Bayesian Inference in IV Regressions[*]

Domenico Giannone[†] Michele Lenza[‡] Giorgio E. Primiceri[§]

First version: August 2025. This version: December 2025

Abstract

It is well known that standard frequentist inference breaks down in IV regressions with weak instruments. Bayesian inference with diffuse priors suffers from the same problem. We show that the issue arises because flat priors on the first-stage coefficients overstate instrument strength. In contrast, inference improves drastically when an uninformative prior is specified directly on the concentration parameter—the key nuisance parameter capturing instrument relevance. The resulting Bayesian credible intervals are asymptotically equivalent to the frequentist confidence intervals based on conditioning approaches, and remain robust to weak instruments.

1 Introduction

In this paper, we study Bayesian inference in the canonical instrumental variables (IV) model. We argue that a flat prior on the concentration parameter—the crucial nuisance parameter that captures instrument strength—best embodies the notion of non-informativeness in this model. We derive such prior and show that it yields posterior inference with both desirable frequentist properties and robustness to weak instruments.

It is well documented that conventional asymptotic approximations to the distribution of standard IV estimators—such as the popular Two-Stage Least Squares (TSLS) estimator—are

[*] We thank Otilia Boltea, Gary Koop, Daniel Lewis, Mike McCracken, Karel Mertens and José Luis Montiel Olea, as well as many seminar and conference participants for comments and suggestions. The views expressed here are those of the authors and do not necessarily represent those of the European Central Bank or the Eurosystem.

[†] Johns Hopkins University and CEPR

[‡] European Central Bank and CEPR

[§] Northwestern University, CEPR and NBER

unreliable when the instruments are weakly correlated with the endogenous regressors. The reason is that the distribution of these estimators depends on a key nuisance parameter, known as the concentration parameter, which reflects the strength of the instruments. When the instruments are weak, there is little sample information about the parameter of interest—a situation analogous to having a small sample size (Rothenberg, 1984). Correspondingly, the small value of the concentration parameter distorts the shape of the estimator's distribution, undermining the adequacy of its Gaussian asymptotic approximation. Moreover, small values of the concentration parameter cannot be consistently estimated, limiting the ability to adjust the asymptotic distribution to better reflect finite-sample behavior.

Bayesian inference based on naive priors suffers from similar problems and can exhibit poor frequentist properties. For example, when instruments are weak, a flat prior on the first-stage coefficients leads to a posterior distribution for the parameter of interest that is excessively concentrated away from the true value. This occurs even in simulation settings where the likelihood function is correctly specified, indicating that the issue lies with the prior. We show that this pathology stems from the induced prior on the concentration parameter. To see why as simply as possible, suppose that the model involves only two unknown parameters: the parameter of interest, call it β, and the (nuisance) concentration parameter, denote it by μ^2. For the sake of clarity, we temporarily abstract from other unknowns such as error variances, covariances, and so on. In this simplified setting, Bayesian inference yields a joint posterior distribution over β and μ^2. To obtain the marginal posterior of β, we need to integrate out μ^2 with respect to its prior. If this prior favors large values of μ^2—that is, regions of the parameter space where the instruments are strong and their variation is very informative about β—the resulting marginal posterior of β is very concentrated. As it turns out, a flat prior on the first-stage coefficients unintentionally induces exactly this kind of prior on μ^2, particularly when the number of instruments is moderate or large. With one instrument, instead, the implied prior on μ^2 is overly concentrated around 0.

Building on this intuition, we argue that this pathology can be entirely resolved by eliciting a prior that places neutral weight over the concentration parameter space, avoiding undue emphasis on regions associated with strong instruments. We derive such a prior and show that it belongs to a class of distributions that reduces overfitting by shrinking the first-stage coefficients toward zero, and that it can be implemented through a hierarchical specification. We also demonstrate that the frequentist properties of Bayesian inference improve substantially when using this prior. Specifically, the resulting credible intervals are conservative when instruments are irrelevant, and they achieve correct coverage when instruments are weak or strong. In sum, this approach yields Bayesian inference that is robust to weak instruments. Moreover, these credible intervals are nearly identical to those obtained by inverting the Conditional Likelihood Ratio test of Moreira (2003).

We establish these results using four complementary approaches: prior elicitation arguments, simulation evidence, a revisit of the classic application of estimating the return to schooling (Angrist and Krueger, 1991), and a theoretical exploration of the connections between Bayesian and classical inference in IV regressions. Regarding the latter, we are able to show that our Bayesian approach has desirable frequentist properties, closely mirroring those of conditional (similar) inference, due to the existence of a one-to-one re-parameterization of the model that satisfies two key properties: (i) the likelihood function is symmetric in the deviation of the parameter of interest from its maximum likelihood estimate; and (ii) the parameter of interest and the nuisance parameter are orthogonal, in the sense that the Fisher information matrix is diagonal. Intuitively, the first property ensures that, conditional on the nuisance parameter, both the posterior and the distribution of the maximum likelihood estimator of the parameter of interest are nearly Gaussian. The second property implies that these distributions are nearly free of the nuisance parameter, because there exist a statistic—depending only on the data—that is approximately sufficient for it. As a result, integrating out the nuisance parameter based on the observed data, as in Bayesian inference, or conditioning inference on a sufficient statistic for it, as in Moreira (2003), leads to very similar inferential outcomes. To be clear, these are not finite-sample results, but are established under a "many-instruments" asymptotic framework, in which the number of instruments k grows with the sample size T. Nevertheless, we show that the rate of convergence in our asymptotic results is twice as fast as that of conventional asymptotics: Our approximations improve at rate $1/k$, rather than the usual $1/\sqrt{k}$. Consequently, as confirmed by our simulations, the results hold remarkably well even with a small number of instruments—including a single one. In sum, our contributions can be viewed as restoring the alignment between classical and Bayesian inference in the non-regular IV setting with weak instruments.

1.1 An overview of the related literature

Nelson and Startz (1990b,a) and Bound et al. (1995) provide early simulation evidence that standard IV estimators are biased and lead to invalid inference when instruments are weak. In particular, Bound et al. (1995) study how the weak instrument problem affects the analysis of the returns to schooling in Angrist and Krueger (1991). It is widely regarded as the first paper to formally diagnose the weak instruments problem in applied econometrics, effectively launching the modern methodological literature on the topic and spurring greater awareness in empirical work.

The seminal contribution by Staiger and Stock (1997) provided the first formal definition of weak instruments in large samples, showing that standard Gaussian-based inference can break down when the concentration parameter is small. They introduced the first-stage F-statistic as a diagnostic and proposed the now-standard rule of thumb: If $F < 10$, instruments may be too weak for reliable inference. Stock et al. (2002) and Stock and Yogo (2005) extended the analysis of Staiger and Stock (1997) to GMM settings and models with multiple endogenous regressors, developing formal weak instrument tests with critical values based on tolerable bias or size distortion. Together, these contributions have established the F-statistic as the standard diagnostic tool for assessing instrument strength in applied work. More recent work has focused on testing for weak instruments under heteroskedasticity, with either single or multiple endogenous regressors (see Montiel Olea and Pflueger, 2013, Lewis and Mertens, 2025, and Andrews et al., 2019, for a comprehensive survey). In contrast to this line of research, we focus on developing a method that yields valid inference regardless of instrument strength.

Our paper is more closely related to studies that propose inference procedures that remain valid under weak identification. The classic test of Anderson and Rubin (1949) is based on a pivotal statistic and retains correct size under the null, but it tends to be conservative and has low power in over-identified settings. The test of Kleibergen (2002) and Kleibergen (2005) is also based on an asymptotically pivotal statistic whose distribution does not depend on the nuisance parameter, and it is typically more powerful than the Anderson-Rubin's test, even though its power function can be non monotonic. A central contribution to this literature is the work of Moreira (2003), which introduces the Conditional Likelihood Ratio test, leveraging the distribution of the likelihood ratio statistic conditional on a sufficient statistic for the nuisance parameter. This test has excellent power properties, and Andrews et al. (2006) show that it is essentially optimal among all similar tests that are invariant to rotations of the instruments. More recently, Lee et al. (2022) have proposed an adjusted t-test for the IV model with a single instrument. Their adjustment of the TSLS asymptotic standard errors is a function of the first-stage F-statistic. Like these procedures, our approach delivers robust inference in the presence of weak instruments, offering a Bayesian alternative to frequentist tests.

Several studies prior to ours have examined Bayesian inference in IV regressions. Kleibergen and Zivot (2003) survey the early literature, including the approach of Drèze (1976) based on uninformative priors, and the subsequent critique by Maddala (1976). The latter highlights that flat priors can lead to misleadingly sharp posteriors for the parameter of interest even when the model is not identified, which is one of the observations that motivate our paper. Kleibergen and van Dijk (1998), Chao and Phillips (1998) and Hoogerheide et al. (2007b) stress that flat priors also lead to a local non-identification problem, because the parameter of interest is not identified if the first-stage coefficients are equal to zero, an issue that renders the posterior improper if the number of instruments and endogenous variables is the same. Kleibergen and Zivot (2003) derive the priors that yield posterior distributions with the same functional form as the sampling distributions of the TSLS and Limited Information Maximum Likelihood (LIML) estimators. The latter, a Jeffreys prior, was also discussed in the analyses of Kleibergen and van Dijk (1998), Chao and Phillips (1998) and Hoogerheide et al. (2007a), among others. While the Jeffreys prior is able to overcome the local non-identification problem, none of these priors exhibit the same robustness to weak instruments that characterizes inference with our flat prior on the concentration parameter. From this perspective, we argue that such prior best embodies the notion of non-informativeness in IV models. This contribution also resolves the ambiguity left in Chamberlain (2007), who acknowledged the challenge of specifying a prior on some parameters in his decision theoretic analysis of the IV model. In addition, as we have stressed above, our prior also yields posteriors with favorable frequentist properties and aligns closely with the conditional inference of Moreira (2003) and the optimal procedures developed by Andrews et al. (2006).

Our paper also relates to work that leverages Bayesian ideas to develop improved frequentist procedures, including Chamberlain and Imbens (2004), Chamberlain (2007) and Montiel Olea (2020). These contributions treat only certain parameters as random, as a device to regularize inference or average loss functions. This perspective has supported the construction of decision rules and confidence intervals with strong frequentist properties under weak identification. In contrast, we adopt a fully Bayesian approach based on a hierarchical prior and study the frequentist properties of the resulting posterior.

We study these properties in a setting where the number of instruments grows large, connecting our work to the literature on IV regressions with many instruments. As emphasized by Mikusheva (2020) and Mikusheva and Sun (2024), adding instruments brings informational gains, but also contributes to overfitting in the first-stage regression. This overfitting, in turn, introduces bias into the second-stage estimation of the parameter of interest. In fact, the TSLS estimator is consistent only when instrument strength grows faster than the number of instruments—that is, when $\mu^2/k \to \infty$ (Bekker, 1994). The literature has addressed this bias through two related but distinct approaches. The first emphasizes bias correction by breaking the endogeneity caused by using the same data in the estimation of both model equations. Examples include sample-splitting, jackknife and deleted-diagonal estimators (Angrist and Krueger, 1995, Angrist et al., 1999, Hansen et al., 2008). The second approach focuses on regularization techniques designed to reduce the effective dimensionality of the instrument set, including principal component methods, ridge regressions, random coefficient models, and sparsity-inducing methods such as the lasso (Bai and Ng, 2009, 2010, Kapetanios and Marcellino, 2010, Kapetanios et al., 2016, Carrasco, 2012, Carriero et al., 2020, Chamberlain and Imbens, 2004, Belloni et al., 2012).

The paper most closely related to ours is Chamberlain and Imbens (2004). As mentioned earlier, their approach is not explicitly Bayesian, but their modeling of the first-stage coefficients as random coefficients resembles our hierarchical shrinkage prior. Despite these similarities, there are two important differences between our work and theirs. First, we show that shrinkage priors can enhance and robustify inference even in settings with very few instruments, not necessarily many, like in their work. Second, we provide both theoretical and simulation-based evidence that our approach aligns with conditional frequentist inference, which has been shown to achieve near-optimality. In fact, a key aspect of our contribution is demonstrating that our Bayesian methodology yields credible intervals that are almost equivalent to the confidence intervals produced by frequentist approaches based on conditioning, and are robust to weak instruments.

To establish these theoretical results, we adopt the simultaneous asymptotic framework introduced by Bekker (1994), and extended by Chao and Swanson (2005), Stock and Yogo (2005), and Andrews and Stock (2007b,a), in which both the sample size and the number of instruments increase. In this setting, Mikusheva and Sun (2022) demonstrate that $\mu^2/\sqrt{k} \to \infty$ is a necessary condition for consistent estimation when the number of instruments is large and no assumptions are made on the form of their optimal combination (Mikusheva and Sun, 2024 survey this type of assumptions commonly imposed in the literature and emphasize that the performance of the associated estimators can deteriorate substantially if the true data-generating process fails to satisfy them; a related point has been made by Giannone et al., 2021 in the context of prediction models). Crudu et al. (2021), Anatolyev and Sølvsten (2023), Matsushita and Otsu (2024) and Mikusheva and Sun (2022, 2024)—propose pre-testing procedures and robust Anderson-Rubin and Lagrange-Multiplier tests based on the jackknife estimator that perform well even in severely adverse scenarios in which the condition $\mu^2/\sqrt{k} \to \infty$ is violated. In comparison, our results require $\mu^2/k \to \underline{c}$, where $\underline{c}$ is an arbitrarily small constant—a condition that accommodates virtually all empirically relevant configurations. Rather than focusing on robustness to asymptotic assumptions, we examine the frequentist validity of Bayesian approaches and their robustness with respect to the prior. Our theoretical contribution is to characterize the conditions under which Bayesian credible intervals converge to frequentist confidence intervals, and to show that this convergence occurs at a faster rate than under conventional asymptotics. In this sense, our results can be viewed as a Bernstein–von Mises–type theorem as they restore the concordance between classical and Bayesian inference in the non-regular IV setting with weak instruments.

1.2 The outline of the paper

The remainder of the paper is organized as follows. Section 2 reviews the problem of weak instruments in IV regressions, with particular attention to the limitations of standard Bayesian inference in this context, which resemble those of frequentist methods. Section 3 introduces a Bayesian approach based on an uninformative prior on the concentration parameter, and shows that it delivers posterior inference with both desirable frequentist properties and robustness to weak instruments. Section 4 revisits the classic application of estimating the return to schooling using multiple quarter-of-birth instruments. Section 5 presents the theoretical results on the connection between our proposed Bayesian procedure and frequentist inference. Section 6 concludes.

2 IV regressions and the problem of weak instruments

Consider the model

$$y = x\beta + \delta\nu + \varepsilon \tag{1}$$

$$x = z\pi + \nu, \tag{2}$$

where $y \in \mathbb{R}^T$ is an observed dependent variable, $x \in \mathbb{R}^T$ is an observed regressor, $z \in \mathbb{R}^{T\times k}$ are observed instrumental variables, $\nu, \varepsilon \in \mathbb{R}^T$ are unobserved shocks uncorrelated with each other, and $\beta \in \mathbb{R}$, $\delta \in \mathbb{R}$ and $\pi \in \mathbb{R}^k$ are unknown parameters.[1] The parameter of interest is β,

[1] Equations (1) and (2) can easily be extended to incorporate additional controls, but we omit them here for clarity, simplifying both the notation and the presentation of our contribution. Appendix B provides full details on the

which captures the causal effect of x on y. If $\delta = 0$, Ordinary Least Squares (OLS) is a consistent estimator of β, because ε is assumed to be uncorrelated with the regressor x. But when $\delta \neq 0$, the OLS estimator of β becomes biased and inconsistent—a well known result.

The key insight behind IV regressions is that, in such cases, we can consistently estimate β by leveraging the variation of x induced by z, as long as these instruments satisfy two crucial conditions: (i) *exogeneity*—z must be uncorrelated with the shocks ν and ε; and (ii) *relevance*—z must be (possibly strongly) correlated with the endogenous regressor x. The most widely used IV estimator of β is the Two-Stage Least Squares (TSLS) estimator, which is obtained by first regressing x on z, and then using the fitted value $\hat{x}$ in place of x in the second-stage regression of y on $\hat{x}$:

$$\hat{\beta}_{TSLS} = \frac{\hat{x}'y}{\hat{x}'\hat{x}},$$

where $\hat{x} = \left(z'z\right)^{-1} z'x$. Under the assumption of conditional homoskedasticity, the asymptotic variance of the TSLS estimator can be estimated by $\hat{\sigma}^2 \left(\hat{x}'\hat{x}\right)^{-1}$, where $\hat{\sigma}^2$ is a consistent estimator of the variance of $\delta\nu + \varepsilon$.

To fix ideas, consider the well-known problem of estimating the returns to schooling (Angrist and Krueger, 1991, AK hereafter). In this example, y represents the logarithm of the wage earned by a typical US male, x denotes his years of schooling, and β captures the causal effect of an additional year education on wages—the object of interest. The major challenge in estimating β using a simple regression of y on x is endogeneity: Individuals with higher innate ability may choose to stay in school longer, leading to a biased estimate of β. To address this concern, AK have proposed using an individual's quarter of birth as an instrumental variable for his years of schooling. They argued that being born in September versus April should have no direct impact on earnings, making the quarter of birth a plausibly exogenous instrument. As for instrument relevance, the quarter of birth influences schooling because state laws on compulsory education require children to start school in September of the year they turn 6, and to remain in school until at least age 16. As a result, individuals born later in the year tend to stay in school longer.

But what if the instruments are only marginally relevant, meaning they are only weakly correlated with the endogenous regressors? Bound et al. (1995) argued that, in such cases, conventional IV methods become unreliable, including TSLS, and that this issue may have affected the analysis of AK. These findings have sparked an extensive literature, leading to the consensus that, with weak instruments, standard "IV estimators can be badly biased, while t-tests may fail to control size, and conventional IV confidence intervals may cover the true parameter value far

estimation procedure, including the treatment of additional controls.

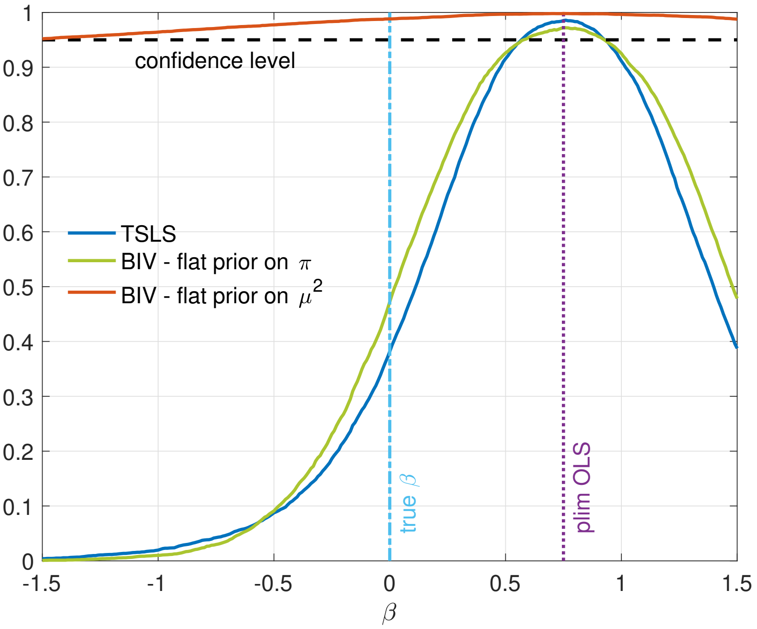

Figure 1: Frequency of inclusion in the 95-percent TSLS confidence interval, the 95-percent Bayesian credible interval based on a flat prior on π, and the 95-percent Bayesian credible interval based on a flat prior on μ^2. The results are based on $10,000$ simulations from model (1)-(2) with $T = 250$, $\beta = 0$, $\varepsilon \sim N(0, I_T)$, $\nu \sim N(0, I_T)$, $\pi = 0_{k\times 1}$, $\delta = 0.75$ and $k = 10$.

less often than intended" (Andrews et al., 2019, p. 728).

To highlight the severity of the problem, let us briefly examine the extreme case of completely irrelevant instruments. Specifically, we simulate $10,000$ datasets from (1)-(2) using the parameter values $T = 250$, $k = 10$, $\beta = 0$, $\delta = 0.75$, $\varepsilon \sim N(0, I_T)$, $\nu \sim N(0, I_T)$, and $\pi = 0$. For each simulation, we construct the 95-percent confidence interval for β based on the conventional asymptotic distribution of the TSLS estimator. The blue line in figure 1 summarizes the results of this Monte Carlo experiment. It represents the frequency across simulations with which each possible $\beta \in [-1.5, 1.5]$ falls within the 95-percent TSLS confidence interval. Unfortunately, the true value of $\beta = 0$ belongs to the 95-percent confidence interval only 40 percent of the times, not 95, revealing a massive size distortion. Instead, values of β between 0.5 and 1 are almost always included in the 95-percent confidence interval. This pattern arises because, with irrelevant instruments, TSLS converges in probability to the (biased) probability limit of the OLS estimator, which in our simulations is given by $\beta + \delta = 0.75$. In addition, the standard errors of the TSLS estimator are way too tight.

The green line in figure 1 represents the frequency across simulations with which each possible $\beta \in [-1.5, 1.5]$ falls within the equal-tailed 95-percent *Bayesian* credible interval, obtained using a flat prior on all the model parameters of (1)-(2), including π. Notably, the green line tracks the blue line, suggesting that the frequentist properties of flat-prior Bayesian methods are just as poor as those of TSLS. The most puzzling aspect of this result is that the Bayesian approach behind the green line in figure 1 relies on the likelihood function, which, according to the likelihood principle, should fully encapsulate all sample evidence relevant for inference. Moreover, the likelihood function is correctly specified in our controlled experiment. So, is the likelihood principle somehow failing in our context? Of course not. In the next section, we demonstrate that the issue stems from an implicit prior that distorts Bayesian inference by unintentionally favoring instrument strength, even when the instruments are weak in practice. From a Bayesian perspective, "correcting" this prior resolves the problem.

3 Robust Bayesian inference

To develop more intuition about the weak instrument problem, suppose that a researcher observes data generated by (1)-(2), knowing that $\pi = 0$. In this situation, x and ν are perfectly collinear, and only $\beta + \delta$ is identified in equation (1), not β and δ separately. Put differently, a researcher who knows that $\pi = 0$ would end up with an infinitely wide confidence interval for β. If the instruments are truly irrelevant, this is the correct conclusion: There is nothing we can say about the causal effect of x on y. This simple argument clarifies why conventional IV confidence intervals are too tight when instruments are irrelevant, as implied by the results displayed in figure 1. It is because researchers do not know that $\pi = 0$ but must estimate it, leading to an estimate of π that differs from the true value of 0—a classic case of overfitting.

To formalize this argument from a Bayesian perspective, note that the posterior variance of β, conditional on the other parameters, is given by

$$\text{var}\left(\beta | y, x, z, \pi, \sigma_\varepsilon^2, \sigma_\nu^2\right) \approx \frac{\sigma_\varepsilon^2}{\sigma_\nu^2} \cdot \frac{1}{\mu^2}, \tag{3}$$

where $\mu^2 = \frac{\pi' z' z \pi}{\sigma_\nu^2}$ is the so-called concentration parameter. The larger μ^2, the greater the share of variation of x induced by the exogenous instruments z, and the lower our posterior uncertainty about β. Conversely, if we knew with certainty that $\pi = 0$, then μ^2 would also be 0, and the posterior variance of β would be infinity. In other words, if we knew that $\pi = 0$, the green line in figure 1 would be flat and equal to 1. The fact that it is not indicates that the posterior of μ^2 does not concentrate on 0, but on large positive value. Why does this happen?

Since the likelihood function is correctly specified in the Monte Carlo experiment of section 2, this distortion in the posterior distribution must originate from the assumed prior. To understand why, consider the possible prior distribution for π given by

$$\pi|\gamma^2, \sigma_\nu^2 \sim N\left(0, \gamma^2 \sigma_\nu^2 \left(z'z\right)^{-1}\right), \tag{4}$$

where γ^2 is a hyperparameter controlling the prior tightness, while σ_ν^2 and $\left(z'z\right)^{-1}$ are convenient scaling factors. Given (4), it is easy to verify that the implied prior on μ^2 is a scaled chi-square distribution with k degrees of freedom, i.e.

$$\mu^2|\gamma^2 \sim \gamma^2 \cdot \chi_k^2.$$

Its expected value is $E\left(\mu^2|\gamma^2\right) = \gamma^2 k$ and its mode is $\gamma^2 \max\left(k-2, 0\right)$. The larger the value of γ^2, the flatter the prior on π, and the more the prior distribution of μ^2 shifts its mass towards large positive values. In addition, the extent of this shift depends on the number of instruments, with more instruments worsening the problem. This is intuitive: As the prior on π becomes more diffuse, it places increasing probability mass on large absolute values of π, with the unintended consequence of concentrating more mass on high values of the concentration parameter μ^2. In sum, the more agnostic the prior is on π, the more informative it becomes on μ^2, which is the parameter that matters most in the weak instruments setting. Finally, notice that the flat prior on π used to generate the green line in figure 1 is a special case of (4), corresponding to infinite variance, i.e. $\gamma^2 \to \infty$.

This insight into the root of the problem—as outlined above—is essential, as it naturally points toward a solution. Since the issue arises from the fact that a flat prior on π favors high values of the concentration parameter, and thus stronger instruments, it is natural to consider a prior that is instead flat on μ^2 directly. The following proposition outlines how to obtain such a prior.

Proposition 1. *The improper prior* $p\left(\pi|\sigma_\nu^2\right) \propto \left(\frac{\pi' z' z \pi}{\sigma_\nu^2}\right)^{-\frac{k-2}{2}} \left(\frac{1}{\sigma_\nu^2}\right)^{\frac{k}{2}}$ *implies a flat prior on the concentration parameter* μ^2*. When* $k > 2$*, this prior is equivalent to a hierarchical specification that combines (4) with a flat hyperprior on* γ^2*. The corresponding posterior is always proper.*

Proof. See appendix A. □

Remark 1. Proposition 1 highlights that the prior on π that induces a flat prior on μ^2 depends on the number of instruments k. A flat prior on π yields a flat prior on μ^2 only when $k = 2$. For $k > 2$, as noted above, a flat prior on π places disproportionate mass on large values of μ^2, effectively favoring strong instruments. Conversely, when $k = 1$, it produces an asymptote at zero in the implied prior for μ^2, making the naive-Bayesian approach overly conservative in the exactly identified case. The prior on π in Proposition 1 offsets these distortions and always yields an uninformative prior on μ^2. When $k = 2$, no adjustment is needed and the prior on π is flat; when $k = 1$, the required prior on π shifts mass away from zero; and when $k > 2$, the prior on π is a shrinkage prior concentrating near zero.

Remark 2. It is useful to compare the prior in Proposition 1 with the Jeffreys prior discussed in the work of Kleibergen and van Dijk (1998), Chao and Phillips (1998), Kleibergen and Zivot (2003) and Hoogerheide et al. (2007a), among others. The Jeffreys prior is proportional to $(\pi' z' z\pi)^{1/2}$ (see, for example, equation (21) in Chao and Phillips, 1998). Therefore, it shifts mass on π in a manner similar to our prior only in the exactly identified case of $k = 1$. Unlike the prior in Proposition 1, however, the Jeffreys prior does not depend on k. Consequently, when $k > 1$, it induces no shrinkage and places excessive mass on large values of the concentration parameter. In this sense, it suffers from precisely the problem that motivates our approach.

In what follows, we adopt a diffuse prior for all other model parameters, β, δ, σ_ε^2 and σ_ν^2, and evaluate the posterior distribution using the simple Gibbs sampling algorithm described in appendix B.[2]

Does this prior—flat on μ^2 rather than on π—lead to substantially different posterior inference? Yes, it does, as illustrated by the behavior of the red line in figure 1. The position and shape of this line indicate that the 95-percent Bayesian credible intervals obtained under a flat prior on μ^2 are very wide across nearly all simulations with irrelevant instruments—almost always covering the entire range of $\beta \in [-1.5, 1.5]$. Put differently, Bayesian inference based on a flat prior on μ^2 is far more effective than both TSLS and the Bayesian approach with a flat prior on π at capturing the substantial uncertainty surrounding β that arises when instruments are truly irrelevant. From now on, for brevity, we will refer to the Bayesian approach with a flat prior on π as the *naive-Bayesian* approach (NB-IV), and to the Bayesian approach with a flat prior on μ^2 as the *weak-instrument-robust-Bayesian* approach (WIRB-IV).

[2] An alternative to (4) could be $\pi|\gamma^2, \sigma_\nu^2 \sim N\left(0, \gamma^2\sigma_\nu^2 I_k\right)$, which closely resembles the distribution of the random coefficients proposed by Chamberlain and Imbens (2004). A drawback of this prior, however, is that when combined with a flat hyperprior on γ^2, it induces only an approximately flat prior on μ^2 , rather than an exactly flat one. That said, the posterior results obtained using the two priors are virtually indistinguishable in our experiments with real and simulated data.

So far, our Monte Carlo experiment has focused exclusively on the case of irrelevant instruments. To evaluate the performance of our approach when instruments are weak—but not entirely irrelevant—we now conduct an additional simulation study. Specifically, we simulate $25,000$ datasets using the same parameter values as in section 2, except that we no longer set $\pi = 0$. Instead, π is drawn from a $\mathcal{N}(0, s^2 I_k)$, where s is uniformly distributed between 0 and 0.25. The exact range of s is not critical; what matters is that the resulting distribution of first-stage F-statistics spans values from 0 up to approximately 20. Since F-statistics below 10 are typically associated with weak instruments (Staiger and Stock, 1997, Stock and Yogo, 2005), this setup allows us to evaluate the performance of WIR-BIV in both weak and moderately strong instrument scenarios.

Figure 2 summarizes the results of this simulation experiment. As before, the solid lines show the frequency with which the confidence (or credible) intervals from various inferential methods cover different values of β. We compare the performance of WIRB-IV to NB-IV, TSLS and Moreira's (2003) conditional likelihood ratio test (CLR)—a frequentist procedure that is robust to weak instruments. The CLR test is based on the likelihood ratio statistic, whose distribution is evaluated conditional on the observed value of another statistic that is sufficient for the nuisance parameters. As a result, the CLR test has always the correct size, as evident from figure 2. The corresponding confidence intervals are obtained by inverting the test, as in Mikusheva and Poi (2006) and Mikusheva (2010).

Panel (a) presents the results of the simulations in which the first-stage F-statistic falls between 0 and 2—a scenario with essentially irrelevant instruments. As expected, TSLS-based tests exhibit substantial size distortion, and the corresponding confidence intervals concentrate around the biased OLS estimate, like in figure 1. As shown earlier, NB-IV performs similarly to TSLS. In contrast, the CLR test maintains correct size, and the confidence intervals obtained by inverting it are generally quite wide. Notice that the red line in panel (a) is above the yellow one, suggesting that our WIRB-IV approach is even more conservative than CLR, yielding even wider credible intervals. The subsequent panels of the figure examine groups of simulations with progressively higher F-statistics. The discrepancy between WIRB-IV and TSLS (as well as NB-IV) remains substantial until the F-statistic approaches 10. Remarkably, in the simulations with F-statistics between 2 and 4 (panel b), the results based on WIRB-IV and CLR are strikingly similar—and for values above 4, the two methods become virtually indistinguishable. Appendix D shows that similar results hold across different simulation settings, with the number of instruments equal to 1, 5, 25, or 100, and with a higher degree of endogeneity ($\delta = 3$, corresponding to

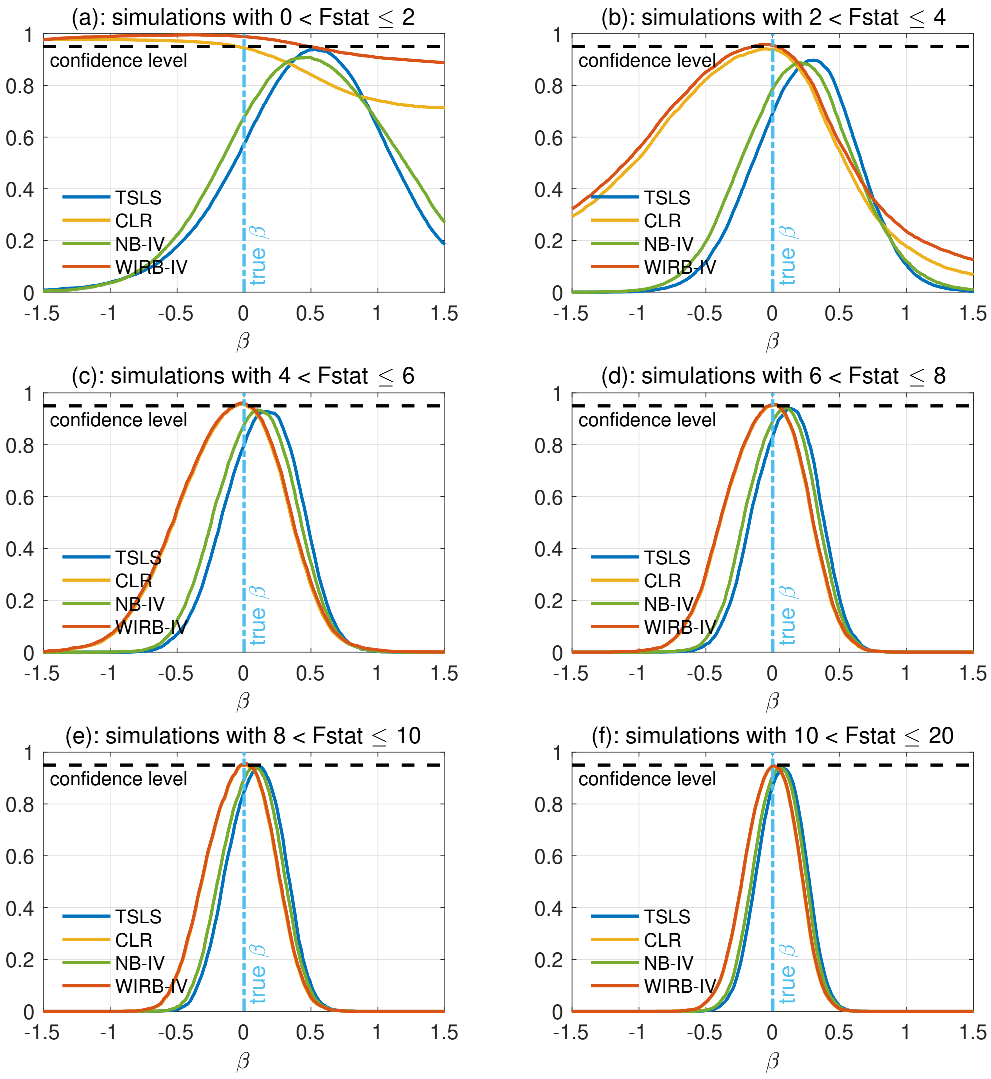

Figure 2: Frequency of inclusion in the 95-percent confidence (credible) interval based on (i) the asymptotic distribution of TSLS; (ii) the inversion of Moreira's (2003) conditional likelihood ratio test (CLR); (iii) the naive-Bayesian approach (NB-IV); and (iv) our weak-instrument-robust Bayesian approach (WIRB-IV). The results are based on $25,000$ simulations from model (1)-(2) with $T = 250$, $\beta = 0$, $\varepsilon \sim N(0, I_T)$, $\nu \sim N(0, I_T)$, $\pi \sim \mathcal{N}(0, s^2 I_k)$, $s \sim Uniform(0, 0.25)$, $\delta = 0.75$ and $k = 10$.

a correlation of the structural shocks of approximately 0.95).

The similarity between WIRB-IV and CLR in these simulations is notable for at least two reasons. First, there is no known theoretical justification for expecting the two procedures to produce such similar results. Second, Andrews et al. (2006) have shown that Moreira's (2003) CLR test is nearly optimal among all invariant similar tests, which implies that the robust Bayesian method proposed in this paper achieves near-optimality as well. In section 5, we will explore these observations in more depth, demonstrating that they are not mere coincidences but reflect strong theoretical links between the two approaches.

It is also instructive to compare these approaches with the popular pre-testing approach, which remains the most widely used method among applied researchers concerned with weak instruments. This approach involves estimating β using TSLS only if the first-stage F-statistic exceeds 10, following the recommendations of Staiger and Stock (1997) and Stock and Yogo (2005). Importantly, applying this rule would lead to infinitely wide confidence intervals in all the scenarios shown in panels (a) through (e), resulting in a substantial loss of information. In contrast, for panel (f), where the F-statistic exceeds 10, the pre-testing approach would yield the same result as TSLS.

4 Estimating the returns to schooling

In this section, we revisit the AK estimates of the returns to schooling using our WIRB-IV approach. The data unambiguously show that higher levels of education are associated with higher earnings (for instance, see U.S. Bureau of Labor Statistics, 2024). However, establishing a causal link between the two has proven challenging, mostly because a simple regression of earnings on years of schooling is likely to suffer from an endogeneity bias (see Card, 1999 and, more recently, Gunderson and Oreopolous, 2020, for a survey of the literature). For example, pre-schooling levels of ability may affect both the level of education, via schooling choices, and earnings, as highlighted by Griliches (1977) and Card (2001), among others.

To address this endogeneity bias, AK have proposed using an individual's quarter of birth as an instrumental variable for years of schooling. The validity of this instrument relies on a specific institutional feature of the U.S. education system: In many states, compulsory schooling laws require students to start going to school in September of the year in which they turn 6, and allow students to leave school upon turning 16. As a result, individuals born early in the calendar year tend to reach the minimum school-leaving age during an earlier grade than those born later, making them more likely to leave school with fewer years of education. Using the quarter of birth as an instrument, AK find that the return to education is both economically and statistically significant.

In an influential paper, however, Bound et al. (1995) have raised concerns about some of the IV regressions in AK, due to the weak correlation between the instrumental variables and the endogenous regressor measuring the level of education. Their criticism was particularly geared toward AK's least parsimonious model, a specification with a large set of instruments obtained by interacting the quarter of birth with the year and the state of birth of an individual. To demonstrate the severity of the problem, Bound et al. (1995) generate 500 artificial datasets by interacting the year and state of birth of the individuals with *random quarters of birth*. Their work show that standard IV estimators, such as TSLS, yield economically and statistically significant estimates of the return to education even with entirely "fake" instrumental variables that are uncorrelated with the endogenous regressor. Moreover, these estimates are similar in magnitude to those obtained using the actual data. These results illustrate how the large-sample approximation of the distribution of conventional IV estimators can be severely misleading when instruments are weak, as we have seen in section 2. Since the publication of Bound et al. (1995), the AK study has become a testing ground for assessing the extent to which various methodologies are robust to the presence of weak instruments (Angrist and Krueger, 1995, Angrist et al., 1999, Kleibergen, 2002, Chamberlain and Imbens, 2004, Imbens and Rosenbaum, 2005, Cruz and Moreira, 2005, Hoogerheide and van Dijk, 2006, Hoogerheide et al., 2007a, Hansen et al., 2008, Andrews and Armstrong, 2017, Mikusheva and Sun, 2024).

In the remainder of this section, we replicate the AK estimates using TSLS and compare them to those obtained with our WIRB-IV approach. The data come from the 1980 U.S. Census and consist of men born between 1930 and 1939.[3] In the notation of equations (1)-(2), the dependent variable (y) is the logarithm of weekly wages in 1979, while the endogenous regressor (x) represents the years of completed schooling. We focus on the specification that was under particular scrutiny in Bound et al. (1995). It includes 177 instruments (z), constructed by interacting individuals' quarter of birth with their year and state of birth. The model also includes control variables for year and state of birth, race, marital status, census region, and residence in a Standard Metropolitan Statistical Area. The dataset comprises a large sample of $329,509$ individuals.

[3] The original dataset and the AK replication package can be downloaded at https://economics.mit.edu/people/faculty/josh-angrist/angrist-data-archive.

	Actual Data	Artificial Data
TSLS	0.083 [0.065 – 0.102]	98.2%
NB-IV	0.083 [0.064 – 0.101]	97.6%
WIRB-IV	0.100 [0.070 – 0.132]	0.2%
CLR	NA [0.065 – 0.127]	4.4%

Table 1: Estimation results of the AK model with actual and artificial data. The column "Actual Data" reports point estimates and 95-percent confidence intervals (credible intervals, in the NB-IV and WIRB-IV cases). The CLR method only yields confidence sets, so we omit its point estimate. The column "Artificial Data" displays the percentage of simulations with random instruments yielding statistically significant estimates at the 5-percent level.

Table 1 reports our estimation results. The first entry of the table replicates the estimates of AK using TSLS, suggesting that an additional year of schooling increases an individual's weekly earnings by 8.3 percent.[4] The corresponding 95-percent confidence interval is relatively tight, ranging from 6.5 to 10.2 percent. However, the first-stage F-statistic is only 2.43, indicating that the instruments are likely to be very weak, and casting doubt on the reliability of these estimates. Indeed, when we repeat the analysis using artificial data (second column of table 1), following the design of Bound et al. (1995), we find that nearly all the 500 simulated datasets produce statistically significant estimates at the 5-percent level—despite the instruments being randomly generated. The Naive-Bayesian approach performs almost exactly like TSLS, as reported in the second row of the table.

The third row of table 1 shows that the point estimate based on WIRB-IV is 10 percent—broadly similar to, if not slightly higher than the TSLS and NB-IV estimates—although the associated credible interval is somewhat wider. Importantly, WIRB-IV correctly recognizes that the artificial instruments provide little information about the return to education. In fact, the estimated effect is not statistically distinguishable from zero in the vast majority of the artificial datasets. For comparison, the final row of the table reports the estimation results based on Moreira's (2003) CLR method, which are nearly identical to those obtained using our Bayesian approach. Like WIRB-IV, CLR yields statistically significant estimates in artificial datasets only a small fraction of the time. This is expected, since CLR is specifically designed to control size in the presence of weak instruments.

[4] These estimates correspond to those of table VII (column 6) in Angrist and Krueger (1991).

5 Theoretical results

In the previous sections, we have used prior elicitation arguments, simulation evidence, and an empirical application to show that Bayesian inference based on a flat prior over the concentration parameter delivers both desirable frequentist properties and robustness to weak instruments. This section complements the earlier analysis by examining the frequentist properties of our approach from a theoretical perspective, and by relating it to frequentist methods for handling nuisance parameters through conditioning.

To derive these theoretical results, we suppose that the observed data are generated by the "reduced-form" version of the model of section 2, given by

$$x = z\pi + \nu \tag{5}$$

$$y = z\pi\beta + e, \tag{6}$$

where $\nu, e \in \mathbb{R}^T$ are the unobserved i.i.d. shocks. Equation (6) can be obtained by substituting (2) into (1). We assume that $u \equiv [\nu, e] \sim \mathcal{MN}(0, I_T, \Sigma)$, or, equivalently, vec $(u) \sim \mathcal{N}(0_{2T\times 1}, \Sigma \otimes I_T)$. We also impose that z has full column rank k, ruling out designs with $k > T$. In addition, we treat Σ as given, addressing the case of an unknown Σ only at the end of the section. This model representation and assumptions are identical to those in Andrews et al. (2006), and all results, lemmas and propositions stated below are derived under these assumptions unless otherwise noted.

To facilitate the exposition of our results, it is useful to consider the inferential problem of three distinct econometricians, each confronted with the same data. While all three believe that the observed data are generated by model (5)-(6), they differ in their assumptions about the underlying model parameters.

5.1 Econometrician B

Econometrician B—short for "Bayesian"—treats the coefficients π and β as realizations of random variables. Specifically, she assumes that π is drawn by nature from a $\mathcal{N}\left(0, \gamma^2\sigma_\nu^2 (z'z)^{-1}\right)$, with $\gamma^2|\beta \sim \mathcal{U}\left(0, M\left(\sigma_\nu^2 b'\Sigma^{-1}b\right)^{-1}\right)$ and $p(\beta) \propto \left(b'\Sigma^{-1}b\right)^{-1}$, where $b \equiv [1,\beta]'$. Here, $\mathcal{U}(\underline{s}, \overline{s})$ denotes the uniform distribution with support on the interval $[\underline{s}, \overline{s}]$, and M is a large positive constant. This joint distribution for (γ^2, β) is similar in spirit to the flat prior in section 3, but not identical. The upper bound for γ^2 depends on β, and the marginal distribution of β is Cauchy. These features will be useful when we later re-parameterize the model in polar coordinates, as they induce a uniform distribution over a rectangular support in those coordinates, greatly simplifying the proofs. In addition, unlike the improper uniform priors of section 3, the current specification is proper.

Under these assumptions, it can be shown that the statistic $\hat{\Gamma} \equiv C'w'z\left(z'z\right)^{-1}z'wC$ is sufficient for inference about β and γ^2, where $w \equiv [x, y]$ and C is any matrix such that $CC' = \Sigma^{-1}$. Moreover, the distribution of $\hat{\Gamma}$ is given by

$$\hat{\Gamma}|\beta, \gamma^2 \sim \mathcal{W}\left(I_2 + \gamma^2\sigma_\nu^2 C'bb'C, k\right),$$

where $\mathcal{W}(S, d)$ denotes the Wishart distribution with scale matrix S and degrees of freedom d.

Finally, to enable a clearer comparison of the inference conducted by our three econometricians (for reasons that will become evident shortly), it is useful to re-parameterize the model in polar coordinates, as in Chamberlain (2007):

$$\hat{\Gamma}|\theta, r \sim \mathcal{W}\left(I_2 + r\phi_\theta\phi_\theta', k\right), \tag{7}$$

where the parameter of interest is now θ, and r is the nuisance parameter. The one-to-one mapping from $\left(\gamma^2, \beta\right)$ to (r, θ) is given by $r = \gamma^2\sigma_\nu^2\left(b'CC'b\right)$ and $\theta = \arctan\left(\frac{[C'b]_2}{[C'b]_1}\right)$, where $[C'b]_j$ denotes the j^{th} element of the vector $C'b$. By choosing C to be upper triangular with positive diagonal entries, the mapping from β to θ becomes not only one-to-one but also monotonically increasing, which is convenient. By the standard change-of-variables formula, the specified distribution for $\left(\gamma^2, \beta\right)$ induces a uniform distribution for (r, θ) over the rectangular support $[0, M] \times \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$.

In sum, econometrician B assumes that the data-generating process (DGP) of $\hat{\Gamma}$ is given by (7), with θ and r randomly drawn from uniform distributions.

5.2 Econometrician RC

Econometrician RC—short for “random coefficients”—makes identical assumptions to B, except that she treats θ as a fixed but unknown parameter, rather than the realization of a random variable. *In sum, econometrician RC assumes that the DGP of $\hat{\Gamma}$ is given by (7), with r randomly drawn from a uniform distribution and $\theta = \theta_0$, where θ_0 is a fixed but unknown value.*

5.3 Econometrician FC

Econometrician FC—short for "fixed coefficients"—is endowed with the true DGP: She correctly believes that none of the coefficients are random, but they are all fixed and unknown. Under these assumptions, Andrews et al. (2006) show that

$$\hat{\Gamma}|\beta,\mu^2 \sim \mathcal{W}^n\left(I_2,k,\mu^2\sigma_\nu^2 C'bb'C\right),$$

where $\mathcal{W}^n(S,d,N)$ is the non-central Wishart distribution with scale matrix S, degrees of freedom d, and non-centrality matrix N. Switching again to polar coordinates, we obtain

$$\hat{\Gamma}|\theta,\rho \sim \mathcal{W}^n\left(I_2,k,\rho\phi_\theta\phi_\theta'\right), \tag{8}$$

where the parameter of interest is now θ, and ρ is the nuisance parameter. The one-to-one mapping from $\left(\mu^2,\beta\right)$ to (ρ,θ) is given by $\rho=\mu^2\sigma_\nu^2\left(b'CC'b\right)$ and $\theta=\arctan\left(\frac{[C'b]_2}{[C'b]_1}\right)$.

In sum, econometrician FC assumes that the DGP of $\hat{\Gamma}$ is given by (8), with $\theta=\theta_0$ and $\rho=\rho_0$, where θ_0 and ρ_0 are fixed but unknown values.

5.4 Additional notation and preliminary results

Before proceeding further, it is useful to establish some additional notation and preliminary results.

Define the eigenvalues of $\hat{\Gamma}$ by $\hat{\lambda}_1$ and $\hat{\lambda}_2$, with $\hat{\lambda}_1>\hat{\lambda}_2$. Let $\hat{v}$ denote the eigenvector of $\hat{\Gamma}$ associated with $\hat{\lambda}_1$, and define the angle of this unit-length vector by $\hat{\theta}=\arctan\left(\frac{\hat{v}_2}{\hat{v}_1}\right)$. By examining the likelihood function of $\hat{\Gamma}$ implied by (7) or (8), it is easy to show that $\hat{\theta}$ is the maximum likelihood estimator (MLE) of θ under the model of all three econometricians. Importantly, $\hat{\theta}$ does not depend on the nuisance parameters—namely, r under (7) or ρ under (8). This property implies that the parameters are orthogonal in the sense of Cox and Reid (1987), a fact that will play an important role later on.

With a slight abuse of notation, we do not explicitly distinguish between estimators and their realizations, using the same symbol for both. The context—particularly the conditioning set—will make clear whether we are referring to a sampling or a posterior density. Finally, when there is potential ambiguity, we will be explicit about whether a given density corresponds to the model of a specific econometrician.

5.5 Comparing inferential approaches

To study the frequentist properties of our Bayesian approach, we now compare the inferential strategies adopted by our three econometricians. The following lemma establishes a relationship between the inference conducted by econometricians B and RC.

Lemma 1. *Let* $p_B\left(\theta|\hat{\Gamma}\right)$ *denote the (posterior) distribution of the random parameter* θ *given the observed data, according to the model of econometrician* B. *Let* $p_{RC}\left(\hat{\theta}|\theta, \hat{\lambda}_1, \hat{\lambda}_2\right)$ *denote the sampling distribution of* $\hat{\theta}$ *conditional on* θ, $\hat{\lambda}_1$ *and* $\hat{\lambda}_2$, *based on the DGP of econometrician* RC. *Then,*

$$p_B\left(\theta|\hat{\Gamma}\right) \propto p_{RC}\left(\hat{\theta}|\theta, \hat{\lambda}_1, \hat{\lambda}_2\right)$$

and both densities depend only on a periodic and even function of $\theta - \hat{\theta}$.

Proof. See appendix C. □

Lemma 1 implies that, under the model of econometrician RC and conditional on the sample realization of the eigenvalues of $\hat{\Gamma}$, the sampling distribution of $\hat{\theta} - \theta$ given θ coincides with the posterior distribution of $\theta - \hat{\theta}$.[5] This result goes a long way toward providing a frequentist interpretation of Bayesian inference, but it does not go all the way for two reasons. First, the lemma relates the posterior to a sampling distribution obtained by integrating out nuisance parameters, thereby adopting a Bayesian treatment of those parameters. Second, the sampling distribution is constructed conditional on the observed eigenvalues, rather than treating them as random. We now address these two limitations by relating the inference conducted by econometricians RC to that of FC, who adopts a more strictly frequentist perspective. Specifically, the next proposition establishes that the asymptotic distribution of $\hat{\theta}$ coincides under the models of econometricians RC and FC, it is invariant to whether one conditions on $\hat{\lambda}_1$ and $\hat{\lambda}_2$, and it is Gaussian.

Proposition 2. *Suppose that there exists a constant* $\underline{c} > 0$ *such that* $\lim_{T,k\to\infty} \frac{\rho}{k} \geq \underline{c}$. *Then,*

$$\frac{\hat{\theta} - \theta}{\hat{S}} = R + O_p\left(\frac{1}{k}\right) \quad \text{as} \quad T, k \to \infty,$$

[5] Because both the posterior of θ and the sampling distribution of $\hat{\theta}$ are supported on the finite interval $[-\pi/2, \pi/2]$, the distributions of $\left(\theta - \hat{\theta}\right)|\hat{\theta}$ and $\left(\hat{\theta} - \theta\right)|\theta$ need not be symmetric around zero simply because their support is not centered there. However, both densities are periodic with period π. Hence, their support can be shifted arbitrarily and re-centered at zero, at which point the two densities coincide. This reflects the familiar "wrapping argument" from circular statistics: A variable defined on the real line is mapped onto a circle of circumference π by treating any two values that differ by an integer multiple of π as the same point. In essence, the posterior of the wrapped difference $(\theta - \hat{\theta}) \bmod \pi$ and the sampling distribution of the wrapped difference $(\theta - \hat{\theta}) \bmod \pi$ coincide.

where $\hat{S} = \frac{\sqrt{\hat{\lambda}_1}}{\hat{\lambda}_1 - \hat{\lambda}_2}$ and $R \sim \mathcal{N}(0,1)$ is a standard Normal random variable. This result holds both conditional on the sample realization of $\hat{\lambda}_1$ and $\hat{\lambda}_2$ and unconditionally, under the models of both econometrician RC and FC.

Proof. See appendix C. □

A few remarks help contextualize and interpret this proposition.

Remark 3. Proposition 2 states that the sampling distribution of the Wald-type statistic $\frac{\hat{\theta}-\theta}{\hat{S}}$ is nearly Gaussian, regardless of whether the nuisance parameters are treated as random (as done by econometrician RC) or fixed (as done by econometrician FC), and regardless of whether inference is conditional on the sample realization of the eigenvalues of $\hat{\Gamma}$. Together with Lemma 1, these results imply that the sampling distribution of $\frac{\hat{\theta}-\theta}{\hat{S}}$ and the posterior distribution of $\frac{\theta-\hat{\theta}}{\hat{S}}$ are asymptotically equivalent. In short, Bayesian inference about θ aligns with frequentist inference.

Remark 4. Given the Gaussian approximation, inference on θ is simple: An $(1-\alpha)$–level confidence (or credible) interval for θ takes the familiar form $\left[\hat{\theta} \pm q_{1-\alpha/2}\hat{S}\right]$, where $q_{1-\alpha/2}$ denotes the $(1-\alpha/2)$ quantile of the standard normal distribution.[6]

Remark 5. The finding that reliable inference can be conducted with a simple Gaussian approximation stands in sharp contrast to the extensive literature on weak instruments, which is built on the premise that conventional asymptotics may perform poorly. To understand this apparent discrepancy, note that Proposition 2 pertains to the Wald statistic for the model reparameterized in polar coordinates. In contrast, the distribution of $\hat{\beta}$ can be substantially skewed and does not exhibit a near-Gaussian shape, unless the number of instruments is very large. Put differently, working in polar coordinates delivers an exact third-order refinement.

Remark 6. The statistic $\frac{\hat{\theta}-\theta}{\hat{S}}$ in Proposition 2 is pivotal—that is, its distribution does not depend on any unknown parameters, including the nuisance parameters. This property ensures that the corresponding test maintains correct size and is robust to weak instruments, and that the associated confidence intervals attain nominal coverage. As shown in appendix C, the nuisance parameter effectively vanishes from the limiting distribution because $\hat{\lambda}_1$, the largest eigenvalue of $\hat{\Gamma}$, provides a very accurate approximation of a sufficient statistic for the nuisance parameter. As elaborated further in Remark 7, this result follows from parameter orthogonality.

[6] Because the likelihood and posterior densities are periodic in θ with period π, any portion of this interval that lies outside the fundamental support of θ can be wrapped onto the circle of circumference π. in this case—which is rare in practice, since $\hat{S}$ converges in probability to zero—the confidence (or credible) interval may appear as two disjoint arcs (see also Mikusheva, 2010).

Remark 7. Proposition 2 is derived under a "many-weak-instruments" asymptotic framework, in which k grows large with the sample size T. This framework has been widely used in the weak-IV literature to derive tractable approximations, including in the seminal work of Staiger and Stock (1997) and Stock and Yogo (2005). What is distinctive here is that the Gaussian approximation is more accurate than in typical settings. In standard cases, the error for approximating the distribution of the Wald statistic with a Gaussian is of order $O_p\left(1/\sqrt{k}\right)$, whereas in our setting it is of order $O_p\left(1/k\right)$. As a consequence, our asymptotic framework offers a good approximation to finite-sample settings not only with many instruments, but also with a modest number of them. The proof in appendix C clarifies that the accelerated convergence of $\frac{\hat{\theta}-\theta}{\hat{S}}$ to a standard Normal random variable is due to two features of the model: (i) the approximate symmetry of the distribution of $\hat{\theta}$ around θ, which enhances the accuracy of the Gaussian approximation; and (ii) the fact that the largest eigenvalue of $\hat{\Gamma}$ ($\hat{\lambda}_1$) approximates very accurately a sufficient statistic for the nuisance parameter ($\phi_\theta'\hat{\Gamma}\phi_\theta$), thereby effectively removing this nuisance parameter from the asymptotic distribution. While our proof of property (ii) in appendix C is model specific, it stems more generally from the orthogonality between the parameter of interest and the nuisance parameter, in the sense that the Fisher information matrix is diagonal. To illustrate the importance of parameter orthogonality, consider the model of econometrician RC. As shown by Cox and Reid (1987), a direct implication of parameter orthogonality is that $\hat{r}_\theta = \hat{r} + O_p\left(\frac{1}{k}\right)$, where $\hat{r}$ is the MLE of r, and $\hat{r}_\theta$ is the MLE of r given θ. Since $\hat{r} = \frac{\hat{\lambda}_1}{k} - 1$ and $\hat{r}_\theta = \frac{\phi_\theta'\hat{\Gamma}\phi_\theta}{k} - 1$, it follows that $\hat{\lambda}_1 = \phi_\theta'\hat{\Gamma}\phi_\theta + O_p\left(\frac{1}{k}\right)$. A similar argument applies in the model of econometrician FC, though the function linking $\hat{\lambda}_1$ to $\hat{\rho}$ and $\phi_\theta'\hat{\Gamma}\phi_\theta$ to $\hat{\rho}_\theta$ is more complex.

Remark 8. The proof of Proposition 2 relies on the assumption that $\lim_{T,k\to\infty}\frac{\rho}{k} \geq \underline{c} > 0$, ensuring that a non-negligible fraction of instruments remain relevant as their number grows. This condition does not require any individual instrument to be strong; it merely rules out degenerate cases in which researchers add an increasing number of purely irrelevant instruments, causing the fraction of relevant ones to vanish and their weak signal to be entirely diluted in noise. In other words, the assumption prevents the model from approaching the non-identification boundary, where the average concentration parameter collapses to zero. As a result, Dufour (1997) impossibility theorem does not apply, and valid inference does not require unbounded confidence sets with positive probability. In practice, this requirement is mild—especially given that the Gaussian approximation in Proposition 2 is already accurate with a moderate number of instruments. Therefore, researchers need not—and should not—cast an excessively wide net in search of instruments. A focused selection of moderately sized instrument sets avoids the risk of accumulating predominantly irrelevant variables that would weaken identification.

Remark 9. Proposition 2 no longer applies when the condition $\lim_{T,k\to\infty} \frac{\rho}{k} \geq \underline{c} > 0$ is violated. In that case, the simulation results in figures 1 and 2a–13a, as well as the estimation of the AK model with fake instruments reported in table 1, indicate that inference becomes conservative, with 95-percent credible intervals containing the true value of the parameter more than 95 percent of the time. In the case of perfectly irrelevant instruments and a known residual covariance matrix, it can be shown that the coverage of our credible intervals depends on the correlation of the structural errors. The intervals remain conservative for correlations as high as 0.935, and even for correlations as large as 0.975 the size distortion does not exceed 5 percent (see appendix C.5).

Remark 10. Proposition 2 and the remarks above concern the asymptotic equivalence between Bayesian and frequentist inference about θ. But do these results translate to inference about β, the primitive parameter? The answer is yes. Because β is a monotone transformation of θ, equal-tailed credible intervals for θ map directly into equal-tailed credible intervals for β. Likewise, the frequentist confidence interval—which coincides with the equal-tailed credible interval in the θ space—can be converted into a confidence interval for β by applying the same monotone mapping to its endpoints. The monotonicity of the mapping and the use of a pivotal statistic ensure that the resulting interval is valid for β as well. Thus, asymptotically, Bayesian inference for β inherits the same desirable frequentist properties as the inference for θ, consistent with our simulation and empirical results in sections 3 and 4.

In sum, Proposition 2 shows that Bayesian inference, when interpreted from a frequentist perspective, achieves correct size and is robust to weak instruments. But is it a powerful approach to inference? Our next proposition explores its connection with the conditional inference strategy of Moreira (2003), which has been shown to achieve maximum power in a class of invariant tests (Andrews et al., 2006).

Proposition 3. *Let $LR = \hat{\lambda}_1 - \phi_\theta' \hat{\Gamma} \phi_\theta$ denote the log-likelihood-ratio statistic of Moreira (2003), obtained as the logarithm of the ratio between the profile likelihoods under the unrestricted and restricted models. Let $\hat{W} = \frac{\hat{\theta}-\theta}{\hat{S}}$ be pivotal statistic defined in Proposition 2. Then, under the same assumption of Proposition 2,*

$$LR = \hat{W}^2 \cdot \zeta + O_p\left(\frac{1}{k}\right) \quad \text{as} \quad T, k \to \infty,$$

where

$$\zeta = \frac{\phi_\theta' \hat{\Gamma} \phi_\theta}{\phi_\theta' \hat{\Gamma} \phi_\theta - J} = \textit{constant} + O_p\left(\frac{1}{\sqrt{k}}\right) \quad \text{as} \quad T, k \to \infty,$$

and J is ancillary with respect to θ and asymptotically independent from $\hat{W}$. These results hold under the models of both econometrician RC and FC.

Proof. See appendix C. □

The following observations serve to clarify and interpret the key aspects of this proposition.

Remark 11. Proposition 3 shows that the likelihood-ratio test statistic of Moreira (2003) is approximately equal to the square of the pivotal statistic from Proposition 2, $\hat{W}^2$, multiplied by a random scalar ζ. Since ζ converges in probability to a constant, inference based on the likelihood-ratio statistic and on $\hat{W}$ is asymptotically equivalent. And since the mapping between θ and β is monotone, this equivalence holds also in the β space.

Remark 12. Moreira's conditional likelihood-ratio test is based on the distribution of LR, conditional on the sample realization of $\phi_\theta' \hat{\Gamma} \phi_\theta$. As a result, the only source of randomness in ζ is the random variable J. Importantly, J is ancillary with respect to θ—thus providing no information about the parameter of interest—and asymptotically independent of $\hat{W}$. Consequently, the LR statistic incorporates an additional source of randomness relative to $\hat{W}$, leading to greater variance. This result implies that, although LR and $\hat{W}^2$ are asymptotically equivalent, inference based on $\hat{W}$ is generally sharper than that based on LR.

To summarize, the combination of Lemma 1 and Propositions 2 and 3 show that Bayesian inference—about either θ or β—has desirable frequentist properties and robustness to weak instruments, and yields results similar to those based on Moreira's conditional inference, consistent with our simulation evidence.

Up to this point, we have worked under the assumption that the covariance matrix of the reduced-form residuals is known. This simplification is justified only if the error in estimating Σ is small relative to the overall estimation uncertainty and the approximation errors of Propositions 2 and 3. The following lemma shows that this is indeed the case, provided that k^2 is of the same order of T.

Lemma 2. *Let $\hat{\Sigma} = \frac{1}{T-k}\left(w - z\hat{\Pi}\right)'\left(w - z\hat{\Pi}\right)$, with $\hat{\Pi} = \left(z'z\right)^{-1} z'w$. Let $\tilde{\Gamma} = \hat{C}'w'z\left(z'z\right)^{-1} z'w\hat{C}$, where $\hat{C} = g\left(\hat{\Sigma}\right)$ and $g\left(\cdot\right)$ is the same function that maps Σ into C. Then, under the assumptions of Proposition 2 and if $\frac{k^2}{T} = O\left(1\right)$,*

$$\frac{1}{k}\tilde{\Gamma} = \frac{1}{k}\hat{\Gamma} + O_p\left(\frac{1}{k}\right) \quad \text{as} \quad T, k \to \infty.$$

This result holds under the model of econometrician FC (i.e. the true DGP).

Proof. See appendix C. □

Lemma 2 shows that replacing C with $\hat{C}$—and hence $\hat{\Gamma}$ with $\tilde{\Gamma}$—introduces only an error of order $\frac{1}{k}$, as long as $k \lesssim \sqrt{T}$. The eigenvalues of symmetric matrices are continuous and differentiable functions of the matrix entries, and the same holds for the eigenvectors when the eigenvalues are well separated—a condition ensured by the assumptions of Proposition 2. It follows that substituting $\tilde{\Gamma}$ for $\hat{\Gamma}$ does not alter the asymptotic distribution of the components of the spectral decomposition. Therefore, all results in Propositions 2 and 3 remain valid, as the difference between $\hat{\Gamma}$ and $\tilde{\Gamma}$ is absorbed within the existing approximation error of order $\frac{1}{k}$. The requirement $k \lesssim \sqrt{T}$ is stronger than the condition established by Andrews and Stock (2007b) as necessary for correct asymptotic size, $k \lesssim T^{2/3}$. However, our $\frac{1}{k}$ convergence rate is faster than their standard $\frac{1}{\sqrt{k}}$. Moreover, Andrews and Stock (2007a) show that their results continue to hold under non-Gaussian errors, provided that the stronger condition $k \lesssim T^{1/3}$ is satisfied.

6 Concluding remarks

In regular statistical models—where the likelihood is well behaved and standard asymptotic approximations apply—classical and Bayesian procedures yield similar conclusions, at least asymptotically, as formalized by the Bernstein–von Mises theorem. This alignment is reassuring: Bayesian inference enjoys desirable frequentist properties, classical inference admits a Bayesian interpretation, and researchers with different philosophical views are nonetheless led to the same empirical findings. For a discussion of the advantages of this equivalence from an a theoretical perspective, see Müller and Norets (2016a,b) and the references therein.

IV regressions with weak instruments, however, are a prominent example of *non-regular* statistical models, in which this near equivalence breaks down and naive inference becomes pathological under both paradigms. In such settings, the presence of a non-vanishing nuisance parameter undermines conventional asymptotic approximations to the distribution of standard IV estimators—such as TSLS—and similarly compromises Bayesian inference based on diffuse or flat priors. Substantial progress has been made in addressing this problem from a frequentist perspective. In particular, Moreira (2003) develops an inferential approach that delivers valid tests by conditioning on a statistic sufficient for the nuisance parameter, and Andrews et al. (2006) show that this procedure is optimal within a broad class of invariant, similar tests.

In this paper, we identify the source of the pathology for Bayesian inference in this class of IV models: Standard diffuse priors inadvertently concentrate mass on regions of the parameter space where the instruments are strong. We then construct a corrected prior that removes this distortion, and show that the resulting Bayesian credible sets are asymptotically equivalent to Moreira's confidence sets. Taken together, these results help restore the concordance between classical and Bayesian inference in the non-regular setting of IV regressions with weak instruments.

A Proof of Proposition 1

This appendix proves each of the three statements that constitute Proposition 1.

A.1 Flat prior on the concentration parameter

We begin by proving the first part of proposition 1, namely that

$$p\left(\pi|\sigma_{\nu}^{2}\right)\propto\left(\frac{\pi'z'z\pi}{\sigma_{\nu}^{2}}\right)^{-\frac{k-2}{2}}\left(\frac{1}{\sigma_{\nu}^{2}}\right)^{\frac{k}{2}} \tag{9}$$

induces a flat prior on the concentration parameter. First, observe that the concentration parameter can be written as

$$\mu^{2}=\frac{\pi'z'z\pi}{\sigma_{\nu}^{2}}=\tilde{\pi}'\tilde{\pi},$$

where $\tilde{\pi}\equiv\frac{1}{\sqrt{\sigma_{\nu}^{2}}}\left(z'z\right)^{\frac{1}{2}}\pi$. Expressing $\tilde{\pi}$ in polar coordinates yields $\tilde{\pi}=q\cdot\varphi(\eta)$, where $q=\|\tilde{\pi}\|=\sqrt{\mu^{2}}$ represents the length (or norm) of the vector, and $\varphi\left(\eta\right)\in S^{k-1}$ is a unit vector that captures its direction on the $(k-1)$-dimensional unit sphere, parameterized by the angular coordinates η. By the change-of-variables formula, the joint density of q and η satisfies

$$p\left(q,\eta|\sigma_{\nu}^{2}\right)=p\left(\pi\left(q,\eta\right)|\sigma_{\nu}^{2}\right)\cdot\left|\frac{\partial\pi\left(q,\eta\right)}{\partial\left(q,\eta\right)}\right|, \tag{10}$$

where the second term on the right-hand side is the determinant of the Jacobian of the transformation from (q,η) to π. This Jacobian can be characterized as

$$\left|\frac{\partial\pi\left(q,\eta\right)}{\partial\left(q,\eta\right)}\right|=\left|\sqrt{\sigma_{\nu}^{2}}\left(z'z\right)^{-\frac{1}{2}}\frac{\partial\tilde{\pi}\left(q,\eta\right)}{\partial\left(q,\eta\right)}\right|$$

$$\propto \left(\sigma_\nu^2\right)^{\frac{k}{2}} \left|\frac{\partial \tilde{\pi}\left(q,\eta\right)}{\partial\left(q,\eta\right)}\right|$$

$$\propto \left(\sigma_\nu^2\right)^{\frac{k}{2}} q^{k-1} \left(\prod_{j=1}^{k-2} \sin^{k-j-1} \eta_j\right), \tag{11}$$

where the last line follows from the fact that $\left|\frac{\partial \tilde{\pi}(q,\eta)}{\partial(q,\eta)}\right|$ is the volume element associated with the transformation from Cartesian to spherical coordinates in $\mathbb{R}^k$, capturing how volume scales under the change of variables. Substituting (9) and (11) into (10), we obtain

$$p\left(q,\eta|\sigma_\nu^2\right) \propto q^{2-k}\left(\frac{1}{\sigma_\nu^2}\right)^{\frac{k}{2}} \cdot \left(\sigma_\nu^2\right)^{\frac{k}{2}} q^{k-1} \left(\prod_{j=1}^{k-2} \sin^{k-j-1} \eta_j\right)$$

$$\propto q \left(\prod_{j=1}^{k-2} \sin^{k-j-1} \eta_j\right),$$

which implies that

$$p\left(q|\sigma_\nu^2\right) \propto q.$$

Finally, applying the change-of-variables formula once more to derive the implied prior on μ^2, we have

$$p\left(\mu^2|\sigma_\nu^2\right) = p\left(q\left(\mu^2\right)|\sigma_\nu^2\right)\left|\frac{\partial q\left(\mu^2\right)}{\partial \mu^2}\right| \propto 1,$$

which shows that the implied prior on the concentration parameter is indeed flat.

A.2 Equivalence to hierarchical prior

The second part of proposition 1 states that, when $k > 2$, (9) is equivalent to a hierarchical specification that combines (4) with a flat hyperprior on γ^2. To show this result, observe that (4) implies

$$p\left(\pi|\gamma^2,\sigma_\nu^2\right) \propto \left|\gamma^2\sigma_\nu^2\left(z'z\right)^{-1}\right|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\pi'\left[\gamma^2\sigma_\nu^2\left(z'z\right)^{-1}\right]^{-1}\pi\right\}$$

$$\propto \left(\sigma_\nu^2\right)^{-\frac{k}{2}}\left(\gamma^2\right)^{-\frac{k}{2}} \exp\left\{-\frac{1}{2}\frac{1}{\gamma^2}\frac{\pi'z'z\pi}{\sigma_\nu^2}\right\}.$$

To compute the marginal density of π, we integrate out γ^2, using a flat hyperprior. This integration yields

$$p\left(\pi|\sigma_\nu^2\right) \propto \int \left(\sigma_\nu^2\right)^{-\frac{k}{2}}\left(\gamma^2\right)^{-\frac{k}{2}} \exp\left\{-\frac{1}{2}\frac{1}{\gamma^2}\frac{\pi'z'z\pi}{\sigma_\nu^2}\right\} d\gamma^2$$

$$\propto \left(\frac{\pi' z' z\pi}{\sigma_\nu^2}\right)^{-\frac{k-2}{2}} \left(\frac{1}{\sigma_\nu^2}\right)^{\frac{k}{2}},$$

where the final proportionality follows from the fact that, when $k > 2$, the integrand is the kernel of an Inverse-Gamma density in γ^2.

A.3 Proper posterior

To prove the last part of proposition 1, we need to demonstrate that the resulting posterior distribution is always integrable. The posterior density is given by the product of the prior (9) and the likelihood of model (1)-(2):

$$p\left(\beta, \delta, \pi, \sigma_\nu^2, \sigma_\varepsilon^2 | y, x, z\right) \propto$$

$$\propto \left(\frac{1}{\sigma_\varepsilon^2}\right)^{\frac{T}{2}} e^{-\frac{1}{2\sigma_\varepsilon^2}(y-x\beta-(x-z\pi)\delta)'(y-x\beta-(x-z\pi)\delta)} \left(\frac{1}{\sigma_\nu^2}\right)^{\frac{T}{2}} e^{-\frac{1}{2\sigma_\nu^2}(x-z\pi)'(x-z\pi)} \cdot \left(\frac{\pi' z' z\pi}{\sigma_\nu^2}\right)^{-\frac{k-2}{2}} \left(\frac{1}{\sigma_\nu^2}\right)^{\frac{k}{2}}$$

$$\propto \left(\frac{1}{\sigma_\varepsilon^2}\right)^{\frac{T}{2}} e^{-\frac{1}{2\sigma_\varepsilon^2}(y-XB)'(y-XB)} \left(\frac{1}{\sigma_\nu^2}\right)^{\frac{T+2}{2}} e^{-\frac{1}{2\sigma_\nu^2}(x-z\pi)'(x-z\pi)} \cdot \left(\pi' z' z\pi\right)^{-\frac{k-2}{2}},$$

where $X \equiv [x, x - z\pi]$ and $B \equiv [\beta, \delta]'$. We can now integrate out β and δ, obtaining

$$p\left(\pi, \sigma_\nu^2, \sigma_\varepsilon^2 | y, x, z\right) = \int\int p\left(\beta, \delta, \pi, \sigma_\nu^2, \sigma_\varepsilon^2 | y, x, z\right) d\beta d\delta$$

$$\propto \left(\pi' z' z\pi\right)^{-\frac{k-2}{2}} \left(\frac{1}{\sigma_\nu^2}\right)^{\frac{T+2}{2}} e^{-\frac{1}{2\sigma_\nu^2}(x-z\pi)'(x-z\pi)} \left(\frac{1}{\sigma_\varepsilon^2}\right)^{\frac{T}{2}} \int e^{-\frac{1}{2\sigma_\varepsilon^2}\left[(y-XB)'(y-XB)\right]} dB$$

$$\propto \left(\pi' z' z\pi\right)^{-\frac{k-2}{2}} \left(\frac{1}{\sigma_\nu^2}\right)^{\frac{T+2}{2}} e^{-\frac{1}{2\sigma_\nu^2}(x-z\pi)'(x-z\pi)} \left(\frac{1}{\sigma_\varepsilon^2}\right)^{\frac{T-2}{2}} e^{-\frac{1}{2\sigma_\varepsilon^2}\left(y-X\hat{B}\right)'\left(y-X\hat{B}\right)} \left|X'X\right|^{-\frac{1}{2}},$$

where $\hat{B} \equiv \left(X'X\right)^{-1} X'y$ and the last expression follows from the fact that the integrand is the kernel of a Gaussian density. We are now ready to integrate out σ_ε^2 and σ_ν^2, which yields

$$p\left(\pi | y, x, z\right) = \int\int p\left(\pi, \sigma_\nu^2, \sigma_\varepsilon^2 | y, x, z\right) d\sigma_\varepsilon^2 d\sigma_\nu^2$$

$$\propto \left(\pi' z' z\pi\right)^{-\frac{k-2}{2}} \left|X'X\right|^{-\frac{1}{2}} \int \left(\frac{1}{\sigma_\nu^2}\right)^{\frac{T+2}{2}} e^{-\frac{1}{2\sigma_\nu^2}(x-z\pi)'(x-z\pi)} d\sigma_\nu^2 \int \left(\frac{1}{\sigma_\varepsilon^2}\right)^{\frac{T-2}{2}} e^{-\frac{1}{2\sigma_\varepsilon^2}\left(y-X\hat{B}\right)'\left(y-X\hat{B}\right)} d\sigma_\varepsilon^2$$

$$\propto \left(\pi' z' z\pi\right)^{-\frac{k-2}{2}} \left|X'X\right|^{-\frac{1}{2}} \left[(x-z\pi)'(x-z\pi)\right]^{-\frac{T}{2}} \left[\left(y - X\hat{B}\right)' \left(y - X\hat{B}\right)\right]^{-\frac{T-4}{2}}, \tag{12}$$

where the last expression follows from the fact that the integrands are kernels of Inverse-Gamma densities.

The marginal posterior of π can be further simplified by observing that

$$\left|X'X\right| = \left|[x, x - z\pi]' \, [x, x - z\pi]\right| = \left(x'x\right) \left[(x - z\pi)' \, M_x \, (x - z\pi)\right] = \left(x'x\right) \left(\pi' z' M_x z\pi\right), \qquad (13)$$

where $M_A \equiv I - A\left(A'A\right)^{-1} A'$ denotes the residual-maker matrix, and we have used the property $|a'ab'b - a'bb'a| = |a'a| \, |b'M_a b| = |b'b| \, |a'M_b a|$, as in Hoogerheide et al. (2005). Using the same property, as well as the Frisch–Waugh–Lovell theorem, we can also show that and

$$\begin{aligned} \left(y - X\hat{B}\right)' \left(y - X\hat{B}\right) &= y' M_X y = \left(M_x y\right)' M_{M_x(x - z\pi)} \left(M_x y\right) \\ &= \frac{y' M_x y}{(x - z\pi)' \, M_x \, (x - z\pi)} \, (x - z\pi)' \, M_x M_{M_x y} M_x \, (x - z\pi) \\ &= \frac{y' M_x y}{\pi' z' M_x z\pi} \left(\pi' z' M_{[y,x]} z\pi\right). \qquad (14) \end{aligned}$$

Substituting (13) and (14) into (12), we obtain

$$p\left(\pi | y, x, z\right) \propto \left(\pi' z' z\pi\right)^{-\frac{k-2}{2}} \left(\pi' z' M_x z\pi\right)^{-\frac{1}{2}} \left[(x - z\pi)' \, (x - z\pi)\right]^{-\frac{T}{2}} \left(\frac{\pi' z' M_{[y,x]} z\pi}{\pi' z' M_x z\pi}\right)^{-\frac{T-4}{2}},$$

This is the density of a proper distribution. Its tails exhibit polynomial decay and are sufficiently thin to ensure integrability as long as $T > 1$. If $k > 1$, the posterior distribution has a singularity at $\pi = 0$, but the integral around $\pi = 0$ always remains finite. Since $\pi = 0$ is a set with measure zero, the Markov chains generated by Markov Chain Monte Carlo algorithms do not violate the requirement of irreducibility, ensuring the validity of these algorithms (see definition 6.13 in Robert and Casella, 2004).

B Algorithm for posterior inference

This appendix provides the details of the Markov Chain Monte Carlo algorithm that we use to evaluate the posterior distribution of the IV model

$$y = x\beta + c\alpha + \nu\delta + \varepsilon \qquad (15)$$

$$x = z\pi + c\rho + \nu, \qquad (16)$$

where $y \in \mathbb{R}^T$ is an observed dependent variable, $x \in \mathbb{R}^T$ is an observed regressor, $c \in \mathbb{R}^{T \times l}$ are observed control variables, $z \in \mathbb{R}^{T \times k}$ are observed instrumental variables, $\nu \sim \mathcal{N}\left(0_{T \times 1}, \sigma_\nu^2 I_T\right)$

and $\nu \sim \mathcal{N}\left(0_{T\times 1}, \sigma_\varepsilon^2 I_T\right)$ are unobserved shocks uncorrelated with each other, and $\beta \in \mathbb{R}$, $\alpha \in \mathbb{R}^l$, $\delta \in \mathbb{R}$, $\pi \in \mathbb{R}^k$, $\rho \in \mathbb{R}^l$, $\sigma_\nu^2 \in \mathbb{R}$ and $\sigma_\varepsilon^2 \in \mathbb{R}$ are unknown parameters. Relative to the model in section 2, equations (15) and (16) include additional controls. As specified in proposition 1, we impose the prior distribution

$$p\left(\pi|\sigma_\nu^2\right) \propto \left(\frac{\pi' z' z \pi}{\sigma_\nu^2}\right)^{-\frac{k-2}{2}} \left(\frac{1}{\sigma_\nu^2}\right)^{\frac{k}{2}}, \tag{17}$$

and flat priors on all remaining model parameters. The remainder of this appendix details the Markov Chain Monte Carlo algorithm used for posterior inference. We begin with the case of $k > 2$, followed by the case where $k = 1$ or $k = 2$.

B.1 The case of $k > 2$

When $k > 2$, the prior (17) is equivalent to

$$\pi|\gamma^2, \sigma_\nu^2 \sim N\left(0, \gamma^2 \sigma_\nu^2 \Omega\right),$$

with a flat hyperprior on γ^2 and $\Omega = (z'z)^{-1}$. The algorithm described in this appendix, however, is also valid for other possible choices of Ω, including $\Omega = I_k$. The latter may be a suitable option when the number of instruments, k, is large relative to the number of observations, T. In both our simulations and empirical application, the results with $\Omega = (z'z)^{-1}$ or I_k are nearly indistinguishable.

The posterior distribution is given by

$$p\left(\beta, \alpha, \delta, \pi, \rho, \sigma_\nu^2, \sigma_\varepsilon^2, \gamma^2 | y, x, z, c\right) \propto p\left(y, x | z, c, \beta, \alpha, \delta, \pi, \rho, \sigma_\nu^2, \sigma_\varepsilon^2, \gamma^2\right) \cdot p\left(\pi|\gamma^2, \sigma_\nu^2\right)$$

$$\propto p\left(y|x, z, c, \beta, \alpha, \delta, \pi, \rho, \sigma_\nu^2, \sigma_\varepsilon^2, \gamma^2\right) \cdot p\left(x|z, c, \beta, \alpha, \delta, \pi, \rho, \sigma_\nu^2, \sigma_\varepsilon^2, \gamma^2\right) \cdot p\left(\pi|\gamma^2, \sigma_\nu^2\right)$$

$$\propto \left(\frac{1}{\sigma_\varepsilon^2}\right)^{\frac{T}{2}} e^{-\frac{1}{2\sigma_\varepsilon^2}[y - x\beta - c\alpha - (x - z\pi - c\rho)\delta]'[y - x\beta - c\alpha - (x - z\pi - c\rho)\delta]} \left(\frac{1}{\sigma_\nu^2}\right)^{\frac{T}{2}} e^{-\frac{1}{2\sigma_\nu^2}(x - z\pi - c\rho)'(x - z\pi - c\rho)} \cdot$$

$$\cdot \left(\frac{1}{\gamma^2 \sigma_\nu^2}\right)^{\frac{k}{2}} e^{-\frac{\pi'\Omega^{-1}\pi}{2\gamma^2\sigma_\nu^2}},$$

and we can sample from it using the following Gibbs sampling with blocks (i) $\left(\beta, \alpha, \delta, \sigma_\varepsilon^2\right)$, (ii) σ_ν^2, (iii) (π, ρ) and (iv) γ^2:

(i) The conditional posterior of $\left(\beta, \alpha, \delta, \sigma_\varepsilon^2\right)$ is given by

$$p\left(\beta, \alpha, \delta, \sigma_\varepsilon^2 | y, x, z, c, \pi, \rho, \sigma_\nu^2, \gamma^2\right) \propto \left(\frac{1}{\sigma_\varepsilon^2}\right)^{\frac{T}{2}} e^{-\frac{1}{2\sigma_\varepsilon^2}[y-x\beta-c\alpha-(x-z\pi-c\rho)\delta]'[y-x\beta-c\alpha-(x-z\pi-c\rho)\delta]},$$

which is the kernel of

$$\sigma_\varepsilon^2 | y, x, z, c, \pi, \rho, \sigma_\nu^2, \gamma^2 \sim IG(\frac{\hat{S}}{2}, \frac{T-l-4}{2})$$

$$\beta, \alpha, \delta | y, x, z, c, \pi, \rho, \sigma_\nu^2, \gamma^2, \sigma_\varepsilon^2 \sim N\left(\hat{\xi}, \left(\tilde{x}'\tilde{x}\right)^{-1} \sigma_\varepsilon^2\right),$$

where $\tilde{x} \equiv [x, c, (x - z\pi - c\rho)]$, $\hat{\xi} = \left(\tilde{x}'\tilde{x}\right)^{-1} \tilde{x}'y$ and $\hat{S} = \left(y - \tilde{x}\hat{\xi}\right)' \left(y - \tilde{x}\hat{\xi}\right)$.[7]

(ii) The conditional posterior of σ_ν^2 is given by

$$p\left(\sigma_\nu^2 | y, x, z, c, \beta, \alpha, \delta, \pi, \rho, \sigma_\varepsilon^2, \gamma^2\right) \propto \left(\frac{1}{\sigma_\nu^2}\right)^{\frac{T+k}{2}} e^{-\frac{1}{2\sigma_\nu^2}\left[(x-z\pi-c\rho)'(x-z\pi-c\rho)+\frac{\pi'\Omega^{-1}\pi}{\gamma^2}\right]},$$

which is the kernel of

$$\sigma_\nu^2 | y, x, z, c, \beta, \alpha, \delta, \pi, \rho, \sigma_\varepsilon^2, \gamma^2 \sim IG\left(\frac{(x - z\pi - c\rho)'(x - z\pi - c\rho)}{2} + \frac{\pi'\Omega^{-1}\pi}{2\gamma^2}, \frac{T+k-2}{2}\right).$$

(iii) The conditional posterior of $\omega \equiv [\pi, \rho]'$ is given by

$$p\left(\omega | y, x, z, c, \beta, \alpha, \delta, \sigma_\nu^2, \sigma_\varepsilon^2, \gamma^2\right) \propto$$

$$\propto e^{-\frac{1}{2\sigma_\varepsilon^2}[y-x\beta-c\alpha-(x-\tilde{z}\omega)\delta]'[y-x\beta-c\alpha-(x-\tilde{z}\omega)\delta]} e^{-\frac{1}{2\sigma_\nu^2}\left[(x-\tilde{z}\omega)'(x-\tilde{z}\omega)+\frac{\omega'P\omega}{\gamma^2}\right]}$$

$$\propto e^{-\frac{1}{2\sigma_\varepsilon^2}(\tilde{y}+\tilde{z}\delta\omega)'(\tilde{y}+\tilde{z}\delta\omega)-\frac{1}{2\sigma_\nu^2}\left[(x-\tilde{z}\omega)'(x-\tilde{z}\omega)+\frac{\omega'P\omega}{\gamma^2}\right]},$$

where $\tilde{z} \equiv [z, c]$, $P \equiv \begin{bmatrix} \Omega^{-1} & 0_{k\times l} \\ 0_{l\times k} & 0_{l\times l} \end{bmatrix}$ and $\tilde{y} \equiv y - x\left(\beta + \delta\right) - c\alpha$. The last expression can be manipulated to show that it is proportional to

$$e^{-\frac{1}{2\sigma_\nu^2}\left[(x-\tilde{z}\omega)'(x-\tilde{z}\omega)+\frac{\sigma_\nu^2}{\sigma_\varepsilon^2}(\tilde{y}+\tilde{z}\delta\omega)'(\tilde{y}+\tilde{z}\delta\omega)+\frac{\omega'P\omega}{\gamma^2}\right]}$$

$$\propto e^{-\frac{1}{2\sigma_\nu^2}\left\{\omega'\left[\left(1+\frac{\sigma_\nu^2}{\sigma_\varepsilon^2}\delta^2\right)\tilde{z}'\tilde{z}+\frac{1}{\gamma^2}P\right]\omega-2\left(x-\frac{\sigma_\nu^2}{\sigma_\varepsilon^2}\delta\tilde{y}\right)'\tilde{z}\omega\right\}},$$

[7] If a random variable, v, has an Inverse-Gamma distribution with scale parameter s and degrees of freedom d, $v \sim IG\left(s, d\right)$, then its pdf is $p\left(v|s, d\right) = \Gamma\left(d\right)^{-1} s^d v^{-d-1} \exp\left\{-\frac{s}{v}\right\}$.

which is the kernel of

$$\omega|y,x,z,c,\beta,\alpha,\delta,\sigma_\nu^2,\sigma_\varepsilon^2,\gamma^2 \sim N\left(\hat{\omega},\sigma_\nu^2\left[\left(1+\frac{\sigma_\nu^2}{\sigma_\varepsilon^2}\delta^2\right)\tilde{z}'\tilde{z}+\frac{1}{\gamma^2}P\right]^{-1}\right),$$

where $\hat{\omega}=\left[\left(1+\frac{\sigma_\nu^2}{\sigma_\varepsilon^2}\delta^2\right)\tilde{z}'\tilde{z}+\frac{1}{\gamma^2}P\right]^{-1}\tilde{z}'\left[x-\frac{\sigma_\nu^2}{\sigma_\varepsilon^2}\delta\left(y-x\left(\beta+\delta\right)-c\alpha\right)\right]$.

(iv) The conditional posterior of γ^2 is given by

$$p\left(\gamma^2|y,x,z,c,\beta,\alpha,\delta,\pi,\rho,\sigma_\nu^2,\sigma_\varepsilon^2\right)\propto\left(\frac{1}{\gamma^2}\right)^{\frac{k}{2}}e^{-\frac{\pi'\Omega^{-1}\pi}{2\gamma^2\sigma_\nu^2}},$$

which is the kernel of

$$\gamma^2|y,x,z,c,\beta,\alpha,\delta,\pi,\rho,\sigma_\nu^2,\sigma_\varepsilon^2\sim IG\left(\frac{\pi'\Omega^{-1}\pi}{2\sigma_\nu^2},\frac{k-2}{2}\right).$$

B.2 The case of $k=1$ or $k=2$

When $k\leq 2$, the previous algorithm requires some slight modifications, as the prior (17) cannot be implemented through the hierarchical specification described in section B.1. In the case of $k\leq 2$, the posterior is given by

$$p\left(\beta,\alpha,\delta,\pi,\rho,\sigma_\nu^2,\sigma_\varepsilon^2|y,x,z,c\right)\propto p\left(y|x,z,c,\beta,\alpha,\delta,\pi,\rho,\sigma_\nu^2,\sigma_\varepsilon^2\right)\cdot p\left(x|z,c,\beta,\alpha,\delta,\pi,\rho,\sigma_\nu^2,\sigma_\varepsilon^2\right)\cdot p\left(\pi|\sigma_\nu^2\right)$$

$$\propto\left(\frac{1}{\sigma_\varepsilon^2}\right)^{\frac{T}{2}}e^{-\frac{1}{2\sigma_\varepsilon^2}[y-x\beta-c\alpha-(x-z\pi-c\rho)\delta]'[y-x\beta-c\alpha-(x-z\pi-c\rho)\delta]}\left(\frac{1}{\sigma_\nu^2}\right)^{\frac{T}{2}}e^{-\frac{1}{2\sigma_\nu^2}(x-z\pi-c\rho)'(x-z\pi-c\rho)}\cdot$$

$$\cdot\left(\frac{\pi'z'z\pi}{\sigma_\nu^2}\right)^{-\frac{k-2}{2}}\left(\frac{1}{\sigma_\nu^2}\right)^{\frac{k}{2}},$$

and we can sample from it using the following Gibbs sampling with blocks (a) $\left(\beta,\alpha,\delta,\sigma_\varepsilon^2\right)$, (b) σ_ν^2, (c) ρ and (d) π:

(a) The conditional posterior of $\left(\beta,\alpha,\delta,\sigma_\varepsilon^2\right)$ is identical to the one derived in section B.1.

(b) The conditional posterior of σ_ν^2 is given by

$$p\left(\sigma_\nu^2|y,x,z,c,\beta,\alpha,\delta,\pi,\rho,\sigma_\varepsilon^2\right)\propto\left(\frac{1}{\sigma_\nu^2}\right)^{\frac{T+2}{2}}e^{-\frac{1}{2\sigma_\nu^2}(x-z\pi-c\rho)'(x-z\pi-c\rho)},$$

which is the kernel of

$$\sigma_\nu^2|y,x,z,c,\beta,\alpha,\delta,\pi,\rho,\sigma_\varepsilon^2 \sim IG\left(\frac{(x-z\pi-c\rho)'\,(x-z\pi-c\rho)}{2},\frac{T}{2}\right).$$

(c) The conditional posterior of ρ is given by

$$p\left(\rho|y,x,z,c,\beta,\alpha,\delta,\sigma_\nu^2,\sigma_\varepsilon^2,\pi\right) \propto$$

$$\propto e^{-\frac{1}{2\sigma_\varepsilon^2}[y-x\beta-c\alpha-(x-z\pi-c\rho)\delta]'[y-x\beta-c\alpha-(x-z\pi-c\rho)\delta]}e^{-\frac{1}{2\sigma_\nu^2}(x-z\pi-c\rho)'(x-z\pi-c\rho)}$$

$$\propto e^{-\frac{1}{2\sigma_\varepsilon^2}(\hat{y}+c\delta\rho)'(\hat{y}+c\delta\rho)-\frac{1}{2\sigma_\nu^2}(\hat{x}-c\rho)'(\hat{x}-c\rho)},$$

where $\hat{x} \equiv x - z\pi$ and $\hat{y} \equiv y - x\left(\beta+\delta\right) - c\alpha + z\pi\delta$. The last expression can be manipulated to show that it is proportional to

$$e^{-\frac{1}{2\sigma_\nu^2}\left[\rho'\left[\left(1+\frac{\sigma_\nu^2}{\sigma_\varepsilon^2}\delta^2\right)c'c\right]\rho-2\left(\hat{x}-\frac{\sigma_\nu^2}{\sigma_\varepsilon^2}\delta\hat{y}\right)'c\rho\right]},$$

which is the kernel of

$$\rho|y,x,z,c,\beta,\alpha,\delta,\sigma_\nu^2,\sigma_\varepsilon^2,\pi \sim N\left(\hat{\rho},\sigma_\nu^2\left[\left(1+\frac{\sigma_\nu^2}{\sigma_\varepsilon^2}\delta^2\right)c'c\right]^{-1}\right),$$

where $\hat{\rho} = \left[\left(1+\frac{\sigma_\nu^2}{\sigma_\varepsilon^2}\delta^2\right)c'c\right]^{-1}c'\left(\hat{x}-\frac{\sigma_\nu^2}{\sigma_\varepsilon^2}\delta\hat{y}\right)$.

(d) The conditional posterior of π is given by

$$p\left(\pi|y,x,z,c,\beta,\alpha,\delta,\pi,\rho,\sigma_\nu^2,\sigma_\varepsilon^2\right) \propto e^{-\frac{1}{2\sigma_\varepsilon^2}[y-x\beta-c\alpha-(x-z\pi-c\rho)\delta]'[y-x\beta-c\alpha-(x-z\pi-c\rho)\delta]}\cdot$$

$$\cdot e^{-\frac{1}{2\sigma_\nu^2}(x-z\pi-c\rho)'(x-z\pi-c\rho)}\left(\pi'z'z\pi\right)^{-\frac{k-2}{2}},$$

which can be manipulated to show that it is proportional to

$$e^{-\frac{1}{2\sigma_\nu^2}\left\{\pi'\left[z'z\left(1+\frac{\sigma_\nu^2}{\sigma_\varepsilon^2}\delta^2\right)\right]\pi-2\left(\bar{x}-\frac{\sigma_\nu^2}{\sigma_\varepsilon^2}\delta\bar{y}\right)'z\pi\right\}}\cdot\left(\pi'z'z\pi\right)^{-\frac{k-2}{2}}, \tag{18}$$

where $\bar{y} \equiv y - x\beta - c\alpha - x\delta + c\rho\delta$ and $\bar{x} \equiv x - c\rho$. Observe that the first term of (18) is the kernel of

$$\pi|y,x,z,c,\beta,\alpha,\delta,\pi,\rho,\sigma_\nu^2,\sigma_\varepsilon^2 \sim N\left(\bar{\pi},\sigma_\nu^2\left[\left(1+\frac{\sigma_\nu^2}{\sigma_\varepsilon^2}\delta^2\right)z'z\right]^{-1}\right). \tag{19}$$

Therefore, we can sample from (18) using a Metropolis-within-Gibbs strategy. First, we draw a candidate π, say π^*, from (19), which we interpret as a proposal density. We accept π^* with probability

$$p = \min\left\{1, \left(\frac{\pi^{*\prime}z'z\pi^*}{\pi^{(j-1)\prime}z'z\pi^{(j-1)}}\right)^{-\frac{k-2}{2}}\right\},$$

where $\pi^{(j-1)}$ is the previous draw of π in the chain. If π^* is not accepted, we set $\pi^{(j)} = \pi^{(j-1)}$. Notice that, when $k = 1$, the acceptance probability simplifies to $p = \min\left\{1, \frac{|\pi^*|}{|\pi^{(j-1)}|}\right\}$. With $k = 2$, instead, $p = 1$ and π^* is always accepted.

C Proofs of the results presented in section 5

This appendix provides the proofs of Lemma 1, and Propositions 2 and 3, as stated in section 5. To that end, we begin by establishing two preliminary lemmas.

Lemma 3. *Let P_θ be a rotation matrix defined by $P_\theta = \left[\phi_\theta \,\vdots\, \phi_\theta^\perp\right]$, with $\phi_\theta \equiv [\cos(\theta), \sin(\theta)]'$ and $\phi_\theta^\perp \equiv [-\sin(\theta), \cos(\theta)]'$. Then, the matrix $\hat{\Gamma}$, defined in section 5.1, can be written as*

$$\hat{\Gamma} = P_\theta \begin{bmatrix} Q & R\sqrt{Q} \\ R\sqrt{Q} & J + R^2 \end{bmatrix} P_\theta'.$$

In this expression, J, R and Q are mutually independent random variables with distributions

$$J \sim \chi^2_{k-1}$$

$$R \sim \mathcal{N}(0,1)$$

$$Q \sim \begin{cases} (1+r)\,\chi^2_k & \text{under the DGP of econometrician } B \text{ and } RC \\ \chi^2_k(\rho) & \text{under the DGP of econometrician } FC, \end{cases}$$

where $\chi^2_k(\rho)$ denotes the non-central chi-squared distribution with k degrees of freedom and non-centrality parameter ρ.

Proof. The result follows directly from equations (7) and (8), combined with the representation theorem 2.2 in Gleser (1976). □

Lemma 4. *Let $\hat{\lambda}_1$ and $\hat{\lambda}_2$ denote the eigenvalues of $\hat{\Gamma}$, with $\hat{\lambda}_1 > \hat{\lambda}_2$. Let $\hat{v}$ denote the eigenvector of*

$\hat{\Gamma}$ *associated with* $\hat{\lambda}_1$*, and define the angle of this unit-length vector by* $\hat{\theta} = \arctan\left(\frac{\hat{v}_2}{\hat{v}_1}\right)$*. Then,* $\hat{\lambda}_1$ *and* $\hat{\lambda}_2$ *are ancillary statistics for* θ*. In addition,* $\hat{\lambda}_1$*,* $\hat{\lambda}_2$ *and* $\hat{\theta}$ *satisfy*

$$\hat{\lambda}_1 = \frac{1}{2}\left[Q + J + R^2 + \sqrt{\left(Q + J + R^2\right)^2 - 4JQ}\right]$$

$$\hat{\lambda}_2 = \frac{1}{2}\left[Q + J + R^2 - \sqrt{\left(Q + J + R^2\right)^2 - 4JQ}\right]$$

$$\begin{aligned}\hat{\theta} &= \theta + \arctan\left(\frac{R\sqrt{Q}}{\hat{\lambda}_1 - J - R^2}\right)\\ &= \theta + \arctan\left(\frac{\hat{\lambda}_1 - Q}{R\sqrt{Q}}\right).\end{aligned}$$

Proof. These expressions are obtained by computing the eigenvalues and eigenvectors of $\hat{\Gamma}$, using the representation of $\hat{\Gamma}$ derived in Lemma 3. $\hat{\lambda}_1$ and $\hat{\lambda}_2$ are ancillary for θ because their distribution does not depend on θ. □

C.1 Proofs of Lemma 1

To prove Lemma 1, recall that $\hat{\Gamma}$ is a sufficient statistic for inference under the model of econometrician B. Therefore, the (posterior) distribution of the random coefficient θ given the observed data satisfies

$$\begin{aligned}p_B\left(\theta|\hat{\Gamma}\right) &= p_B\left(\theta|\hat{\lambda}_1, \hat{\lambda}_2, \hat{\theta}\right)\\ &\propto p_B\left(\hat{\lambda}_1, \hat{\lambda}_2, \hat{\theta}|\theta\right)\\ &\propto p_B\left(\hat{\theta}|\theta, \hat{\lambda}_1, \hat{\lambda}_2\right)\\ &\propto p_{RC}\left(\hat{\theta}|\theta, \hat{\lambda}_1, \hat{\lambda}_2\right),\end{aligned}$$

where the first equality follows from the one-to-one mapping between $\hat{\Gamma}$ and its spectral decomposition; the second line from the uniform prior on θ; the third line from the ancillarity of $\hat{\lambda}_1$ and $\hat{\lambda}_2$; and the last line from the fact that econometrician B and RC share the same marginal likelihood function of θ, obtained by integrating out the random coefficient r.

To show that these functions are symmetric in $\theta - \hat{\theta}$, observe that

$$\begin{aligned} p_B\left(\theta, r|\hat{\Gamma}\right) &\propto p_B\left(\hat{\Gamma}|\theta, r\right) \\ &\propto \left|I_2 + r\phi_\theta\phi_\theta'\right|^{-\frac{k}{2}} \exp\left\{-\frac{1}{2}\text{tr}\left[\left(I_2 + r\phi_\theta\phi_\theta'\right)^{-1}\hat{\Gamma}\right]\right\} \\ &\propto (1+r)^{-\frac{k}{2}} \exp\left\{-\frac{1}{2}\text{tr}\left[\left(I_2 + r\phi_\theta\phi_\theta'\right)^{-1}\hat{\Gamma}\right]\right\} \\ &\propto (1+r)^{-\frac{k}{2}} \exp\left\{\frac{1}{2}\frac{r}{1+r}\phi_\theta'\hat{\Gamma}\phi_\theta\right\} \\ &\propto (1+r)^{-\frac{k}{2}} \exp\left\{\frac{1}{2}\frac{r}{1+r}\left[\hat{\lambda}_1 - \left(\hat{\lambda}_1 - \hat{\lambda}_2\right)\sin^2\left(\theta - \hat{\theta}\right)\right]\right\} \end{aligned}$$

where the first line follows from the uniform prior on (θ, r); the second line from (7); the third line from the fact that the eigenvalues of $I_2 + r\phi_\theta\phi_\theta'$ are 1 and $1 + r$; the fourth line from the Woodbury matrix identity; and the last line from the spectral decomposition of $\hat{\Gamma}$, given by $\hat{\Gamma} = P_{\hat{\theta}}\begin{bmatrix}\hat{\lambda}_1 & 0 \\ 0 & \hat{\lambda}_2\end{bmatrix}P_{\hat{\theta}}'$, and from the fact that $\phi_\theta' P_{\hat{\theta}} = \phi_{\theta-\hat{\theta}}'$. Since $p_B\left(\theta, r|\hat{\Gamma}\right)$ depends only on a periodic and even function of $\theta - \hat{\theta}$ for any given r, the marginal density of θ, obtained as

$$p_B\left(\theta|\hat{\Gamma}\right) = \int_0^M p_B\left(\theta, r|\hat{\Gamma}\right) dr,$$

depends only on a periodic and even function of $\theta - \hat{\theta}$ as well.

Although not needed for this proof, it can be shown that the posterior density admits the closed-form representation

$$p_B(\theta \mid \hat{\Gamma}) \propto {}_1F_1\left(1, \frac{k}{2}, \phi_\theta'\hat{\Gamma}\phi_\theta\right),$$

where ${}_1F_1$ denotes the confluent hypergeometric function of the first kind.

C.2 Proof of Proposition 2

To prove Proposition 2, we require the following lemma describing the asymptotic behavior of the eigenvalues of $\hat{\Gamma}$. Unless otherwise stated, all limits are taken as $T, k \to \infty$.

Lemma 5. *Suppose that there exists a constant* $\underline{c} > 0$ *such that* $\lim_{T,k\to\infty} \frac{\rho}{k} \geq \underline{c}$. *Then,*

$$\frac{\hat{\lambda}_1}{k} = \frac{Q}{k} + O_p\left(\frac{1}{k}\right)$$

and

$$\frac{\hat{\lambda}_2}{k} = \frac{J}{k} + O_p\left(\frac{1}{k}\right),$$

under both the DGP of econometrician B and FC.

Proof. From Lemma 4,

$$\hat{\lambda}_1 = \frac{1}{2}\left[Q + J + R^2 + \sqrt{\left(Q + J + R^2\right)^2 - 4JQ}\right].$$

Rearranging the terms inside the square root, we obtain

$$\hat{\lambda}_1 = \frac{1}{2}\left[Q + J + R^2 + \sqrt{\left(Q - J - R^2\right)^2 + 4QR^2}\right]$$

$$= \frac{1}{2}\left[Q + J + R^2 + \left|Q - J - R^2\right|\sqrt{1 + \frac{4QR^2}{\left(Q - J - R^2\right)^2}}\right].$$

Notice that $\frac{R^2}{k} = O_p\left(\frac{1}{k}\right)$, $\frac{J}{k} = 1 + O_p\left(\frac{1}{\sqrt{k}}\right)$, and $\frac{Q}{k} = 1 + \frac{\rho}{k} + O_p\left(\frac{1}{\sqrt{k}}\right)$ under the model of econometrician FC, and $\frac{Q}{k} = 1 + E_B\left(r\right) + O_p\left(\frac{1}{\sqrt{k}}\right)$ under the model of econometricians B and RC. Since

$$\frac{4QR^2}{\left(Q - J - R^2\right)^2} = \frac{4\frac{Q}{k}\frac{R^2}{k}}{\left(\frac{Q}{k} - \frac{J}{k} - \frac{R^2}{k}\right)^2} = O_p\left(\frac{1}{k}\right),$$

we can perform a Taylor expansion of $\sqrt{\left[1 + \frac{4QR^2}{\left(Q-J-R^2\right)^2}\right]}$ around 1, which yields

$$\hat{\lambda}_1 = \frac{1}{2}\left[Q + J + R^2 + \left|Q - J - R^2\right|\left(1 + \frac{1}{2}\frac{4QR^2}{\left(Q - J - R^2\right)^2} + O_p\left(\left(\frac{QR^2}{\left(Q - J - R^2\right)^2}\right)^2\right)\right)\right]$$

$$= \frac{1}{2}\left[Q + J + R^2 + \left|Q - J - R^2\right|\left(1 + O_p\left(\frac{1}{k}\right)\right)\right].$$

This expression implies

$$\frac{\hat{\lambda}_1}{k} = \frac{1}{2}\left[\frac{Q}{k} + \frac{J}{k} + \frac{R^2}{k} + \left|\frac{Q}{k} - \frac{J}{k} - \frac{R^2}{k}\right|\left(1 + O_p\left(\frac{1}{k}\right)\right)\right]$$

$$= \frac{1}{2}\left[\frac{Q}{k} + \frac{J}{k} + \frac{R^2}{k} + \left|\frac{Q}{k} - \frac{J}{k} - \frac{R^2}{k}\right| + O_p\left(\frac{1}{k}\right)\right]$$

$$= \frac{Q}{k} + O_p\left(\frac{1}{k}\right),$$

where the last equality follows from the fact that $\left|\frac{Q}{k}-\frac{J}{k}-\frac{R^2}{k}\right|=\frac{Q}{k}-\frac{J}{k}-\frac{R^2}{k}+O_p\left(\frac{1}{k}\right)$.[8]

Similar steps can be used to establish the second part of the lemma, namely that

$$\frac{\hat{\lambda}_2}{k}=\frac{J}{k}+O_p\left(\frac{1}{k}\right).$$

□

In sum, Lemma 5 states that $\hat{\lambda}_1$ and $\hat{\lambda}_2$ are extremely accurate approximations of Q and J, which is a key input into the proof of the main claim of Proposition 2. To prove this proposition, recall from Lemma 4 that

$$\hat{\theta}-\theta=\arctan\left(\frac{R\sqrt{Q}}{\hat{\lambda}_1-J-R^2}\right).$$

Since $\frac{R\sqrt{Q}}{\hat{\lambda}_1-J-R^2}=O_p\left(\frac{1}{\sqrt{k}}\right)$, we can perform a Taylor expansion of $\arctan\left(\frac{R\sqrt{Q}}{\hat{\lambda}_1-J-R^2}\right)$ around 0. Given that the second derivative of the $\arctan$ function evaluated at the origin is equal to 0, we have

$$\begin{aligned}\hat{\theta}-\theta&=\frac{R\sqrt{Q}}{\hat{\lambda}_1-J-R^2}+O_p\left(\left(\frac{R\sqrt{Q}}{\hat{\lambda}_1-J-R^2}\right)^3\right)\\&=\frac{R\sqrt{Q}}{\hat{\lambda}_1-J-R^2}+O_p\left(\frac{1}{k\sqrt{k}}\right),\end{aligned}$$

which implies

$$\begin{aligned}\sqrt{k}\left(\hat{\theta}-\theta\right)&=R\frac{\sqrt{\frac{Q}{k}}}{\frac{\hat{\lambda}_1}{k}-\frac{J}{k}-\frac{R^2}{k}}+O_p\left(\frac{1}{k}\right)\\&=R\frac{\sqrt{\frac{\hat{\lambda}_1}{k}+O_p\left(\frac{1}{k}\right)}}{\frac{\hat{\lambda}_1}{k}-\frac{\hat{\lambda}_2}{k}+O_p\left(\frac{1}{k}\right)}+O_p\left(\frac{1}{k}\right)\\&=R\frac{\sqrt{\frac{\hat{\lambda}_1}{k}}}{\frac{\hat{\lambda}_1}{k}-\frac{\hat{\lambda}_2}{k}}+O_p\left(\frac{1}{k}\right),\end{aligned}$$

where the second line follows from applying the results of Lemma 5, and the last line from the fact that $\frac{\sqrt{\frac{\hat{\lambda}_1}{k}+O_p\left(\frac{1}{k}\right)}}{\frac{\hat{\lambda}_1}{k}-\frac{\hat{\lambda}_2}{k}+O_p\left(\frac{1}{k}\right)}=\frac{\sqrt{\frac{\hat{\lambda}_1}{k}}}{\frac{\hat{\lambda}_1}{k}-\frac{\hat{\lambda}_2}{k}}+O_p\left(\frac{1}{k}\right)$.

[8] To show that $\left|\frac{Q}{k}-\frac{J}{k}-\frac{R^2}{k}\right|-\left(\frac{Q}{k}-\frac{J}{k}-\frac{R^2}{k}\right)\equiv D=O_p\left(\frac{1}{k}\right)$, notice that $D=\begin{cases}0 & \text{if } Q\geq J+R^2\\ 2\left(\frac{J}{k}+\frac{R^2}{k}-\frac{Q}{k}\right) & \text{if } Q<J+R^2\end{cases}$. Applying Chebyshev inequality, we obtain $\Pr\left\{\left|D>\frac{\bar{M}}{k}\right|\right\}=\Pr\left\{J+R^2-Q>\frac{\bar{M}}{2}\right\}\leq\frac{4(k+\rho)}{\left(\bar{M}/2+\rho\right)^2}$. It follows that $D=O_p\left(\frac{1}{k}\right)$, since, for any $\epsilon>0$, there exists an $\bar{M}$ such that $\Pr\left(\left|D>\frac{\bar{M}}{k}\right|\right)<\epsilon$ for large k.

Finally, since $\frac{\sqrt{\frac{\hat{\lambda}_1}{k}}}{\frac{\hat{\lambda}_1}{k}-\frac{\hat{\lambda}_2}{k}} = O_p(1)$, we can divide both the left- and right-hand sides of the last expression by this term to obtain

$$\frac{\hat{\theta}-\theta}{\frac{\sqrt{\hat{\lambda}_1}}{\hat{\lambda}_1-\hat{\lambda}_2}} = R + O_p\left(\frac{1}{k}\right),$$

which concludes the proof of Proposition 2.

C.3 Proof of Proposition 3

The log-likelihood-ratio test statistic of Moreira (2003) is defined as

$$LR = 2\left[\max_{\beta,\pi} \log p\left(y,x|z,\beta,\pi\right) - \max_{\pi} \log p\left(y,x|z,\beta,\pi\right)\right],$$

where

$$p\left(y,x|z,\beta,\pi\right) \propto \exp\left\{-\frac{1}{2}\mathrm{tr}\left[\Sigma^{-1}\left(w-z\pi b'\right)'\left(w-z\pi b'\right)\right]\right\} \tag{20}$$

is the likelihood function implied by equations (5)-(6), and $w \equiv [x,y]$. To compute the profile likelihood, we maximize (20) with respect to π, by setting π equal to

$$\hat{\pi}_\beta = \frac{\left(z'z\right)^{-1} z'w\Sigma^{-1}b}{b'\Sigma^{-1}b}.$$

Substituting $\pi = \hat{\pi}_\beta$ into $p\left(y,x|z,\beta,\pi\right)$, and simplifying the resulting expression, we obtain

$$\max_{\pi} p\left(y,x|z,\beta,\pi\right) \propto \exp\left\{\frac{1}{2}\frac{b'C}{\left(b'\Sigma^{-1}b\right)^{\frac{1}{2}}}\hat{\Gamma}\frac{C'b}{\left(b'\Sigma^{-1}b\right)^{\frac{1}{2}}}\right\}.$$

Switching to polar coordinates yields the simpler expression

$$\max_{\pi} p\left(y,x|z,\theta,\pi\right) \propto \exp\left\{\frac{1}{2}\phi_\theta'\hat{\Gamma}\phi_\theta\right\},$$

which reaches its maximum, $\exp\left\{\frac{1}{2}\hat{\lambda}_1\right\}$, when $\theta = \hat{\theta}$, the MLE defined in section 5.4. In sum, the log-likelihood-ratio test statistic of Moreira (2003) is

$$\begin{aligned} LR &= \hat{\lambda}_1 - \phi_\theta'\hat{\Gamma}\phi_\theta \\ &= \hat{\lambda}_1 - Q, \end{aligned}$$

where the second line follows from Lemma 3.

To prove Proposition 3, recall from Lemma 4 that

$$\hat{\theta}-\theta=\arctan\left(\frac{\hat{\lambda}_1-Q}{R\sqrt{Q}}\right).$$

Since $\frac{\hat{\lambda}_1-Q}{R\sqrt{Q}}=O_p\left(\frac{1}{\sqrt{k}}\right)$, we can perform a Taylor expansion of $\arctan\left(\frac{\hat{\lambda}_1-Q}{R\sqrt{Q}}\right)$ around 0. Given that the second derivative of the $\arctan$ function evaluated at the origin is equal to 0, we have

$$\begin{aligned}\hat{\theta}-\theta&=\frac{\hat{\lambda}_1-Q}{R\sqrt{Q}}+O_p\left(\left(\frac{\hat{\lambda}_1-Q}{R\sqrt{Q}}\right)^3\right)\\&=\frac{\hat{\lambda}_1-Q}{R\sqrt{Q}}+O_p\left(\frac{1}{k\sqrt{k}}\right),\end{aligned}$$

which implies

$$\sqrt{k}\left(\hat{\theta}-\theta\right)=\frac{\hat{\lambda}_1-Q}{R\sqrt{\frac{Q}{k}}}+O_p\left(\frac{1}{k}\right),$$

or, equivalently,

$$\hat{\lambda}_1-Q=\sqrt{k}\left(\hat{\theta}-\theta\right)R\sqrt{\frac{Q}{k}}+O_p\left(\frac{1}{k}\right).$$

Multiplying and dividing the right-hand-side of this expression by $\frac{\sqrt{\frac{\hat{\lambda}_1}{k}}}{\frac{\hat{\lambda}_1}{k}-\frac{\hat{\lambda}_2}{k}}$, and substituting the expression for R derived in Proposition 2, we obtain

$$\begin{aligned}\hat{\lambda}_1-Q&=\sqrt{k}\frac{\left(\hat{\theta}-\theta\right)}{\frac{\sqrt{\frac{\hat{\lambda}_1}{k}}}{\frac{\hat{\lambda}_1}{k}-\frac{\hat{\lambda}_2}{k}}}\frac{\frac{\hat{\lambda}_1}{k}}{\frac{\hat{\lambda}_1}{k}-\frac{\hat{\lambda}_2}{k}}R+O_p\left(\frac{1}{k}\right)\\&=\sqrt{k}\frac{\left(\hat{\theta}-\theta\right)}{\frac{\sqrt{\frac{\hat{\lambda}_1}{k}}}{\frac{\hat{\lambda}_1}{k}-\frac{\hat{\lambda}_2}{k}}}\frac{\frac{\hat{\lambda}_1}{k}}{\frac{\hat{\lambda}_1}{k}-\frac{\hat{\lambda}_2}{k}}\left[\sqrt{k}\frac{\left(\hat{\theta}-\theta\right)}{\frac{\sqrt{\frac{\hat{\lambda}_1}{k}}}{\frac{\hat{\lambda}_1}{k}-\frac{\hat{\lambda}_2}{k}}}+O_p\left(\frac{1}{k}\right)\right]+O_p\left(\frac{1}{k}\right)\\&=\left[\sqrt{k}\frac{\left(\hat{\theta}-\theta\right)}{\frac{\sqrt{\frac{\hat{\lambda}_1}{k}}}{\frac{\hat{\lambda}_1}{k}-\frac{\hat{\lambda}_2}{k}}}\right]^2\frac{\frac{\hat{\lambda}_1}{k}}{\frac{\hat{\lambda}_1}{k}-\frac{\hat{\lambda}_2}{k}}+O_p\left(\frac{1}{k}\right)\end{aligned}$$

$$= \hat{W}^2 \frac{\frac{\hat{\lambda}_1}{k}}{\frac{\hat{\lambda}_1}{k} - \frac{\hat{\lambda}_2}{k}} + O_p\left(\frac{1}{k}\right),$$

where $\hat{W} = \frac{\hat{\theta}-\theta}{\frac{\sqrt{\hat{\lambda}_1}}{\hat{\lambda}_1-\hat{\lambda}_2}}$ is the pivotal statistic defined in Proposition 2. Finally, using the results of Lemma 5, we have

$$\hat{\lambda}_1 - Q = \hat{W}^2 \frac{Q}{Q-J} + O_p\left(\frac{1}{k}\right).$$

Recall from Lemma 3 that $J \sim \chi^2_{k-1}$, and it is thus ancillary for θ. In addition, J is asymptotically independent from $\hat{W}$, since $\hat{W}$ converges in probability to R.

C.4 Proof of Lemma 2

To prove Lemma 2, we first show that

$$\hat{\Sigma} = \Sigma + O_p\left(\frac{1}{k}\right).$$

To see this, note that $\hat{\Sigma} = \frac{1}{T-k}\left(w - z\hat{\Pi}\right)'\left(w - z\hat{\Pi}\right)$ can be decomposed as

$$\hat{\Sigma} = \frac{1}{T-k}(w - z\Pi)'(w - z\Pi) - \frac{2}{T-k}\left(\hat{\Pi} - \Pi\right)' z'(w - z\Pi) + \frac{1}{T-k}\left(\hat{\Pi} - \Pi\right)' z'z\left(\hat{\Pi} - \Pi\right)$$

$$= \frac{1}{T-k}u'u - \frac{1}{T-k}u'z\left(z'z\right)^{-1}z'u$$

$$= \frac{1}{T-k}u'\left[I - z\left(z'z\right)^{-1}z'\right]u,$$

where $\Pi = [\pi, \pi\beta]$ and $u = (w - z\Pi)$. Under the assumption that $z'z$ is full rank, the matrix $M_z = I - z\left(z'z\right)^{-1}z'$ is symmetric and idempotent, and it has rank equal to $T-k$. Since

$$u \sim \mathcal{MN}\left(0, I_2, \Sigma\right),$$

standard results (e.g. Anderson, 2003) imply that

$$\hat{\Sigma} \sim \frac{1}{T-k}\mathcal{W}\left(\Sigma, T-k\right).$$

It follows that

$$\hat{\Sigma} = \Sigma + O_p\left(\frac{1}{\sqrt{T-k}}\right) = \Sigma + O_p\left(\frac{1}{k}\right), \tag{21}$$

where the last equality uses the assumption that $\frac{k^2}{T} = O_p(1)$.

We now use this result to prove the main claim of Lemma 2. Given the definitions of $\hat{\Gamma}$ and $\tilde{\Gamma}$, note that we can re-write $\tilde{\Gamma}$ as

$$\begin{aligned}\tilde{\Gamma} &= \left(C^{-1}\hat{C}\right)'\hat{\Gamma}\left(C^{-1}\hat{C}\right)\\ &= (I_2+E)'\,\hat{\Gamma}\,(I_2+E),\end{aligned} \tag{22}$$

where

$$E \equiv C^{-1}\left(\hat{C} - C\right).$$

Since $C = g(\Sigma)$, $\hat{C} = g\left(\hat{\Sigma}\right)$, g is continuous, and Σ and C are of full rank, expression (21) implies that $E = O_p\left(\frac{1}{k}\right)$. It follows from (22) that

$$\frac{1}{k}\tilde{\Gamma} = \frac{1}{k}\hat{\Gamma} + O_p\left(\frac{1}{k}\right).$$

C.5 The case of irrelevant instruments

When instruments are perfectly irrelevant, the credible intervals produced by our Bayesian approach are typically conservative, and small size distortions arise only in very extreme cases. To make this statement precise, recall that the mapping from θ to β is given by $\theta = \arctan\left(\frac{[C'b]_2}{[C'b]_1}\right)$, where C is upper triangular with positive diagonal entries and satisfies $C'C = \Sigma^{-1}$. The inverse mapping is therefore

$$\beta = \frac{\Sigma_{12}}{\Sigma_{11}} + \frac{[\det(\Sigma)]^{\frac{1}{2}}}{\Sigma_{11}}\tan(\theta).$$

If the data are generated by the structural model (1)-(2) with "true" parameters β_0, δ_0, $\sigma_{0,\varepsilon}$ and $\sigma_{0,\nu}$, then $\Sigma_{11} = \sigma^2_{0,\nu}$, $\Sigma_{12} = (\beta_0+\delta_0)\sigma^2_{0,\nu}$, $\Sigma_{22} = (\beta_0+\delta_0)^2\sigma^2_{0,\nu} + \sigma^2_{0,\varepsilon}$, and $\det(\Sigma) = \sigma^2_\nu\sigma^2_\varepsilon$. Hence, the mapping can equivalently be written as

$$\begin{aligned}\beta &= \beta_0 + \delta_0 + \frac{\sigma_{0,\varepsilon}}{\sigma_{0,\nu}}\tan(\theta)\\ &= \beta_0 + \frac{\sigma_{0,\varepsilon}}{\sigma_{0,\nu}}\left[\sqrt{\frac{\varrho_0^2}{1-\varrho_0^2}} + \tan(\theta)\right],\end{aligned} \tag{23}$$

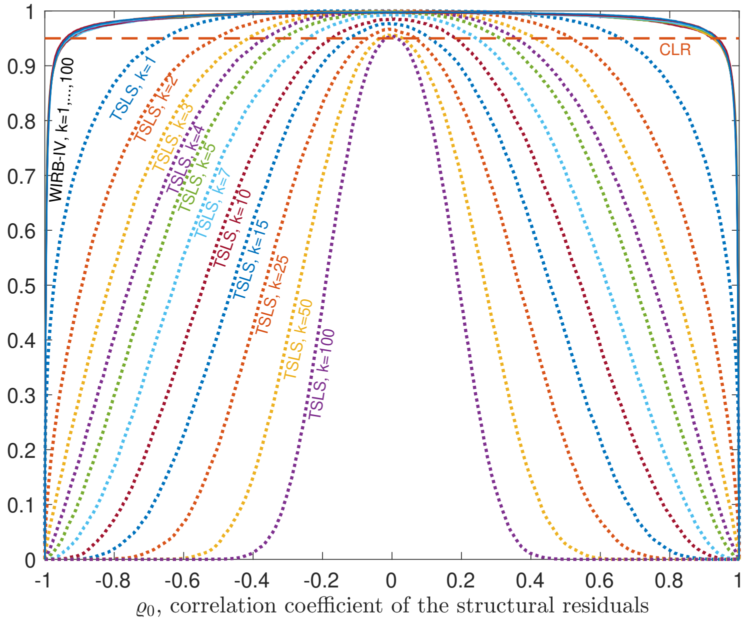

Figure 3: Coverage of 95-percent confidence (credible) intervals based on (i) the asymptotic distribution of TSLS; (ii) the inversion of Moreira's (2003) conditional likelihood ratio test (CLR) with known Σ; (iii) our weak-instrument-robust Bayesian approach (WIRB-IV) with known Σ. The results are based on $10,000$ simulations for each value of k from model (1)-(2) with irrelevant instruments (i.e. $\pi = 0_{k\times 1}$). The coverage of CLR and WIRB-IV does not depend on any other parameters. For TSLS, the simulations use $T = 250$, but its dependence on T is negligible.

where the second equality follows from the definition of the correlation coefficient of the structural errors, $\varrho_0 \equiv \frac{\text{cov}(\nu,\delta_0\nu+\varepsilon)}{\sqrt{\text{var}(\nu)\text{var}(\delta_0\nu+\varepsilon)}} = \frac{\delta_0\sigma^2_{0,\nu}}{\sqrt{\sigma^2_{0,\nu}\left(\delta_0^2\sigma^2_{0,\nu}+\sigma^2_{0,\varepsilon}\right)}}$.

Expression (23) shows that the credible interval for β contains the true value β_0 if and only if the corresponding credible interval for $\tan(\theta)$ contains $-\sqrt{\varrho_0^2/(1-\varrho_0^2)}$. The posterior distribution of θ admits the closed-form expressions derived in appendix C.1. When instruments are irrelevant, $\hat{\Gamma} \sim \mathcal{W}(I_2, k)$, and its behavior in repeated samples only depends on k, so coverage can easily be evaluated by simulation. Figure 3 reports the coverage of the 95-percent equal-tailed credible intervals for β as a function of ϱ_0, for different values of k. Even for large values of this correlation coefficient, the Bayesian credible intervals remain valid and are indeed conservative. Size distortions appear only when $|\varrho_0|$ exceeds about 0.935, and even then remain below 5 percent as long as $|\varrho_0| \leq 0.975$. Finally, these results are essentially independent of the number of instruments.

For comparison, the figure also reports the coverage of standard TSLS 95-percent confidence intervals. With a single irrelevant instrument, coverage falls below the nominal level as soon as the degree of endogeneity implies a correlation exceeding about 0.64. Moreover, as is well known, adding instruments substantially worsens performance. The figure also shows the coverage of 95-percent confidence intervals obtained by inverting the CLR test, which has exact nominal size by construction. Relative to WIRB-IV, CLR is almost always less conservative, and becomes comparatively preferable only when endogeneity is extremely high (roughly $|\varrho_0| > 0.935$).

D Additional simulation evidence

This appendix presents additional simulation evidence on the performance of WIRB-IV. Relative to the Monte Carlo experiment in the main text, we consider settings with varying numbers of instruments and different degrees of endogeneity. For completeness, recall that the baseline simulation results in the main text are obtained using model (1)-(2) with $T = 250$, $\beta = 0$, $\varepsilon \sim N(0, I_T)$, $\nu \sim N(0, I_T)$, $\pi \sim \mathcal{N}(0, s^2 I_k)$, $s \sim Uniform\,(0, 0.25)$, $\delta = 0.75$ and $k = 10$. Here, we experiment with $k = 1$, 5, 10, 25 and 100, and with $\delta = 0.75$ and 3. The results are presented in figures 4-13, which have the same format of figure 2 in section 3.

References

Anatolyev, S. and M. Sølvsten (2023): "Testing many restrictions under heteroskedasticity," *Journal of Econometrics*, 236.

Anderson, T. W. (2003): *An Introduction to Multivariate Statistical Analysis*, Hoboken, NJ: John Wiley & Sons, 3 ed.

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

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

Andrews, D. W. K., M. J. Moreira, and J. H. Stock (2006): "Optimal Two-Sided Invariant Similar Tests for Instrumental Variables Regression," *Econometrica*, 74, 715–752.

Andrews, D. W. K. and J. H. Stock (2007b): "Inference with Weak Instruments," in *Advances in Economics and Econometrics: Theory and Applications, Ninth World Congress, Vol. III*, ed. by R. Blundell, W. K. Newey, and T. Persson, Cambridge, UK: Cambridge University Press, 122–173.

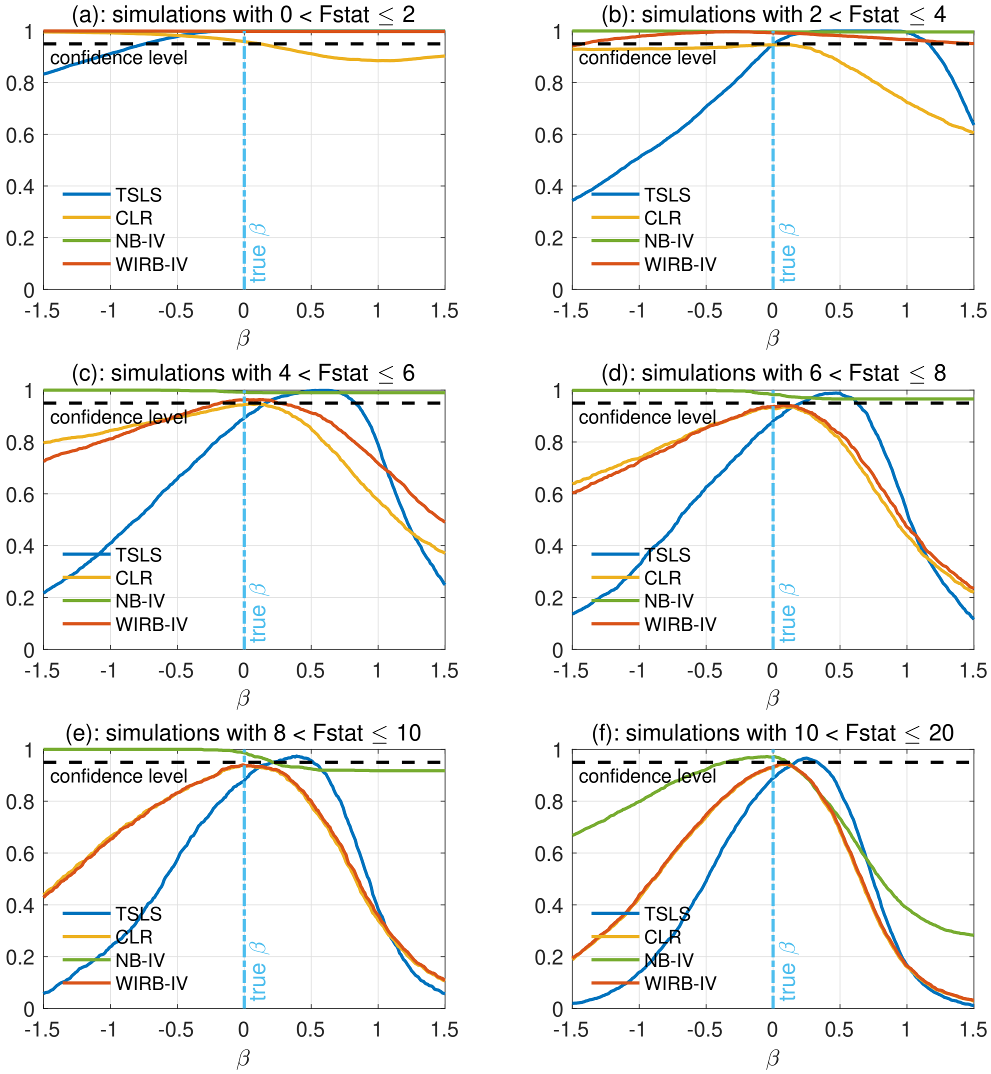

Figure 4: Frequency of inclusion in the 95-percent confidence (credible) interval based on (i) the asymptotic distribution of TSLS; (ii) the inversion of Moreira's (2003) conditional likelihood ratio test (CLR); (iii) the naive-Bayesian approach (NB-IV); and (iv) our weak-instrument-robust Bayesian approach (WIRB-IV). The results are based on $25,000$ simulations from model (1)-(2) with $T = 250$, $\beta = 0$, $\varepsilon \sim N(0, I_T)$, $\nu \sim N(0, I_T)$, $\pi \sim \mathcal{N}(0, s^2 I_k)$, $s \sim Uniform\,(0, 0.25)$, $\delta = 0.75$ and $k = 1$. With $k = 1$, NB-IV yields an improper posterior unless additional restrictions are imposed. To obtain a proper posterior in this case, we truncate the uniform prior over a sufficiently wide interval.

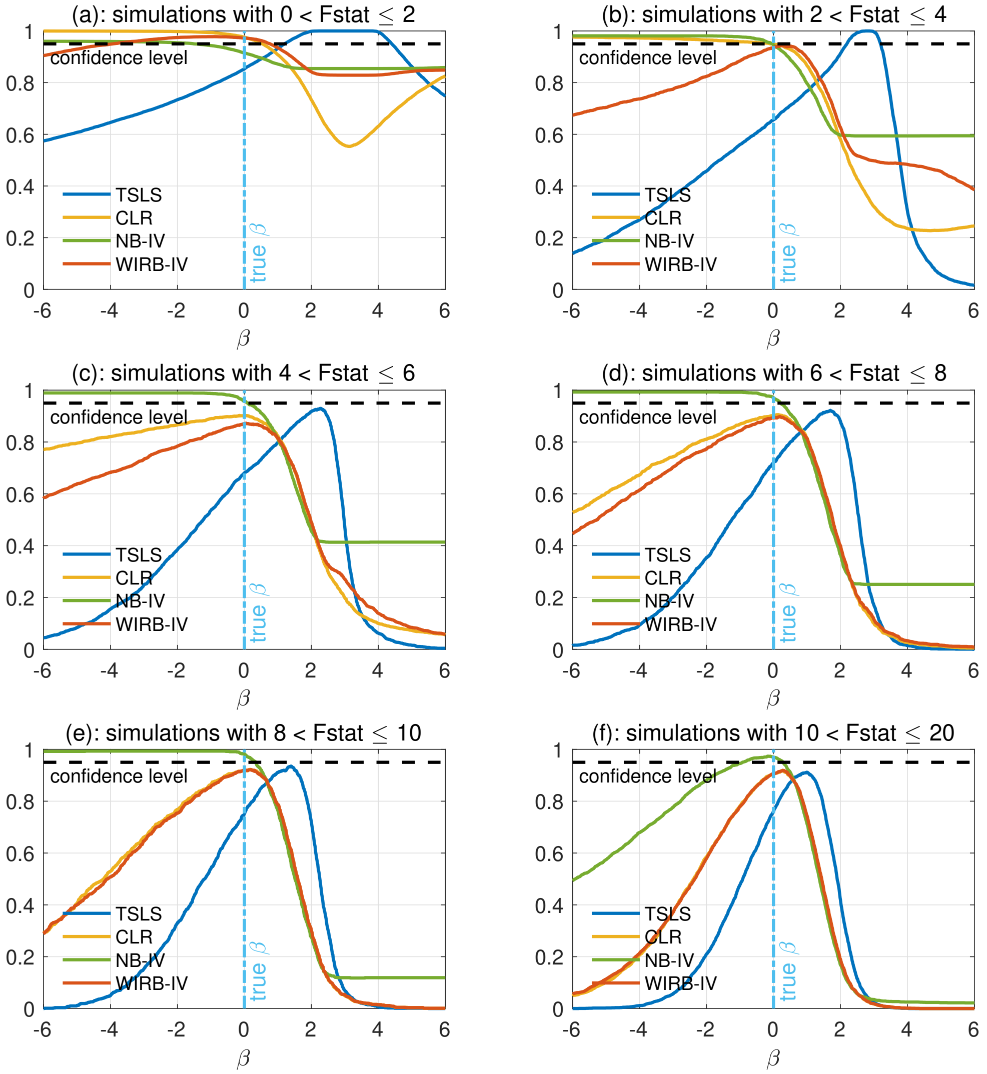

Figure 5: Frequency of inclusion in the 95-percent confidence (credible) interval based on (i) the asymptotic distribution of TSLS; (ii) the inversion of Moreira's (2003) conditional likelihood ratio test (CLR); (iii) the naive-Bayesian approach (NB-IV); and (iv) our weak-instrument-robust Bayesian approach (WIRB-IV). The results are based on $25,000$ simulations from model (1)-(2) with $T = 250$, $\beta = 0$, $\varepsilon \sim N(0, I_T)$, $\nu \sim N(0, I_T)$, $\pi \sim \mathcal{N}(0, s^2 I_k)$, $s \sim Uniform\,(0, 0.25)$, $\delta = 3$ and $k = 1$. With $k = 1$, NB-IV yields an improper posterior unless additional restrictions are imposed. To obtain a proper posterior in this case, we truncate the uniform prior over a sufficiently wide interval.

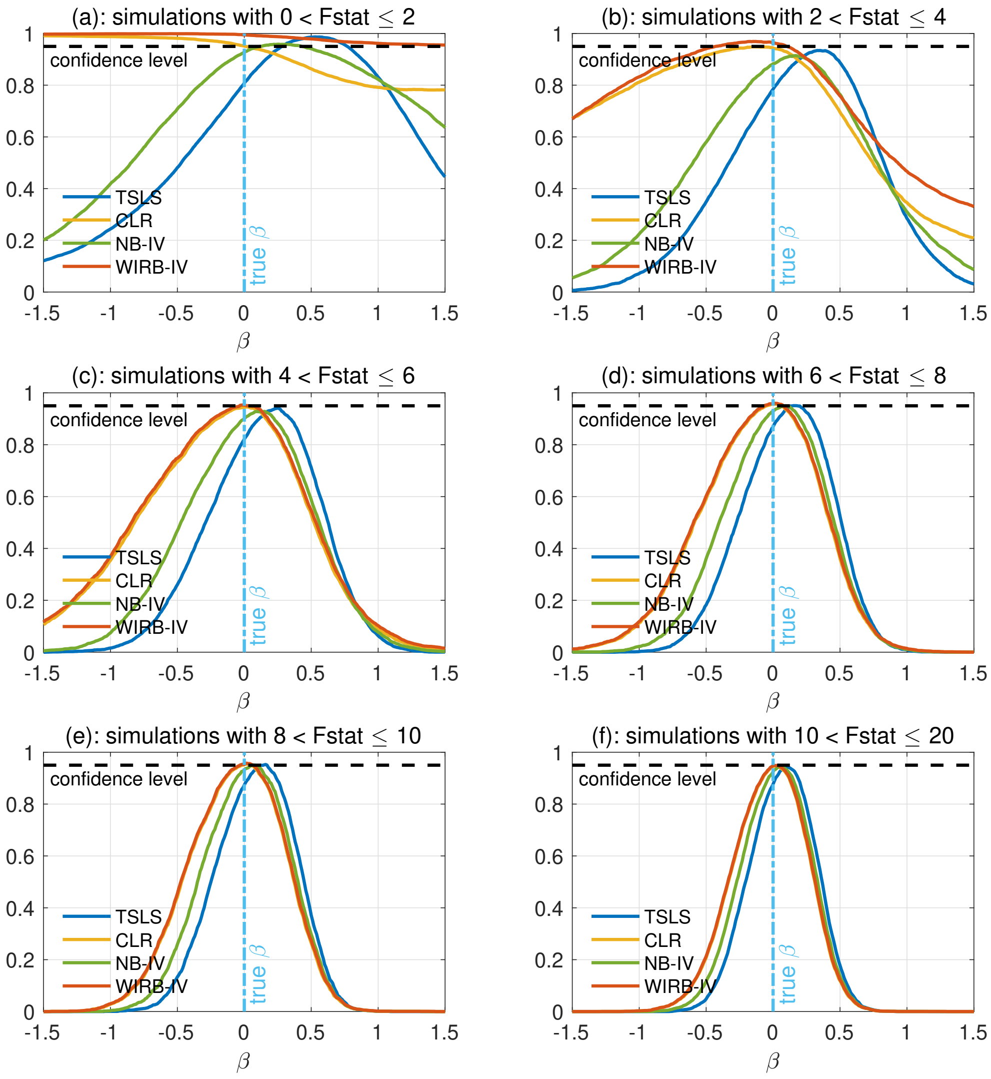

Figure 6: Frequency of inclusion in the 95-percent confidence (credible) interval based on (i) the asymptotic distribution of TSLS; (ii) the inversion of Moreira's (2003) conditional likelihood ratio test (CLR); (iii) the naive-Bayesian approach (NB-IV); and (iv) our weak-instrument-robust Bayesian approach (WIRB-IV). The results are based on $25,000$ simulations from model (1)-(2) with $T = 250$, $\beta = 0$, $\varepsilon \sim N(0, I_T)$, $\nu \sim N(0, I_T)$, $\pi \sim \mathcal{N}(0, s^2 I_k)$, $s \sim Uniform(0, 0.25)$, $\delta = 0.75$ and $k = 5$.

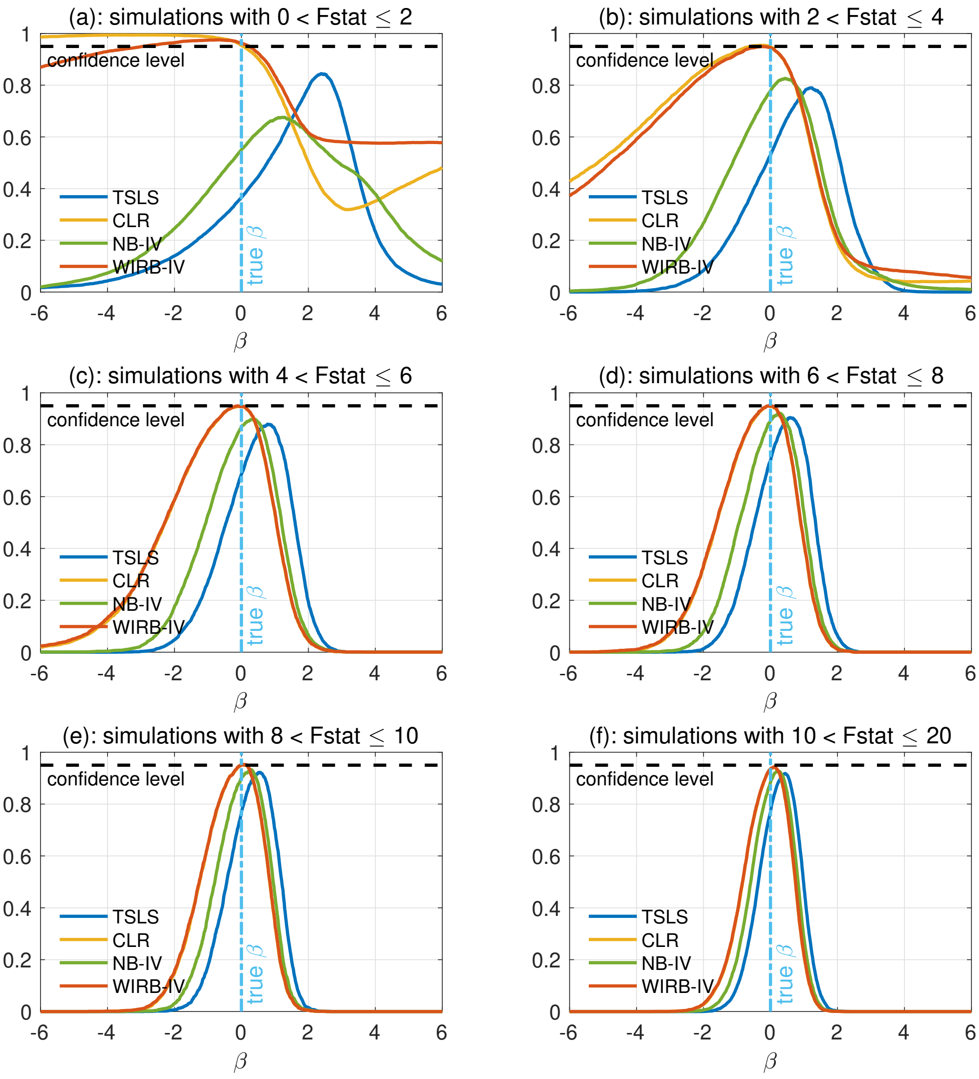

Figure 7: Frequency of inclusion in the 95-percent confidence (credible) interval based on (i) the asymptotic distribution of TSLS; (ii) the inversion of Moreira's (2003) conditional likelihood ratio test (CLR); (iii) the naive-Bayesian approach (NB-IV); and (iv) our weak-instrument-robust Bayesian approach (WIRB-IV). The results are based on $25,000$ simulations from model (1)-(2) with $T = 250$, $\beta = 0$, $\varepsilon \sim N(0, I_T)$, $\nu \sim N(0, I_T)$, $\pi \sim \mathcal{N}(0, s^2 I_k)$, $s \sim Uniform\,(0, 0.25)$, $\delta = 3$ and $k = 5$.

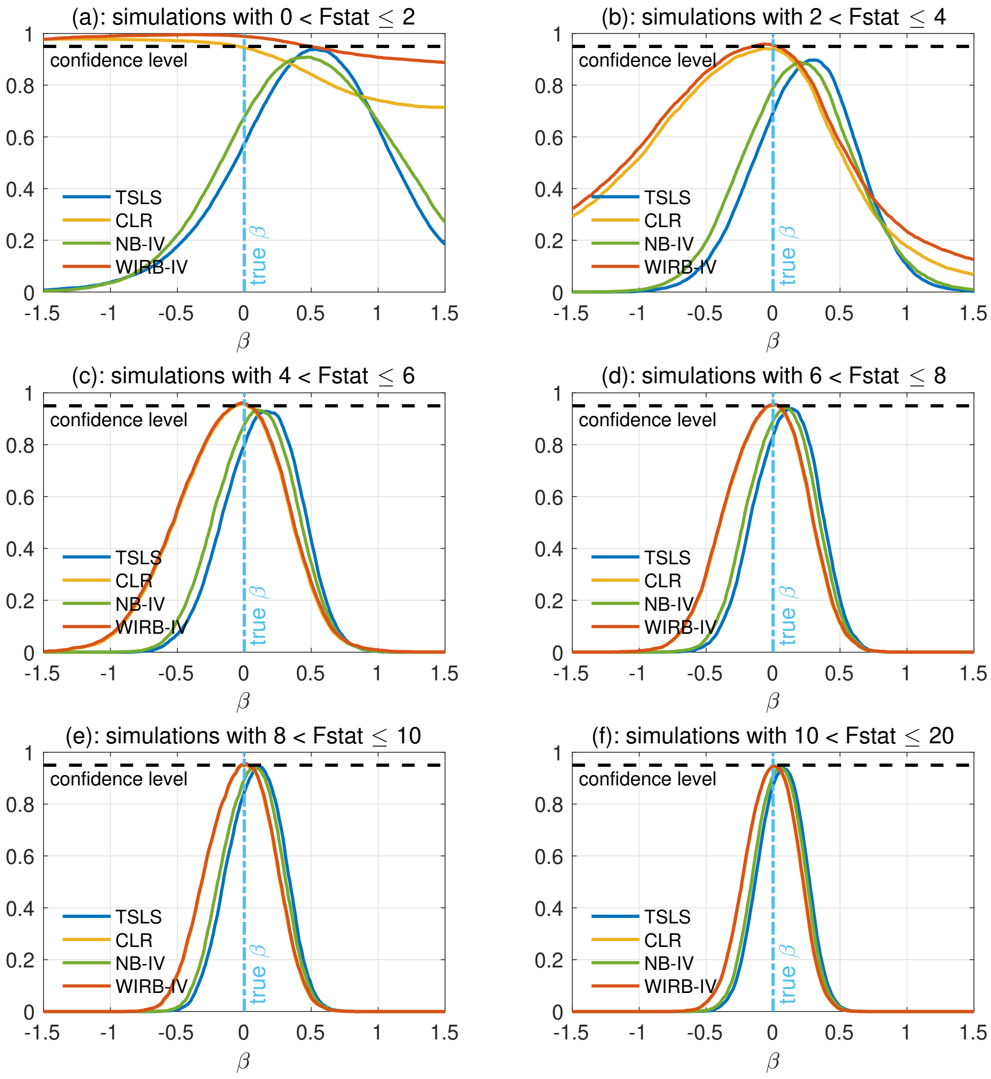

Figure 8: Frequency of inclusion in the 95-percent confidence (credible) interval based on (i) the asymptotic distribution of TSLS; (ii) the inversion of Moreira's (2003) conditional likelihood ratio test (CLR); (iii) the naive-Bayesian approach (NB-IV); and (iv) our weak-instrument-robust Bayesian approach (WIRB-IV). The results are based on $25,000$ simulations from model (1)-(2) with $T = 250$, $\beta = 0$, $\varepsilon \sim N\left(0, I_T\right)$, $\nu \sim N\left(0, I_T\right)$, $\pi \sim \mathcal{N}(0, s^2 I_k)$, $s \sim Uniform\left(0, 0.25\right)$, $\delta = 0.75$ and $k = 10$.

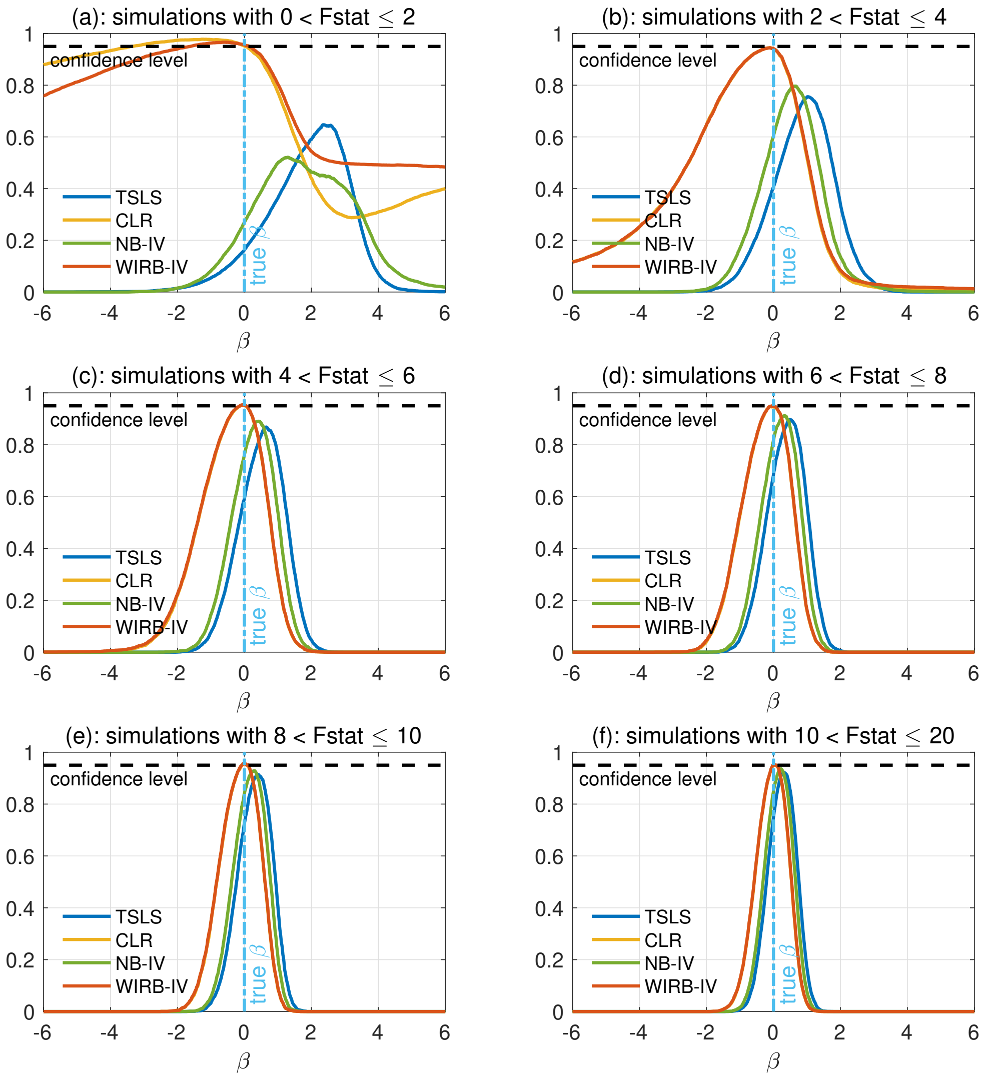

Figure 9: Frequency of inclusion in the 95-percent confidence (credible) interval based on (i) the asymptotic distribution of TSLS; (ii) the inversion of Moreira's (2003) conditional likelihood ratio test (CLR); (iii) the naive-Bayesian approach (NB-IV); and (iv) our weak-instrument-robust Bayesian approach (WIRB-IV). The results are based on $25,000$ simulations from model (1)-(2) with $T = 250$, $\beta = 0$, $\varepsilon \sim N(0, I_T)$, $\nu \sim N(0, I_T)$, $\pi \sim \mathcal{N}(0, s^2 I_k)$, $s \sim Uniform(0, 0.25)$, $\delta = 3$ and $k = 10$.

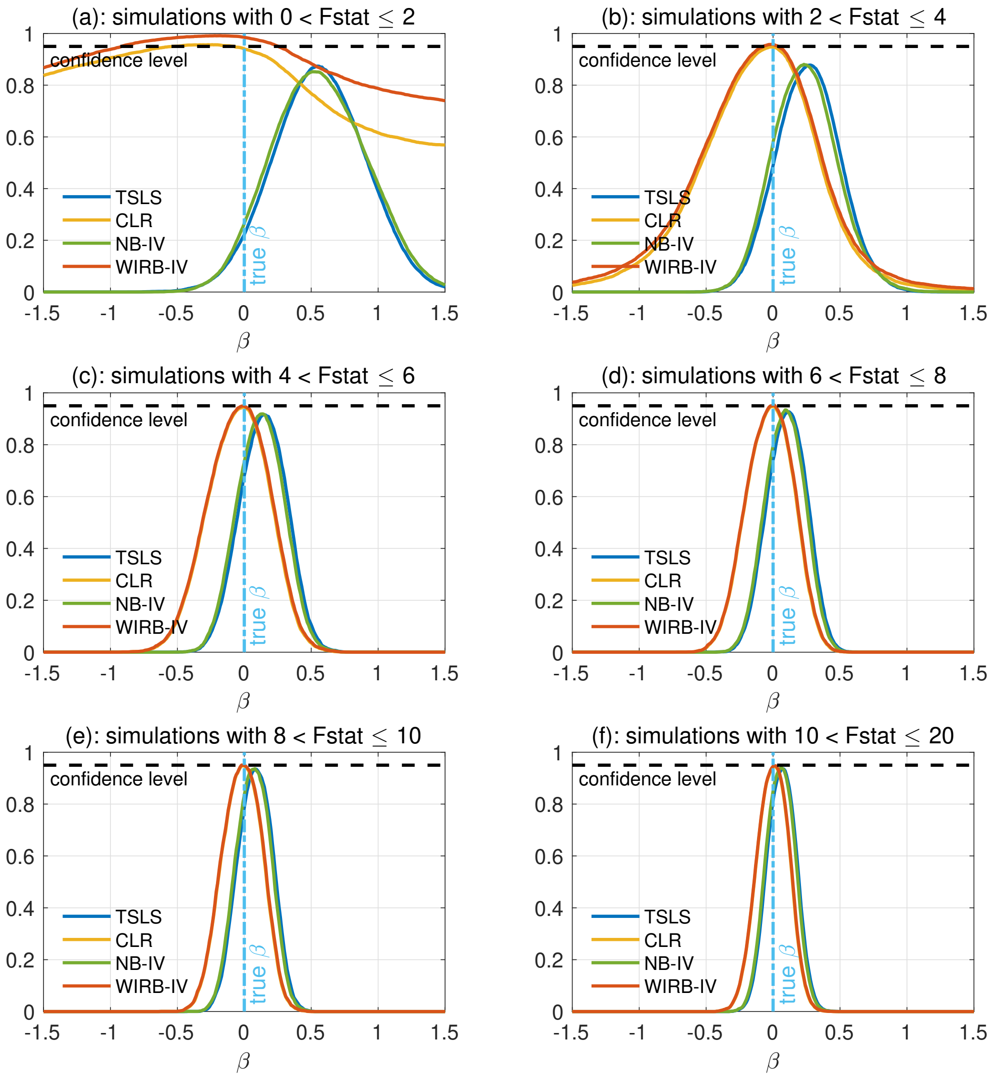

Figure 10: Frequency of inclusion in the 95-percent confidence (credible) interval based on (i) the asymptotic distribution of TSLS; (ii) the inversion of Moreira's (2003) conditional likelihood ratio test (CLR); (iii) the naive-Bayesian approach (NB-IV); and (iv) our weak-instrument-robust Bayesian approach (WIRB-IV). The results are based on $25,000$ simulations from model (1)-(2) with $T = 250$, $\beta = 0$, $\varepsilon \sim N(0, I_T)$, $\nu \sim N(0, I_T)$, $\pi \sim \mathcal{N}(0, s^2 I_k)$, $s \sim Uniform(0, 0.25)$, $\delta = 0.75$ and $k = 25$.

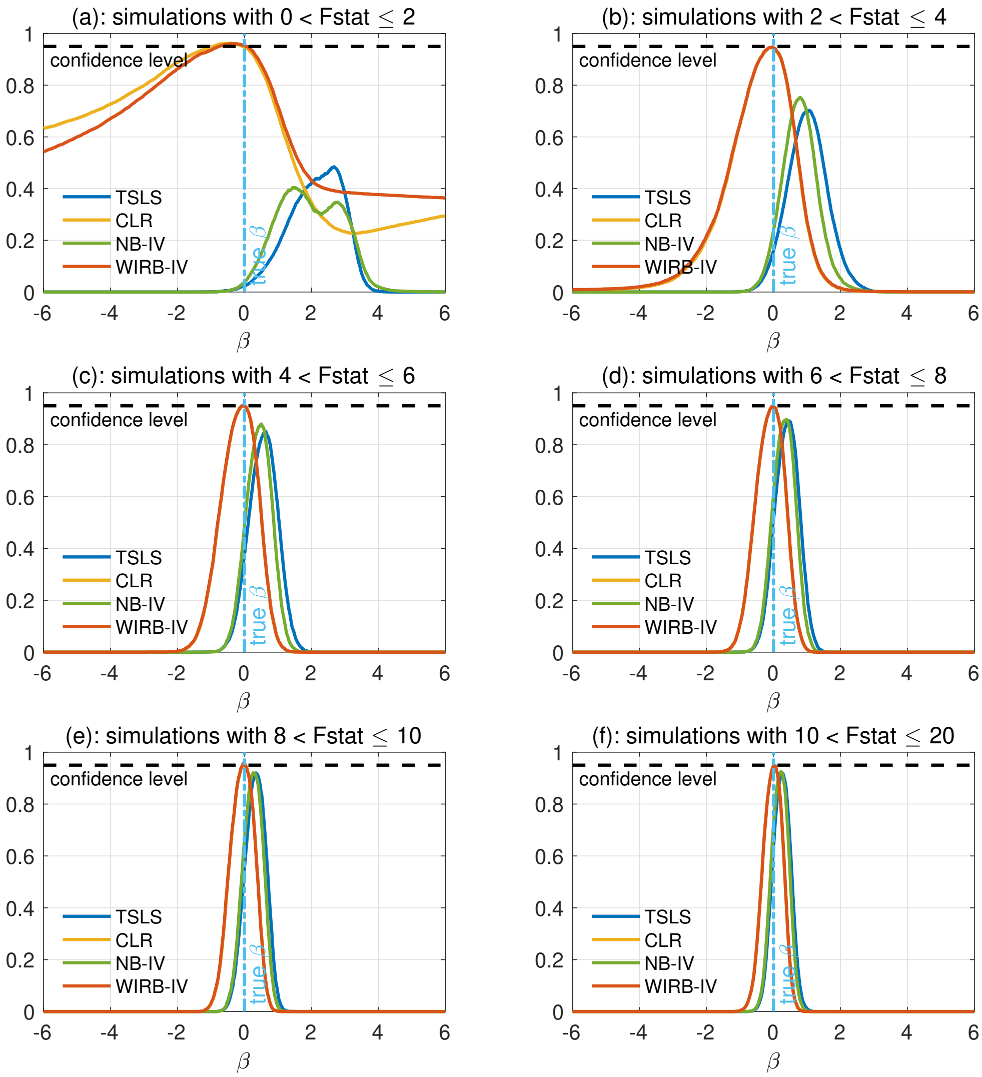

Figure 11: Frequency of inclusion in the 95-percent confidence (credible) interval based on (i) the asymptotic distribution of TSLS; (ii) the inversion of Moreira's (2003) conditional likelihood ratio test (CLR); (iii) the naive-Bayesian approach (NB-IV); and (iv) our weak-instrument-robust Bayesian approach (WIRB-IV). The results are based on $25,000$ simulations from model (1)-(2) with $T = 250$, $\beta = 0$, $\varepsilon \sim N(0, I_T)$, $\nu \sim N(0, I_T)$, $\pi \sim \mathcal{N}(0, s^2 I_k)$, $s \sim Uniform(0, 0.25)$, $\delta = 3$ and $k = 25$.

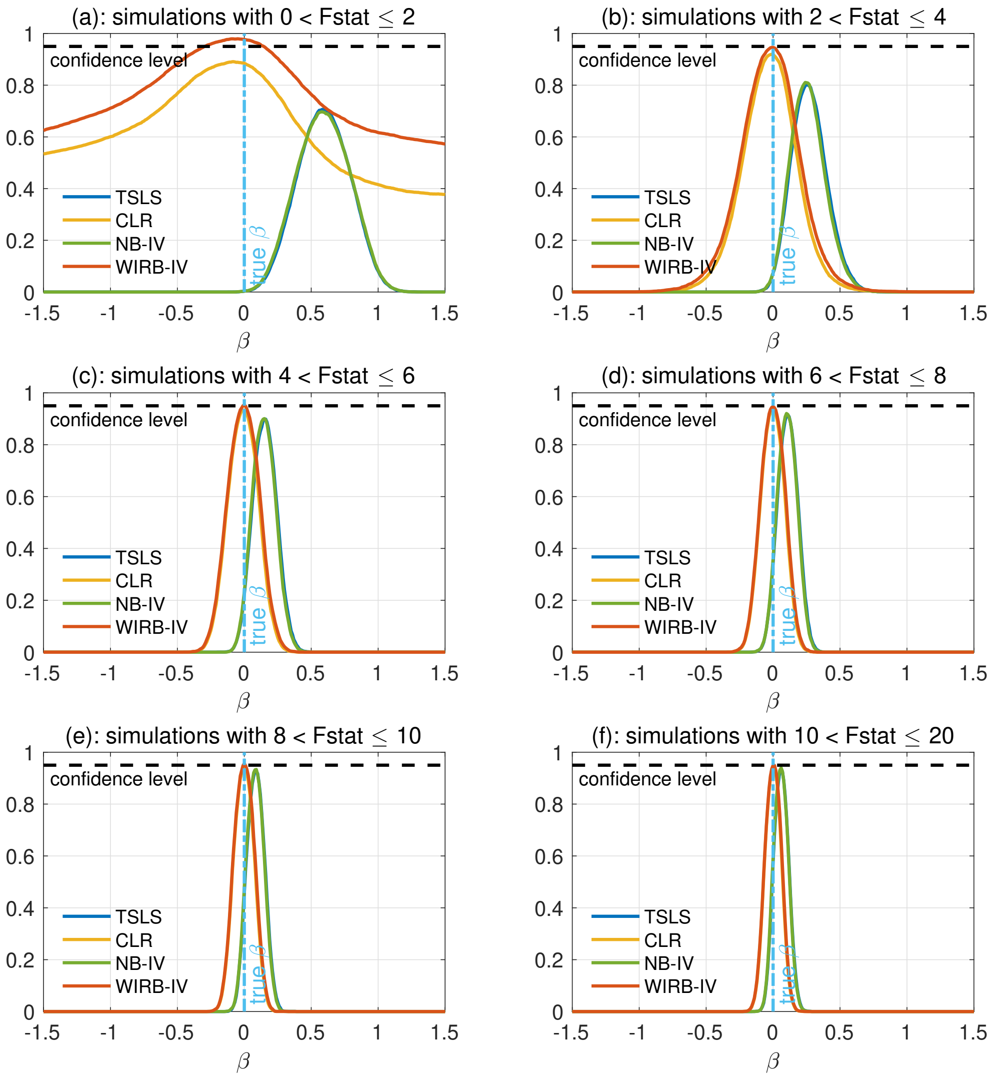

Figure 12: Frequency of inclusion in the 95-percent confidence (credible) interval based on (i) the asymptotic distribution of TSLS; (ii) the inversion of Moreira's (2003) conditional likelihood ratio test (CLR); (iii) the naive-Bayesian approach (NB-IV); and (iv) our weak-instrument-robust Bayesian approach (WIRB-IV). The results are based on $25,000$ simulations from model (1)-(2) with $T = 250$, $\beta = 0$, $\varepsilon \sim N(0, I_T)$, $\nu \sim N(0, I_T)$, $\pi \sim \mathcal{N}(0, s^2 I_k)$, $s \sim Uniform(0, 0.25)$, $\delta = 0.75$ and $k = 100$.

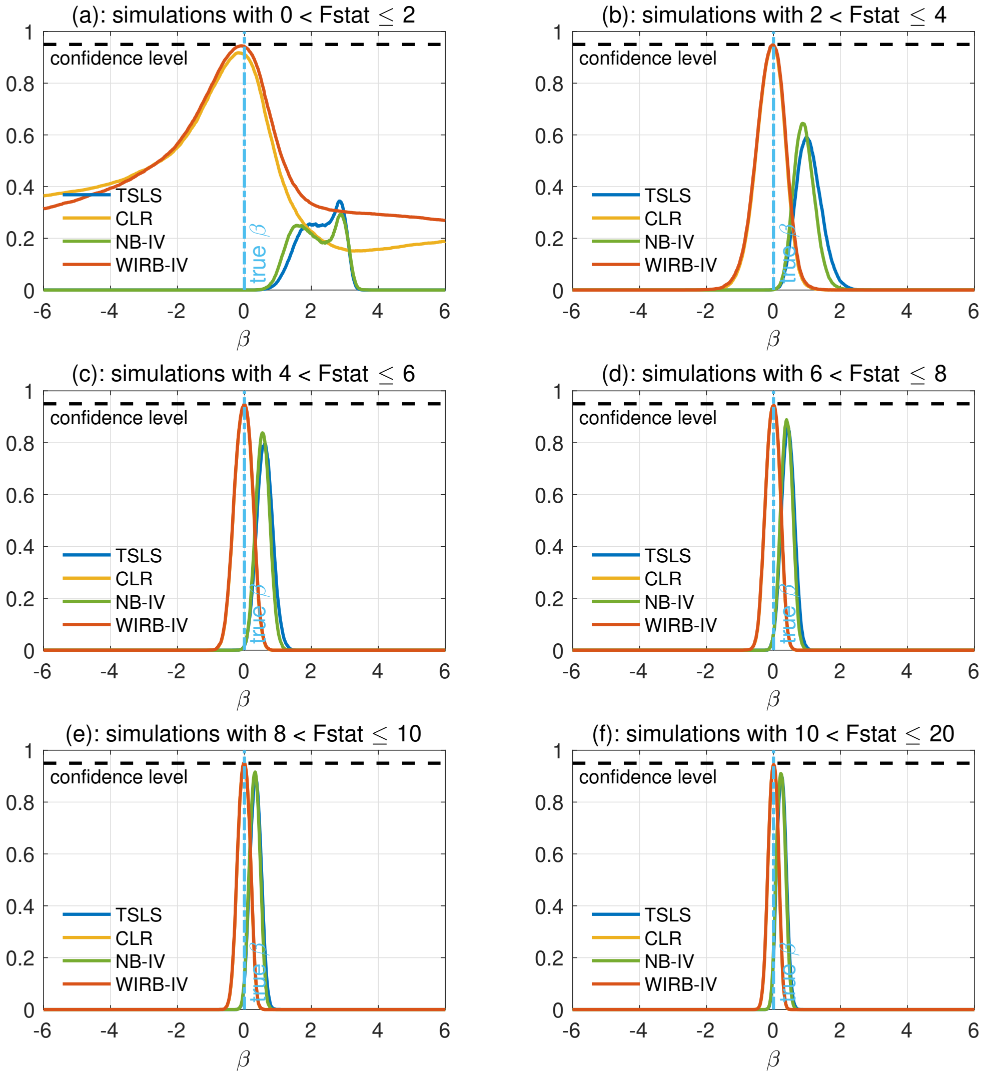

Figure 13: Frequency of inclusion in the 95-percent confidence (credible) interval based on (i) the asymptotic distribution of TSLS; (ii) the inversion of Moreira's (2003) conditional likelihood ratio test (CLR); (iii) the naive-Bayesian approach (NB-IV); and (iv) our weak-instrument-robust Bayesian approach (WIRB-IV). The results are based on $25,000$ simulations from model (1)-(2) with $T = 250$, $\beta = 0$, $\varepsilon \sim N(0, I_T)$, $\nu \sim N(0, I_T)$, $\pi \sim \mathcal{N}(0, s^2 I_k)$, $s \sim Uniform(0, 0.25)$, $\delta = 3$ and $k = 100$.

ANDREWS, I. AND T. B. ARMSTRONG (2017): "Unbiased instrumental variables estimation under known first-stage sign," *Quantitative Economics*, 8, 479–503.

ANDREWS, I., J. H. STOCK, AND L. SUN (2019): "Weak Instruments in Instrumental Variables Regression: Theory and Practice," *Annual Review of Economics*, 11, 727–753.

ANGRIST, J. D., G. W. IMBENS, AND A. B. KRUEGER (1999): "Jackknife Instrumental Variables Estimation," *Journal of Applied Econometrics*, 14, 57–67.

ANGRIST, J. D. AND A. B. KRUEGER (1991): "Does Compulsory School Attendance Affect Schooling and Earnings?" *The Quarterly Journal of Economics*, 106, 979–1014.

——— (1995): "Split-Sample Instrumental Variables Estimates of the Return to Schooling," *Journal of Business & Economic Statistics*, 13, 225–235.

BAI, J. AND S. NG (2009): "Selecting Instrumental Variables in a Data Rich Environment," *Journal of Time Series Econometrics*, 1, 1–34.

——— (2010): "Instrumental Variables Estimation in a Data Rich Environment," *Econometric Theory*, 26, 1577–1606.

BEKKER, P. A. (1994): "Alternative Approximations to the Distributions of Instrumental Variable Estimators," *Econometrica*, 62, 657–681.

BELLONI, A., D. CHEN, V. CHERNOZHUKOV, AND C. HANSEN (2012): "Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain," *Econometrica*, 80, 2369–2429.

BOUND, J., D. A. JAEGER, AND R. M. BAKER (1995): "Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogeneous Explanatory Variable is Weak," *Journal of the American Statistical Association*, 90, 443–450.

CARD, D. (1999): "The causal effect of education on earnings," in *Handbook of Labor Economics*, ed. by O. Ashenfelter and D. Card, Elsevier, vol. 3 of *Handbook of Labor Economics*, chap. 30, 1801–1863.

——— (2001): "Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems," *Econometrica*, 69, 1127–1160.

CARRASCO, M. (2012): "A Regularization Approach to the Many Instruments Problem," *Journal of Econometrics*, 170, 383–398.

CARRIERO, A., G. KAPETANIOS, AND M. MARCELLINO (2020): "A Shrinkage Instrumental Variable Estimator for Large Datasets," *L'Actualité Économique: Revue d'Analyse Économique*, 91, 67–87, special Issue in Honor of Jean-Marie Dufour.

CHAMBERLAIN, G. (2007): "Decision Theory Applied to an Instrumental Variables Model," *Econometrica*, 75, 609–652.

CHAMBERLAIN, G. AND G. IMBENS (2004): "Random Effects Estimators with many Instrumental Variables," *Econometrica*, 72, 295–306.

CHAO, J. C. AND P. C. B. PHILLIPS (1998): "Posterior distributions in limited information analysis of the simultaneous equations model using the Jeffreys prior," *Journal of Econometrics*, 87, 49–86.

CHAO, J. C. AND N. R. SWANSON (2005): "Consistent Estimation with a Large Number of Weak Instruments," *Econometrica*, 73, 1673–1692.

COX, D. R. AND N. REID (1987): "Parameter Orthogonality and Approximate Conditional Inference," *Journal of the Royal Statistical Society. Series B (Methodological)*, 49, 1–39.

CRUDU, F., G. MELLACE, AND Z. SÁNDOR (2021): "Inference In Instrumental Variable Models With Heteroskedasticity And Many Instruments," *Econometric Theory*, 37, 281–310.

CRUZ, L. M. AND M. J. MOREIRA (2005): "On the Validity of Econometric Techniques with Weak Instruments: Inference on Returns to Education Using Compulsory School Attendance Laws," *The Journal of Human Resources*, 40, 393–410.

DRÈZE, J. H. (1976): "Bayesian Limited Information Analysis of the Simultaneous Equations Model," *Econometrica*, 44, 1045–1075.

DUFOUR, J.-M. (1997): "Some Impossibility Theorems in Econometrics, with Applications to Structural and Dynamic Models," *Econometrica*, 65, 1365–1388.

GIANNONE, D., M. LENZA, AND G. E. PRIMICERI (2021): "Economic Predictions With Big Data: The Illusion of Sparsity," *Econometrica*, 89, 2409–2437.

GLESER, L. J. (1976): "A Canonical Representation for the Noncentral Wishart Distribution Useful for Simulation," *Journal of the American Statistical Association*, 71, 690–695.

GRILICHES, Z. (1977): "Estimating the Returns to Schooling: Some Econometric Problems," *Econometrica*, 45, 1–22.

GUNDERSON, M. AND P. OREOPOLOUS (2020): "Returns to education in developed countries," in *The Economics of Education (Second Edition)*, ed. by S. Bradley and C. Green, Academic Press, chap. 3, 39–51, second edition ed.

HANSEN, C., J. HAUSMAN, AND W. NEWEY (2008): "Estimation With Many Instrumental Variables," *Journal of Business & Economic Statistics*, 26, 398–422.

HOOGERHEIDE, L., F. KLEIBERGEN, AND H. K. VAN DIJK (2007a): "Natural conjugate priors for the instrumental variables regression model applied to the Angrist-Krueger data," *Journal of Econometrics*, 138, 63–103.

HOOGERHEIDE, L. AND H. VAN DIJK (2006): "A reconsideration of the Angrist-Krueger analysis on returns to education," Econometric Institute Research Papers EI 2006-15.

HOOGERHEIDE, L. F., J. F. KAASHOEK, AND H. K. VAN DIJK (2005): "On the shape of posterior densities and credible sets in instrumental variable regression models with reduced rank: An application of flexible sampling methods using neural networks," LIDAM Discussion Papers CORE 2005029.

HOOGERHEIDE, L. F., J. F. KAASHOEK, AND H. K. VAN DIJK (2007b): "On the shape of posterior densities and credible sets in instrumental variable regression models with reduced rank: An application of flexible sampling methods using neural networks," *Journal of Econometrics*, 139, 154–180, endogeneity, instruments and identification.

IMBENS, G. AND P. ROSENBAUM (2005): "Robust, Accurate Confidence Intervals with a Weak Instrument: Quarter of Birth and Education," *Journal of the Royal Statistical Society Series A*, 168, 109–126.

KAPETANIOS, G., L. KHALAF, AND M. MARCELLINO (2016): "Factor-Based Identification-Robust Inference in IV Regressions," *Journal of Applied Econometrics*, 31, 821–842.

KAPETANIOS, G. AND M. MARCELLINO (2010): "Factor-GMM Estimation with Large Sets of Possibly Weak Instruments," *Computational Statistics and Data Analysis*, 54, 2655–2675.

KLEIBERGEN, F. (2002): "Pivotal Statistics for Testing Structural Parameters in Instrumental Variables Regression," *Econometrica*, 70, 1781–1803.

——— (2005): "Testing Parameters in GMM Without Assuming that They Are Identified," *Econometrica*, 73, 1103–1123.

KLEIBERGEN, F. AND H. K. VAN DIJK (1998): "Bayesian Simultaneous Equations Analysis Using Reduced Rank Structures," *Econometric Theory*, 14, 701–743.

KLEIBERGEN, F. AND E. ZIVOT (2003): "Bayesian and classical approaches to instrumental variable regression," *Journal of Econometrics*, 114, 29–72.

LEE, D. S., J. MCCRARY, M. J. MOREIRA, AND J. PORTER (2022): "Valid t-Ratio Inference for IV," *American Economic Review*, 112, 3260–3290.

LEWIS, D. J. AND K. MERTENS (2025): "A Robust Test for Weak Instruments for 2SLS withMultiple Endogenous Regressors," Unpublished manuscript, University College London.

MADDALA, G. S. (1976): "Weak Priors and Sharp Posteriors in Simultaneous Equation Models," *Econometrica*, 44, 345–351.

MATSUSHITA, Y. AND T. OTSU (2024): "A Jackknife Lagrange Multiplier Test With Many Weak Instruments," *Econometric Theory*, 40, 447–470.

MIKUSHEVA, A. (2010): "Robust Confidence Sets in the Presence of Weak Instruments," *Journal of Econometrics*, 157, 236–247.

——— (2020): "Instrumental Variables and Weak Instruments: A Survey," in *Advances in Economics and Econometrics: Eleventh World Congress, Volume III*, ed. by R. Blundell, W. Dickens, and C. Meghir, Cambridge: Cambridge University Press.

MIKUSHEVA, A. AND B. P. POI (2006): "Tests and Confidence Sets with Correct Size in the Simultaneous Equations Model with Potentially Weak Instruments," *The Stata Journal*, 6, 335–347.

MIKUSHEVA, A. AND L. SUN (2022): "Inference with Many Weak Instruments," *The Review of Economic Studies*, 89, 2663–2686.

——— (2024): "Weak identification with many instruments," *The Econometrics Journal*, 27(2), C1–C28.

MONTIEL OLEA, J. L. (2020): "Admissible, Similar Tests: A Characterization," *Econometric Theory*, 36, 347–366.

MONTIEL OLEA, J. L. AND C. PFLUEGER (2013): "A Robust Test for Weak Instruments," *Journal of Business & Economic Statistics*, 31, 358–369.

MOREIRA, M. J. (2003): "A Conditional Likelihood Ratio Test for Structural Models," *Econometrica*, 71, 1027–1048.

MÜLLER, U. K. AND A. NORETS (2016a): "Coverage Inducing Priors in Nonstandard Inference Problems," *Journal of the American Statistical Association*, 111, 1233–1241.

——— (2016b): "Credibility of Confidence Sets in Nonstandard Econometric Problems," *Econometrica*, 84, 2183–2213.

NELSON, C. R. AND R. STARTZ (1990a): "Some Further Results on the Exact Small Sample Properties of the Instrumental Variable Estimator," *Econometrica*, 58, 967–976.

——— (1990b): "The Distribution of the Instrumental Variables Estimator and Its t-Ratio When the Instrument Is a Poor One," *The Journal of Business*, 63, 125–140.

ROBERT, C. P. AND G. CASELLA (2004): *Monte Carlo Statistical Methods*, Springer Texts in Statistics, New York: Springer-Verlag, 2nd ed.

ROTHENBERG, T. J. (1984): "Approximating the distributions of econometric estimators and test statistics," Elsevier, vol. 2 of *Handbook of Econometrics*, chap. 15, 881–935.

STAIGER, D. AND J. H. STOCK (1997): "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 65, 557–586.

STOCK, J. H., J. H. WRIGHT, AND M. YOGO (2002): “A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments,” *Journal of Business & Economic Statistics*, 20, 518–529.

STOCK, J. H. AND M. YOGO (2005): “Testing for Weak Instruments in Linear IV Regression,” in *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, ed. by D. W. K. Andrews and J. H. Stock, Cambridge: Cambridge University Press, 80–108.

U.S. BUREAU OF LABOR STATISTICS (2024): “Education pays, 2023,” *Career Outlook*, suggested citation: ”Education pays, 2023,” *Career Outlook*, U.S. Bureau of Labor Statistics, April 2024.