Online Appendix to "The Causal Interpretation of Two-Stage Least Squares with Multiple Instrumental Variables"

Magne Mogstad, Alexander Torgovitsky, and Christopher Walters

March 2021

A Survey Results By Journal

Table A.1: *Multiple IVs by top economics journal*

	(1)	(2)	(3)	(4)
	All papers	Papers that use multiple IVs	Papers with more IVs than endogenous variables	Papers with more IVs than endogenous variables using 2SLS/GMM
American Economic Review	100% 44	39% 17	36% 16	34% 15
Quarterly Journal of Economics	100% 28	54% 15	46% 13	43% 12
Journal of Political Economy	100% 23	65% 15	48% 11	43% 10
Econometrica	100% 15	67% 10	60% 9	53% 8
Review of Economic Studies	100% 12	67% 8	67% 8	58% 7
All	100% 122	53% 65	47% 57	43% 52

This table displays the results of a survey on the use of instrumental variables separately for each top economics journal. Column (1) includes articles published between January 2000 and October 2018 containing the words "instrument" or "instrumental variable" in the abstract, title, or topic words and using at least one instrumental variable in an empirical application.

Table A.2: *Multiple IV papers by journal and relationship between their IVs*

	(1)	(2)	(3)	(4)
	Papers with more IVs than endogenous variables using 2SLS/GMM	Case **A**	Case **B**	Case **C**
American Economic Review	100% 15	60% 9	13% 2	27% 4
Quarterly Journal of Economics	100% 12	75% 9	8% 1	17% 2
Journal of Political Economy	100% 10	70% 7	0% 0	30% 3
Econometrica	100% 8	63% 5	38% 3	0% 0
Review of Economic Studies	100% 7	71% 5	14% 1	14% 1
All	100% 52	67% 35	13% 7	19% 10

*This table classifies the papers from column (4) of Table 1 into three cases of multiple instruments separately by journal. Case **A** are studies that use multiple economically distinct instruments. Case **B** are studies that use covariate interactions with a single instrument. Case **C** are studies that use multiple functions of a single instrument. These cases are mutually exclusive and exhaustive; some proportions do not sum to 100% due to rounding.*

Table A.3: *Multiple IV papers that use IVs separately by journal*

	(1)	(2)	(3)
	Papers that ever use IVs separately	Papers that do not control for omitted IVs	Papers that always use IVs separately and never control for omitted IVs
American Economic Review	7	6	2
Quarterly Journal of Economics	5	5	4
Journal of Political Economy	6	5	2
Econometrica	1	1	0
Review of Economic Studies	1	1	1
All	20	18	9

This table includes the subset of the papers from Table 1 that use multiple instruments and fit a separate 2SLS model using a single IV in at least one specification in the main body of the paper, separately by journal.

B Additional Examples of Assumptions IAM and PM

In this appendix, we consider three example papers from our survey, each of which combines multiple economically distinct IVs using 2SLS. In each example, we discuss the content of both Assumptions IAM and PM. We conclude in each that Assumption IAM is unlikely to hold, whereas the weaker Assumption PM is more plausible.

B.1 Thornton (2008)

The author evaluates an experiment in rural Malawi in which individuals who were screened for HIV were then randomly assigned incentives to receive their results. Two incentives were used: a cash-redeemable voucher and the distance to the nearest results center. Thornton uses these randomly assigned incentives as instruments for the decision to obtain the results, with the goal of estimating the causal effect of learning HIV status on the demand for condoms.

Assumption IAM would require an individual's decision to obtain their HIV test results to either be influenced more by the voucher, or by the distance to the results center. It is not difficult to imagine two individuals who similarly value learning their HIV status, but who differ in their preferences over distance versus a monetary incentive. For example, a low-wage worker with a lower opportunity cost of time might be more affected by a voucher than by the distance to the center. On the other hand, a high-wage worker with a higher opportunity cost might value the reduced travel time to a closer center more than a voucher.

While Assumption IAM monotonicity is hard to justify, Assumption PM is likely to hold. Keeping distance fixed, a voucher incentive should increase the likelihood that any individual obtains his or her results. Keeping the voucher incentive fixed, being randomly assigned a closer results facility should decrease the opportunity cost and thus also increase the likelihood of obtaining the test results.

B.2 Currie and Moretti (2003)

The authors examine the impact of a mother's educational attainment on her infant's health. As instruments for mother's education, the authors use the number of two- and four-year colleges present in the mother's county of residence when she was 17 years old. The argument is that the availability of college may have induced some women to obtain more education.

Assumption IAM would require all mothers to be influenced more by the presence of a two-year college or by the presence of a four-year college when deciding how much education to obtain. This is unlikely. For example, two-year colleges may provide a cheaper option, while four-year colleges may provide a better education. There is likely to be substantial heterogeneity in preferences over the cost and quality of education.

However, Assumption PM is a more reasonable assumption. A mother's educational attainment should increase as her opportunities for education increase. If this is true, then the *ceteris paribus* impact of the presence of a four-year college on attained education would be positive regardless of whether there is also a two-year college, and vice-versa. This is all that is required for Assumption PM to be satisfied.

B.3 Dippel (2014)

The author considers the long-term effects on economic growth of forcibly integrating Native American communities in the United States. As instruments for integration, the author uses the value of gold and silver mining activity in these communities. The rationale behind these instruments is that when land was found to be more valuable, the federal government would free up larger portions of the land by forming fewer, more concentrated reservations.

Assumption IAM would require the likelihood of forced integration across all tribes and reservations to either be affected more by the value of silver activity or by the value of gold activity. This would not hold if the officials responsible for forced integration had different beliefs about the value of each metal. This could occur if these beliefs vary by the location of the reservation or by the official responsible, among other reasons. In contrast, Assumption PM follows the logic of the author's argument: holding the value of one metal fixed, increasing the value of the other increases the likelihood of forced integration.

C Tests about the Signs of the 2SLS Weights

C.1 Estimation

Suppose that we exclude a priori the possibility that $\pi_{1\text{c}} = 0$ or $\pi_{2\text{c}} = 0$. Then Proposition 5 implies that the signs of the 2SLS weights in the case with two binary instruments are determined by the signs of

$$\theta_j(x) \equiv \mathbb{P}[D_i = 1 | Z_{i,j} = 1, X_i = x] - \mathbb{P}[D_i = 1 | Z_{i,j} = 0, X_i = x] \quad \text{for } j = 1, 2,$$

where we have conditioned on covariates, X_i, since they are included in our application in Section 5. To develop our tests, we assume that the conditional probability of treatment given $Z_{i,j}$ and X_i is additively separable in $Z_{i,j}$. This implies that $\theta_j(x) \equiv \theta_j$ does not depend on x, and that θ_j is identified as the population regression coefficient on $Z_{i,j}$ in a regression of D_i on $Z_{i,j}$ and X_i.[1]

We estimate $\theta \equiv (\theta_1, \theta_2)$ with two ordinary least squares estimators, $\hat{\theta} \equiv (\hat{\theta}_1, \hat{\theta}_2)$, based on a sample of size n. We assume that these estimators are jointly asymptotically normal and denote the limiting variance–covariance matrix of $\sqrt{n}(\hat{\theta} - \theta)$ by Σ. Let $\hat{\Sigma}$ be a consistent estimator of Σ. Let σ_1^2 and σ_2^2 be the diagonal components of Σ, let σ_{12} be the off–diagonal component, and denote the corresponding components of $\hat{\Sigma}$ by $\hat{\sigma}_1^2$, $\hat{\sigma}_2^2$, and $\hat{\sigma}_{12}$, respectively.

C.2 Testing the Null Hypothesis of Positive Weights

We first consider tests of the null hypothesis that the weights are positive (non-negative), that is of

$$H_0^+ : \theta_1 \geq 0 \quad \text{and} \quad \theta_2 \geq 0,$$

against the complementary alternative.

The first and simplest approach is to treat $\theta_1 \geq 0$ and $\theta_2 \geq 0$ as separate hypotheses and then apply a Bonferroni correction. Letting $\hat{p}_1$ and $\hat{p}_2$ denote the p–values from the corresponding one–sided t–tests, the Bonferroni–corrected p–value is then

$$\hat{p}_0 \equiv \min\{2\hat{p}_1, 2\hat{p}_2, 1\}.$$

This test will typically be conservative.

The second approach is to consider the test statistic

$$\hat{T} \equiv \min_{j=1,2} \frac{\sqrt{n}\hat{\theta}_j}{\hat{\sigma}_j}, \tag{1}$$

that is, the minimum of the individual t–statistics, and reject H_0^+ if this quantity is too small.

[1] It is straightforward to extend the following to test joint hypotheses about $\theta_j(x)$ across different pre–specified x values. It is also possible to test joint hypotheses about $\theta_j(X_i)$ as a random variable using tools from the literature on conditional moment inequalities (e.g. Andrews and Shi, 2013; Chernozhukov, Lee, and Rosen, 2013; Armstrong, 2015; Chetverikov, 2018). We focus on the separable case as it enables the construction of simple tests which can be implemented easily in statistical software, and which require no additional choices or tuning parameters on the part of the researcher.

If H_0^+ is true, then the distribution of $\hat{T}$ has the following lower bound asymptotically:

$$\mathbb{P}\left[\hat{T} \leq t\right] = \mathbb{P}\left[\min_{j=1,2} \frac{\sqrt{n}(\hat{\theta}_j - \theta_j)}{\hat{\sigma}_j} + \frac{\sqrt{n}\theta_j}{\hat{\sigma}_j} \leq t\right] \leq \mathbb{P}\left[\min_{j=1,2} W_j \leq t\right], \tag{2}$$

where $W \equiv (W_1, W_2)$ is a bivariate normal distribution with mean zero, unit variances, and correlation $\sigma_{1,2}\sigma_1^{-1}\sigma_2^{-1}$. Thus, the test that rejects H_0^+ when $\hat{T}$ is smaller than the α quantile of the distribution of $\min_{j=1,2} W_j$ has size no greater than α. In the Monte Carlo simulation in Section C.4, we refer to this test as the "Mintest." Implementing it requires simulating the distribution of $\min_{j=1,2} W_j$ using its estimated correlation, $\hat{\sigma}_{12}\sigma_1^{-1}\hat{\sigma}_2^{-1}$.

The third approach uses the quasi–likelihood ratio statistic

$$\hat{Q} = \min_{t \geq 0} n\left(\hat{\theta} - t\right)' \hat{\Sigma}^{-1}\left(\hat{\theta} - t\right),$$

and rejects if $\hat{Q}$ is too large. Let $\hat{t}^\star$ denote the minimizer of this problem and let $\hat{k}^\star \in \{0, 1, 2\}$ denote the number of components of $\hat{t}^\star$ that are zero, that is, where the non-negativity constraint is binding. Cox and Shi (2019) show that the test that rejects when $\hat{Q}$ is larger than the $1-\alpha$ quantile of a chi–squared distribution with $\hat{k}^\star$ degrees of freedom controls size at level α. We call this the Cox–Shi test in Section C.4.

The fourth test is from Romano, Shaikh, and Wolf (2014, "RSW"), and also uses the test statistic $\hat{T}$ from (1). Their approach improves on the Mintest described above by estimating a $1-\beta$ joint confidence interval for $\min_{j=1,2} \sqrt{n}(\hat{\theta}_j - \theta_j)$ and using this to improve the coarse bound used in (2) to obtain a critical value. Both this first step and computing the resulting critical value require bootstrapping the linear regression estimators, $\hat{\theta}_1$ and $\hat{\theta}_2$. For a level α test, RSW recommend setting $\beta = \alpha/10$, which means that the number of bootstraps used in the first step confidence interval needs to be rather large to get an accurate approximation of the $1-\beta$ quantile. This can make the RSW test somewhat computationally demanding compared to the other three tests.

C.3 Testing the Null Hypothesis of Negative Weights

We also consider the opposite null hypothesis that one or more 2SLS weight is negative, that is, of

$$H_0^- : \theta_1 \leq 0 \quad \text{or} \quad \theta_2 \leq 0,$$

against the complementary alternative. We use an intersection–union test (IUT) based on the theory in Berger (1982). In the current context, the IUT argument is simple: reject H_0^- at level α if both $\theta_1 \leq 0$ and $\theta_2 \leq 0$ are rejected at level α using one-sided t–tests. This controls size because the probability that both $\theta_1 \leq 0$ and $\theta_2 \leq 0$ are rejected under the null is by construction smaller than the probability that either $\theta_1 \leq 0$ or $\theta_2 \leq 0$ are rejected. Perhaps more surprisingly, Berger (1982, Theorem 2) provides conditions under which the IUT test is size–correct, which is confirmed in our simulations. See Berger and Hsu (1996, Section 3) and Casella and Berger (2002, Section 8.3) for more detail.

C.4 A Monte Carlo Simulation

The Monte Carlo simulation has the following data generating process. The group shares are set at $\pi_{\mathrm{at}} = 2/12$, $\pi_{\mathrm{ec}} = 1/12$, $\pi_{\mathrm{rc}} = 1/12$, $\pi_{\mathrm{nt}} = 2/12$, $\pi_{\mathrm{1c}} = 5/12$, and $\pi_{\mathrm{2c}} = 1/12$. First, $Z_{i,2} \in \{0, 1\}$ is drawn with probability 1/2. Then, $Z_{i,1}$ is drawn conditional on $Z_{i,2}$ with probability

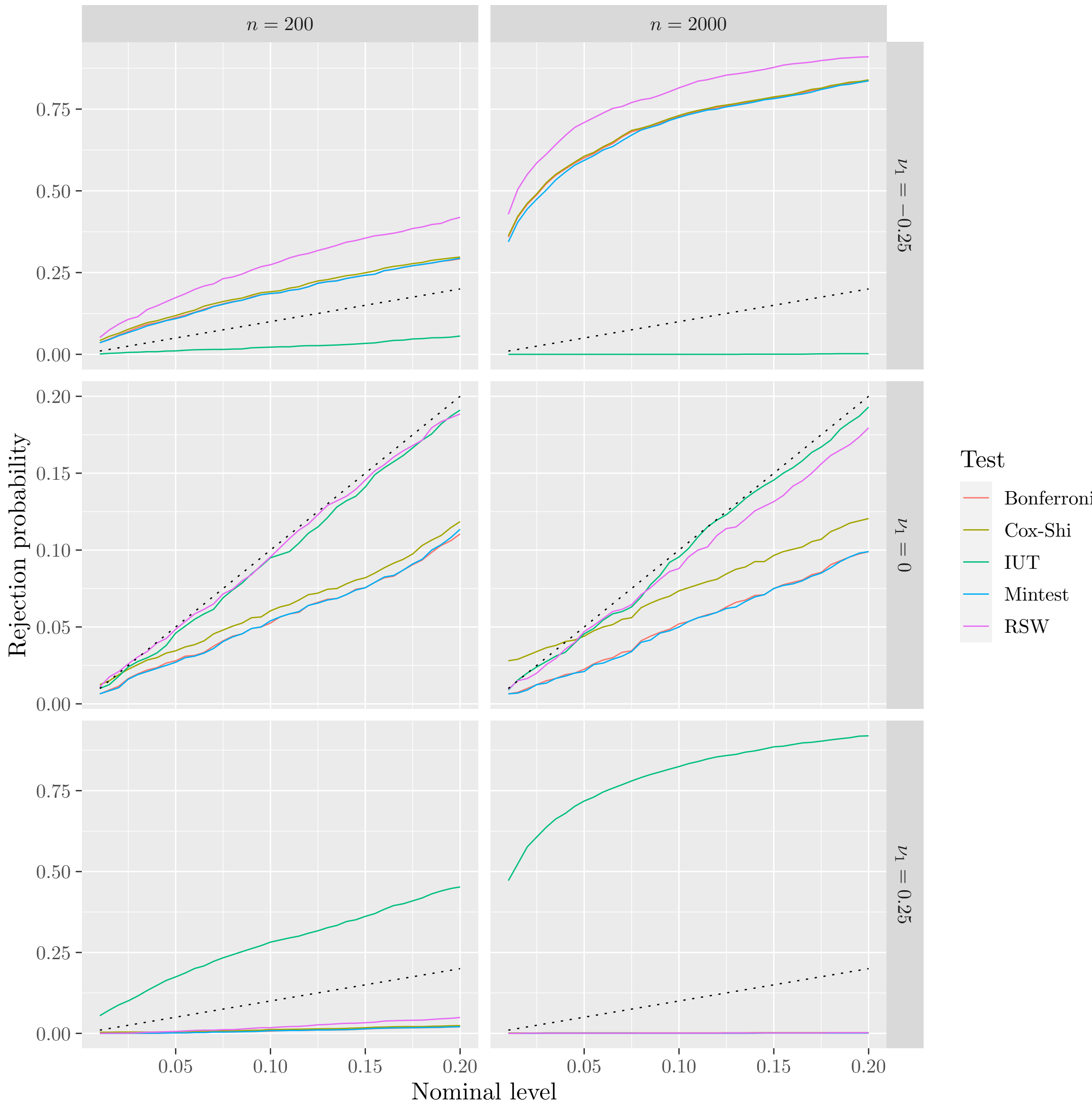
$$\mathbb{P}[Z_{i,1} = 1 | Z_{i,2} = z_2] = \Phi(\nu_0(1 - z_2) + \nu_1 z_2),$$

where Φ is the standard normal distribution function, and ν_0, ν_1 are design parameters. The parameter ν_0 is set such that H_0^+ is true if and only if $\nu_1 \geq 0$, H_0^- is true if and only if $\nu_1 \leq 0$, and both null hypotheses are true when $\nu_1 = 0$.

Figure C.1 shows QQ–plots of the p–values from our various tests against the uniform distribution for three values of ν_1.[2] The middle row with $\nu_1 = 0$ shows that all tests control size when both H_0^+ and H_0^- are true, with the IUT and RSW tests being closest to size–correct. When $\nu_1 = -.25$, so that H_0^- is true, the three simpler-to-implement tests (Bonferroni, Mintest, Cox–Shi) all have roughly the same power, while the RSW test is substantially more powerful. Power for the IUT test when $\nu_1 = .25$ is difficult to gauge, since there is no point of comparison, but one might expect that it is fairly good given its performance at the boundary of the null hypothesis ($\nu_1 = 0$). Figure C.2 reports power curves for a 5% level test which confirm the superior power of the RSW test.

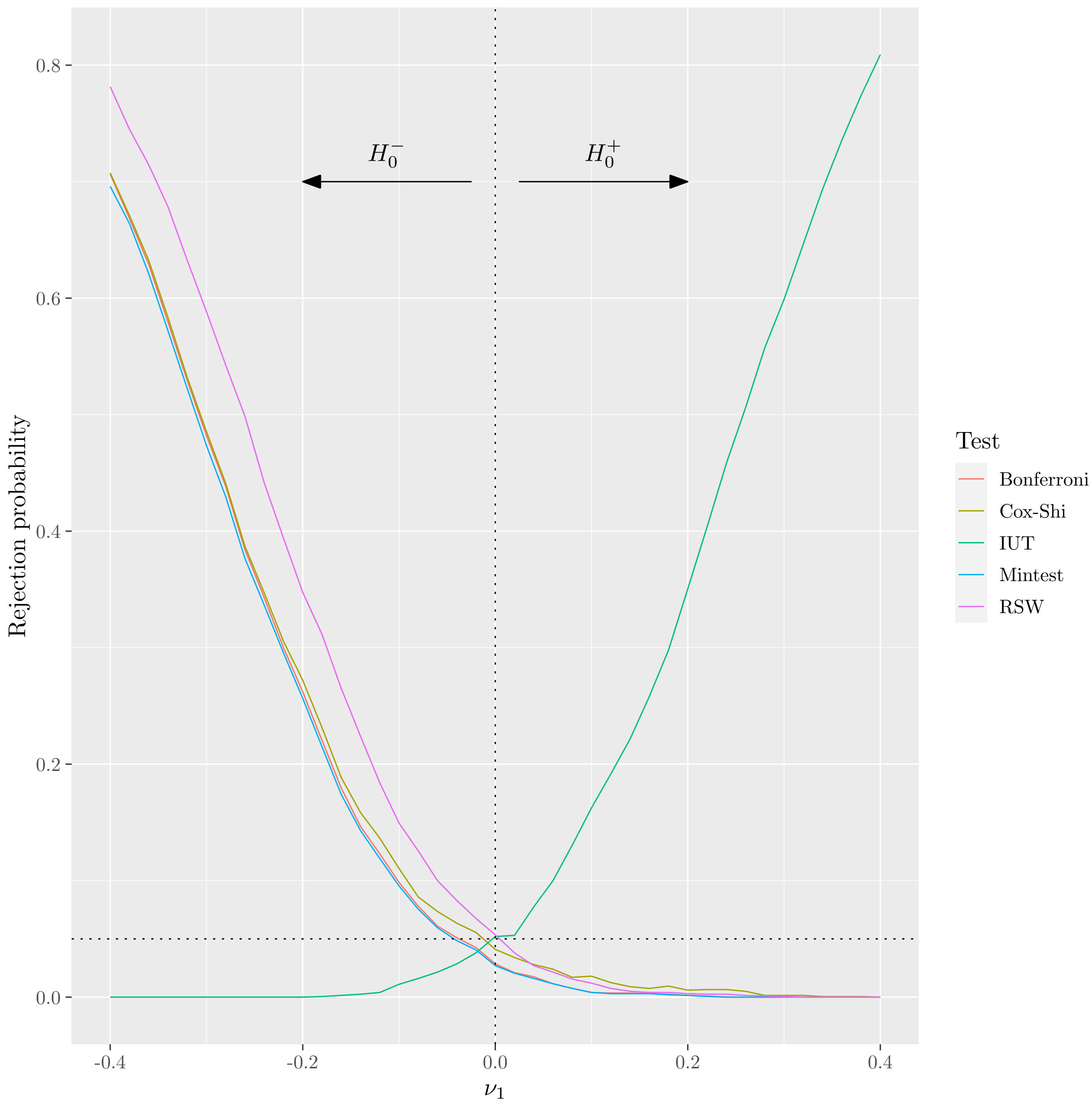
[2] All simulations are based on 2,000 replications. We used 2,000 bootstrap draws for the RSW test.

Figure C.1: *Size and power for five tests*



The dotted line is the 45 degree line. The top row ($\nu_1 = -.25$) is where H_0^- is true, the bottom row is where H_0^+ is true, and both hypothesis are true in the middle row.

Figure C.2: *Power curves for five tests*



The dotted horizontal line indicates the nominal level of .05. The dotted vertical line indicates the boundary between where H_0^+ and H_0^- are true. The sample size is $n = 1000$.

References

Andrews, D. W. K. and X. Shi (2013): "Inference Based on Conditional Moment Inequalities," *Econometrica*, 81, 609–666. 6

Armstrong, T. B. (2015): "Asymptotically Exact Inference in Conditional Moment Inequality Models," *Journal of Econometrics*, 186, 51–65. 6

Berger, R. L. (1982): "Multiparameter Hypothesis Testing and Acceptance Sampling," *Technometrics*, 24, 295–300. 8

Berger, R. L. and J. C. Hsu (1996): "Bioequivalence Trials, Intersection-Union Tests and Equivalence Confidence Sets," *Statistical Science*, 11, 283–319. 8

Casella, G. and R. L. Berger (2002): *Statistical Inference*, Pacific Grove, Calif. [u.a.]: Duxbury/Thomson Learning. 8

Chernozhukov, V., S. Lee, and A. M. Rosen (2013): "Intersection Bounds: Estimation and Inference," *Econometrica*, 81, 667–737. 6

Chetverikov, D. (2018): "Adaptive Tests of Conditional Moment Inequalities," *Econometric Theory*, 34, 186–227. 6

Cox, G. and X. Shi (2019): "A Simple Uniformly Valid Test for Inequalities," *arXiv:1907.06317 [econ, math, stat]*. 7

Currie, J. and E. Moretti (2003): "Mother's Education and the Intergenerational Transmission of Human Capital: Evidence from College Openings," *The Quarterly Journal of Economics*, 118, 1495–1532. 4

Dippel, C. (2014): "Forced Coexistence and Economic Development: Evidence From Native American Reservations: Forced Coexistence and Economic Development," *Econometrica*, 82, 2131–2165. 5

Romano, J. P., A. M. Shaikh, and M. Wolf (2014): "A Practical Two-Step Method for Testing Moment Inequalities," *Econometrica*, 82, 1979–2002. 7

Thornton, R. L. (2008): "The Demand for, and Impact of, Learning HIV Status," *American Economic Review*, 98, 1829–63. 4