma(x)$  is the digamma function (the logarithmic derivative of the gamma function  $\Gamma(x)$ ). From Alzer (1997) (equation 2.2), we know that

$$(D.41) \quad 1/\nu < \ln(\nu/2) - \psi(\nu/2) < 2/\nu.$$

For  $n = 1$ , the LHS of (D.35) is always positive to the right of the mode, which means the constraint on the mean ( $n = 1$ ) is always binding. The Alzer bounds imply that in the right tail of any optimal distribution, the LHS of (D.35) is always strictly positive for  $n = 2, 3$ , which means that the constraints are also binding as long as  $\alpha$  is sufficiently small.