The Case Against JIVE

by

Russell Davidson

Department of Economics
McGill University
Montreal, Quebec, Canada
H3A 2T7

GREQAM
Centre de la Vieille Charité
2 rue de la Charité
13002 Marseille, France

email: **russell.davidson@mcgill.ca**

and

James G. MacKinnon

Department of Economics
Queen's University
Kingston, Ontario, Canada
K7L 3N6

email: **jgm@qed.econ.queensu.ca**

Abstract

We perform an extensive series of Monte Carlo experiments to compare the performance of the "Jackknife Instrumental Variables Estimator," or JIVE, with that of the more familiar 2SLS and LIML estimators. We find no evidence to suggest that JIVE should ever be used. It is always more dispersed than 2SLS, often very much so, and it is almost always inferior to LIML in all respects. Interestingly, JIVE seems to perform particularly badly when the instruments are weak.

Keywords: JIVE, two-stage least squares, LIML

JEL codes: C100, C150, C200

This research was supported, in part, by grants from the Social Sciences and Humanities Research Council of Canada. We are grateful to participants at the Montreal Econometric Workshop for comments on a very early version.

February, 2004

1. Introduction

The finite-sample properties of instrumental variables estimators can be very poor, especially when the sample size is small and/or the instruments are weak. See, among many others, Nelson and Startz (1990a, 1990b), Staiger and Stock (1997), and Hahn, Hausman, and Kuersteiner (2002). One problem, but not the only one, is that the IV estimator can be severely biased in the same direction as the OLS estimator. This bias arises largely because of the correlation between the fitted values from first-stage estimation of the reduced form regression(s) and the error terms of the structural equation. It is therefore natural to conjecture that modifying the fitted values so as to reduce this correlation will result in an estimator with better finite-sample properties.

One way to accomplish this is to replace the usual fitted values from the reduced form regression(s) with omit-one fitted values. The omit-one fitted value for observation t is really a forecast based on OLS estimates that omit observation t. This leads to an IV estimator that was proposed by Phillips and Hale (1977) and later rediscovered by Angrist, Imbens, and Krueger (1999) and Blomquist and Dahlberg (1999). The two more recent papers called it the "Jackknife Instrumental Variables Estimator" or JIVE. The terminology is unfortunate, because this estimator is emphatically not a jackknife estimator. It appears to have been given this misleading name merely because, like some but not all jackknife estimators, it makes use of estimates that omit one observation at a time.

The only reason to use JIVE rather than the classic generalized IV, or 2SLS, estimator is that JIVE allegedly has better finite-sample properties. However, there is little evidence that this is actually the case. Phillips and Hale (1977), which was written at a time when computation was extremely expensive by modern standards, provided no simulation evidence at all. Angrist, Imbens, and Krueger (1999) performed just four experiments. Blomquist and Dahlberg (1999) considered a somewhat larger, but still very limited, number of cases. More recently, in a paper that is primarily concerned with other issues, Hahn, Hausman, and Kuersteiner (2002) provide evidence that JIVE can perform poorly. In this paper, we make a much stronger empirical case against the use of JIVE.

The main contribution of this paper is an extensive series of Monte Carlo experiments that compare the finite-sample properties of JIVE with those of the more familiar 2SLS and LIML estimators. In order to keep the experimental design manageable, we limit ourselves to a model with one structural and one reduced-form equation. Within this limitation, our experiments are much more comprehensive than any comparable ones that we are aware of in the econometric literature. We vary the sample size, the number of instruments, the weakness of the instruments, and the correlation between the errors of the reduced-form and structural equations. All of these factors affect the performance of the three estimators we study, often dramatically so.

One undesirable property of JIVE is that it has no moments. Evidence of this fact from simulations has been observed by Hahn, Hausman, and Kuersteiner (2002) and others, but, to our knowledge, the first proof is in Davidson and MacKinnon (2004). While having no moments is not, by itself, a fatal flaw for an estimator, it is certainly

a serious weakness. If an investigator is willing to use an estimator with no moments, then LIML is readily available. Therefore, unless JIVE can be shown to be superior to LIML for at least some interesting classes of problem, there is no reason ever to use JIVE. The extensive simulation experiments reported in this paper find only very limited parts of the parameter space for which JIVE is systematically better than LIML, but they find extensive regions in which the opposite is true. Moreover, even though LIML is usually better than JIVE, it is not at all clear that LIML is better than 2SLS. Thus we conclude that the case against JIVE is a very strong one.

In the next section, we establish notation and briefly discuss the three estimators that we study, as well as some variants of JIVE for which we will not present results. In Section 3, we discuss the design of the simulation experiments. In Section 4, we discuss the results of the experiments, all of which are presented graphically for ease of comprehension. In Section 5, we argue that it makes no sense to compare estimators on the basis of the performance of tests or confidence intervals. Section 6 contains concluding remarks.

2. IV, LIML, and JIVE Estimators

We are interested in one structural equation from a linear simultaneous equations model. The structural equation can be written as

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\beta}_1 + \boldsymbol{Y}\boldsymbol{\beta}_2 + \boldsymbol{u} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}, \tag{1}$$

where $\boldsymbol{Z}$ is an $n \times k_1$ matrix of observations on exogenous variables, $\boldsymbol{Y}$ is an $n \times k_2$ matrix of observations on endogenous variables, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are vectors of unknown parameters,

$$\boldsymbol{X} \equiv [\boldsymbol{Z} \;\; \boldsymbol{Y}], \;\text{ and }\; \boldsymbol{\beta} \equiv \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix}.$$

The rest of the system is treated as an unrestricted reduced form:

$$\boldsymbol{Y} = \boldsymbol{W}\boldsymbol{\Pi} + \boldsymbol{V}, \tag{2}$$

where there are l instruments in the matrix $\boldsymbol{W} \equiv [\boldsymbol{Z} \;\; \boldsymbol{W}_2]$. The error terms have mean zero conditional on $\boldsymbol{W}$, and they are assumed to be IID. For any observation,

$$\mathrm{E}\left(\begin{bmatrix} u_t \\ \boldsymbol{v}_t \end{bmatrix} [\, u_t \quad \boldsymbol{v}_t^\top] \right) = \begin{bmatrix} \sigma^2 & \boldsymbol{\omega}^\top \\ \boldsymbol{\omega} & \boldsymbol{\Omega} \end{bmatrix} \equiv \boldsymbol{\Sigma}, \tag{3}$$

where $\boldsymbol{v}_t$ is the t^{th} row of $\boldsymbol{V}$ rewritten as a column vector. The covariance matrix $\boldsymbol{\Sigma}$ is $(k_2 + 1) \times (k_2 + 1)$.

The 2SLS, or generalized IV, estimator can be written compactly as

$$\hat{\boldsymbol{\beta}}^{\text{IV}} = (\boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{y}, \tag{4}$$

where $\boldsymbol{P_W} \equiv \boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top$ projects on to $\mathcal{S}(\boldsymbol{W})$, the subspace spanned by the columns of $\boldsymbol{W}$. The usual covariance matrix estimator is

$$\hat{\sigma}^2_{\text{IV}}(\boldsymbol{X}^\top\boldsymbol{P_W}\boldsymbol{X})^{-1}, \text{ where } \hat{\sigma}^2_{\text{IV}} \equiv \frac{1}{n}\|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^{\text{IV}}\|^2. \tag{5}$$

A classic alternative to the 2SLS estimator is the limited-information maximum likelihood, or LIML, estimator, for which $\hat{\boldsymbol{\beta}}_2^{\text{LIML}}$ minimizes

$$\frac{(\boldsymbol{y} - \boldsymbol{Y}\boldsymbol{\beta}_2)^\top\boldsymbol{M_Z}(\boldsymbol{y} - \boldsymbol{Y}\boldsymbol{\beta}_2)}{(\boldsymbol{y} - \boldsymbol{Y}\boldsymbol{\beta}_2)^\top\boldsymbol{M_W}(\boldsymbol{y} - \boldsymbol{Y}\boldsymbol{\beta}_2)}, \tag{6}$$

where $\boldsymbol{M_Z}$ projects on to $\mathcal{S}^\perp(\boldsymbol{Z})$, and $\boldsymbol{M_W}$ projects on to $\mathcal{S}^\perp(\boldsymbol{W})$. If the minimized value of (6) is denoted by $\hat{\kappa}$, then the complete LIML estimator can be written as

$$\hat{\boldsymbol{\beta}}^{\text{LIML}} = \big(\boldsymbol{X}^\top(\mathbf{I} - \hat{\kappa}\boldsymbol{M_W})\boldsymbol{X}\big)^{-1}\boldsymbol{X}^\top(\mathbf{I} - \hat{\kappa}\boldsymbol{M_W})\boldsymbol{y}. \tag{7}$$

The usual covariance matrix estimator is

$$\hat{\sigma}^2_{\text{LIML}}\big(\boldsymbol{X}^\top(\mathbf{I} - \hat{\kappa}\boldsymbol{M_W})\boldsymbol{X}\big)^{-1}, \text{ where } \hat{\sigma}^2_{\text{LIML}} \equiv \frac{1}{n}\|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^{\text{LIML}}\|^2.$$

The matrix of fitted values that is implicitly used by the IV estimator as instruments for the endogenous explanatory variables $\boldsymbol{Y}$ is

$$\hat{\boldsymbol{Y}} = \boldsymbol{P_W}\boldsymbol{Y} = \boldsymbol{W}\hat{\boldsymbol{\Pi}} = \boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top\boldsymbol{Y}, \tag{8}$$

where $\hat{\boldsymbol{\Pi}}$ is the matrix of OLS estimates from the first-stage regressions (2). In contrast, the matrix of omit-one fitted values that is used by JIVE is the matrix $\tilde{\boldsymbol{Y}}$, of which the t^{th} row is $\tilde{\boldsymbol{Y}}_t = \boldsymbol{W}_t\hat{\boldsymbol{\Pi}}^{(t)}$, where $\hat{\boldsymbol{\Pi}}^{(t)}$ is the matrix of OLS estimates computed without observation t. The omit-one fitted values $\tilde{\boldsymbol{Y}}_t$ are easy to compute without actually running n linear regressions, by use of the formula

$$\tilde{\boldsymbol{Y}}_t = \frac{(\boldsymbol{P_W}\boldsymbol{Y})_t - h_t\boldsymbol{Y}_t}{1 - h_t},$$

where h_t is the t^{th} diagonal element of $\boldsymbol{P_W}$. The JIVE estimator that we will consider is

$$\hat{\boldsymbol{\beta}}^{\text{JIVE}} = (\tilde{\boldsymbol{X}}^\top\boldsymbol{X})^{-1}\tilde{\boldsymbol{X}}^\top\boldsymbol{y}, \text{ where } \tilde{\boldsymbol{X}} \equiv [\boldsymbol{Z} \;\; \tilde{\boldsymbol{Y}}]. \tag{9}$$

This estimator was called JIVE1 by Angrist, Imbens, and Krueger (1999) and UJIVE by Blomquist and Dahlberg (1999). We will simply call it JIVE.

Angrist, Imbens, and Krueger (1999) also propose a very similar estimator, which they call JIVE2. It uses the matrix with typical row

$$\ddot{\boldsymbol{Y}}_t \equiv \frac{(\boldsymbol{P_W}\boldsymbol{Y})_t - h_t\boldsymbol{Y}_t}{1 - 1/n} \tag{10}$$

instead of $\tilde{\boldsymbol{Y}}$ in (9). It can be seen readily that the fitted values defined in (10) are obtained using reduced-form estimates $\hat{\boldsymbol{\Pi}}^{(t)}$ in which the t^{th} observation is omitted in the factor $\boldsymbol{W}^\top\boldsymbol{Y}$ but not in the factor $\boldsymbol{W}^\top\boldsymbol{W}$. Our simulation results show that JIVE2 is always extremely similar to JIVE1. To the extent that these estimators differ in any systematic way, however, JIVE2 generally seems to perform a little bit less well. We therefore do not present results for JIVE2 or discuss it further.

Blomquist and Dahlberg (1999) also consider the estimator

$$\hat{\boldsymbol{\beta}}^{\text{OLSJ}} = (\tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^\top\boldsymbol{y}, \tag{11}$$

which is not an IV estimator at all, but is rather the OLS estimator from the regression of $\boldsymbol{y}$ on $\tilde{\boldsymbol{X}}$. As such, it suffers from a sort of errors-in-variables problem, since $\tilde{\boldsymbol{X}}$ can be thought of as $\boldsymbol{X}$ measured with error. In fact, OLSJ has extremely different finite-sample properties from the two JIVE estimators. In our experiments, $\hat{\boldsymbol{\beta}}^{\text{OLSJ}}$ was often even more biased than $\hat{\boldsymbol{\beta}}^{\text{IV}}$, while also being much less precise. Further, unlike the 2SLS estimator, which is biased towards the inconsistent OLS estimator, OLSJ is biased towards zero, just like the OLS estimator when the explanatory variable is measured with error. It appears then that OLSJ is, at least in many cases, very much inferior to 2SLS in all respects. In order to keep the figures in Section 4 from becoming unreadable, we therefore do not report results for it.

3. The Simulation Experiments

The design of our experiments is just about as simple as possible. There is one structural equation, with dependent variable $\boldsymbol{y}$ generated by the equation

$$\boldsymbol{y} = \beta_1 + \beta_2\boldsymbol{x} + \boldsymbol{u}. \tag{12}$$

The endogenous explanatory variable $\boldsymbol{x}$ is generated as a function of the exogenous instruments in the matrix $\boldsymbol{W}$ by

$$\boldsymbol{x} = \boldsymbol{W}\boldsymbol{\pi} + \boldsymbol{v}. \tag{13}$$

This data-generating process can be thought of as representing a model in which (12) is the structural equation of interest and (13) is the single reduced-form equation. In fact, however, it can be shown that the distributions of all the estimators covered in our experiments are the same for that model, with n observations, and a model with k_1 exogenous explanatory variables in the structural equation, like that constituted by the equations (1) and (2), with the number of endogenous explanatory variables, k_2, equal to 1, but with a sample size of $n - k_1 + 1$. See Davidson and MacKinnon (2004) on this point.

The first column of $\boldsymbol{W}$ is a vector of 1s, and the remaining $l - 1$ columns are IID standard normal random variables which are redrawn for each replication. The number of overidentifying restrictions is $r = l - 2$. Each row $[u_t \;\; v_t]$ of the $n \times 2$ matrix of error terms $[\boldsymbol{u} \;\; \boldsymbol{v}]$ is IID normal with mean vector $\boldsymbol{0}$ and covariance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_u^2 & \rho\sigma_u\sigma_v \\ \rho\sigma_u\sigma_v & \sigma_v^2 \end{bmatrix}.$$

In all experiments, $\beta_1 = 1$, $\beta_2 = 1$, and $\sigma_u^2 = 1$. It can be shown that the t statistic for β_2 to be equal to its true value is invariant to these parameters, and that the bias of the IV estimator (the only one for which bias is defined) is proportional to σ_u. Thus it is no interest to change the values of these parameters.

Since all the estimators we consider depend on the instrument matrix $\boldsymbol{W}$ only through the orthogonal projection $\boldsymbol{P_W}$, it is equally of no interest to change the specification of $\boldsymbol{W}$ in ways that do not affect $\boldsymbol{P_W}$; for this reason there is no need to investigate the consequences of greater or lesser correlation among the instruments. It can also be seen that multiplying σ_v and the elements of the vector $\boldsymbol{\pi}$ by the same constant affects only the scale of the variable $\boldsymbol{x}$, and so has only a scale effect on the distribution of estimators of β_2, and no effect at all on the t statistics.

What does have an important effect is the ratio of $\|\boldsymbol{\pi}\|^2$ to σ_v^2. This ratio can be interpreted as the signal-to-noise ratio in the reduced-form equation (13). Small values of the ratio imply that the instruments are weak. In order to capture the effect the ratio has on the distributions we study, we normalize the quantity

$$\sigma_2^2 \equiv \|\boldsymbol{\pi}\|^2 + \sigma_v^2, \tag{14}$$

to unity. In addition, for convenience and without loss of generality, the values of the π_j, for $j \geq 2$, were constrained to be equal, with $\pi_1 = 0$. We may then vary a parameter that we denote by R_∞^2. It is the limiting R^2 of the reduced-form regression as $n \to \infty$. It follows from equation (14) and the fact that the matrix $n^{-1}\boldsymbol{W}^\top\boldsymbol{W}$ tends to an $l \times l$ identity matrix as $n \to \infty$, that

$$R_\infty^2 = \frac{\|\boldsymbol{\pi}\|^2}{\|\boldsymbol{\pi}\|^2 + \sigma_v^2}.$$

This is just a deterministic function of the signal-to-noise ratio, with the advantage that takes values restricted to the $[0, 1]$ interval.

In our experiments, therefore, we vary the sample size and three parameters: the number of overidentifying restrictions, $r = l - 2$, the correlation between the structural and reduced-form error terms, ρ, and R_∞^2. Because the only exogenous variable is the constant, nR_∞^2 is a reasonable measure of the weakness of the instruments, in the sense that a small value of the parameter implies weak instruments. It is a monotonically increasing function of the "concentration parameter" that is often used in the weak instruments literature; see Stock, Wright, and Yogo (2002). If there were nonconstant exogenous explanatory variables explicitly included in the structural equation, it would be necessary to replace R_∞^2 by a partial R^2, that is, the limiting R^2 of the reduced-form regression in which all variables had been multiplied by the orthogonal projection on to the orthogonal complement of the exogenous explanatory variables.

All experiments used 500,000 replications. When either ρ or R_∞^2 was varied, we performed experiments for every value from 0.00 to 0.99 at intervals of 0.01. We did not investigate negative values of ρ because results would have been identical to

those for the corresponding positive values, except for the sign of any bias. The case $R^2_\infty = 0$ is evidently an extreme one, for which β_2 is not asymptotically identified. It is included to show what happens in the limit as the instruments become infinitely weak. When r was varied, we considered all values from 0 to 16. Every experiment was performed six times, for samples of size 25, 50, 100, 200, 400, and 800.

It would be impractical either to obtain or to present graphically results for 100 values of ρ, 100 values of R^2_∞, 17 values of r, and six sample sizes. This would require 1,020,000 experiments. Therefore, when we varied one parameter, we held the other two fixed at certain base values. We chose the base values to make estimation challenging. The base value of ρ was 0.9, because all three estimators generally performed worse as ρ was increased. The base value of R^2_∞ was 0.1, which implies that the instruments are very weak when n is small. The base value of r was 5. Because results for very small values of r were sometimes quite different from results for larger values, we did not want to use a very small value for the base case. However, because computational cost increased sharply with r, we also did not want to use a large value. The choice $r = 5$ seems a reasonable compromise.

Because the LIML and JIVE estimators have no moments, it makes no sense to report either bias or root mean squared error for these estimators, although some other authors have done so. As one would expect, both bias and RMSE were highly variable across experiments, extremely sensitive to the random numbers used, and essentially meaningless, for the two estimators that do not have moments.[1] Instead, we report two quantile-based summary statistics. As a measure of the central tendency of each estimator, we report *median bias*, that is, the 0.5 quantile of the $\hat{\beta}_2$ estimates minus the true value β_{20}. As a measure of dispersion, we report the *nine decile range*, that is, the 0.95 quantile minus the 0.05 quantile. The nine decile range is the width of an interval in which the estimates lie 90% of the time. It seems to us a more reasonable measure of dispersion than the interquartile range that is commonly used.

4. Results of the Experiments

Results are reported in Figures 1 through 9. Each figure contains six panels, one for each of the sample sizes 25, 50, 100, 200, 400, and 800. Readers should be careful to check the vertical scales of the various panels, which are not always the same for different sample sizes.

The first three figures concern median bias. Figure 1 shows that the median bias of the 2SLS estimator appears to be proportional to ρ and inversely proportional to n. These results are consistent with the theoretical results on the distribution of this estimator discussed by Phillips (1984). In contrast, the median biases of LIML and JIVE are evidently nonlinear functions of ρ. For LIML, the median bias is indiscernible for

[1] In one case, simply changing compilers, while using the same sequence of random numbers, produced dramatically different results for bias and RMSE. Results for the quantile-based summary statistics that we actually report were unaffected by the change.

$n \geq 200$. In contrast, the median bias of JIVE appears to be $O(n^{-1})$, like that of 2SLS, but it is positive for small samples and negative for large ones.

Figure 2 shows that the median biases of all three estimators generally decline as R^2_∞ increases. For 2SLS and LIML, the decline is monotonic. For JIVE, however, the sign of the median bias always changes from positive to negative as R^2_∞ increases, before it finally approaches zero in absolute value. When nR^2_∞ is very small, that is, when the instruments are extremely weak, JIVE is actually a little more biased than 2SLS.

Figure 3 shows that median bias depends on the number of overidentifying restrictions. For 2SLS and LIML, it increases monotonically, although, for $n \geq 100$, the median bias of LIML appears to be essentially zero. The median bias of JIVE can be of either sign and can be larger in absolute value than that of 2SLS. It is always negative for the larger sample sizes. Notice that 2SLS and LIML are identical when $r = 0$, but JIVE is quite different, and always more biased.

We conclude that, except for a few carefully chosen parameter values, JIVE is always worse than LIML in terms of median bias. For larger sample sizes, it is almost always very much worse, because LIML is essentially median unbiased. Although JIVE usually outperforms 2SLS according to this criterion, there are numerous cases in which it fails to do so. Unlike LIML, JIVE may be biased in either the same direction as 2SLS or in the opposite direction.

Figures 4 through 6 deal with the dispersion of the three estimators, as measured by the nine decile range. In Figure 4, where ρ varies on the horizontal axes, JIVE is always more dispersed then LIML, which in turn is always more dispersed than 2SLS. The differences are very large when n is small, although they are quite modest when $n = 800$. Figure 5 shows that the same pattern holds for most values of R^2_∞. For moderately small values, JIVE is often much more dispersed than LIML. For extremely small values, however, LIML is more dispersed than JIVE. Since the set of values for which this is true becomes smaller as n increases, it evidently occurs only when the instruments are extremely weak. Note that this is the case in which the median bias of JIVE is very large.

Further evidence is provided by Figure 6, which shows that JIVE is always more dispersed than LIML except for large values of r when $n = 25$. Once again, the differences among the three estimators can be striking. 2SLS, the only estimator with moments, is always less dispersed than LIML, which is often very much less dispersed than JIVE. It is clear from all three figures that JIVE and LIML are much more dispersed than 2SLS when the instruments are weak. Thus it is evident that, in line with a conclusion reached in Hahn, Hausman, and Kuersteiner (2002), using an estimator with no moments is not a way to solve the weak instruments problem.

The final three figures concern rejection frequencies for two-tailed asymptotic tests based on pseudo-t statistics for β_2 to equal its true value. We include these figures because many authors, including Blomquist and Dahlberg (1999), appear to believe that the performance of tests or, equivalently, confidence intervals is an important criterion for choosing among estimators. As we argue in the next section, this is actually not a sensible criterion at all. Nevertheless, the results are of some interest.

On average, in Figures 7 through 9, LIML seems to yield the most reliable inferences, although there are a number of cases in which JIVE is more reliable than LIML. Because the base case value of ρ is 0.9, overrejection is more common in the figures than underrejection. However, it can be seen from Figure 7 that all three estimators underreject for smaller values of ρ. This is particularly severe for JIVE. For values of ρ less than about 0.35, even 2SLS yields more reliable inferences than JIVE.

Figures 7 through 9 provide little support for using JIVE rather than LIML. But they may appear to suggest that LIML should be used instead of 2SLS if we are concerned primarily with obtaining accurate inferences. This means using a less efficient estimator (in the sense of dispersion as measured by the nine decile range) in order to obtain more accurate inferences. In the next section, however, we argue that this is generally not a sensible thing to do.

5. Accuracy of Inference and Choice of Estimator

The fundamental reason for not choosing an estimator on the basis of the accuracy of asymptotic inference when the null hypothesis is true is that we can always obtain more accurate asymptotic t tests or, equivalently, more accurate asymptotic confidence intervals simply by adding noise to an efficient estimator. Suppose, for purposes of illustration, that the finite-sample distribution of some estimator is well approximated by the equation

$$\hat{\theta} = \theta_0 + n^{-1}b + n^{-1/2}e. \tag{15}$$

Here θ is a parameter to be estimated, θ_0 is the true value, $\hat{\theta}$ is an asymptotically efficient estimator, b is a constant that determines bias, and the random variable e is $\mathrm{N}(0, \sigma_e^2)$. In many cases, (15) should provide a reasonable approximation to the true distribution of $\hat{\theta}$.

An asymptotic t statistic for $\theta = \theta_0$ is

$$t_\theta = \frac{\hat{\theta} - \theta_0}{n^{-1/2}\sigma_e} = \frac{n^{-1}b + n^{-1/2}e}{n^{-1/2}\sigma_e} = n^{-1/2}\frac{b}{\sigma_e} + \varepsilon, \tag{16}$$

where $\varepsilon \sim \mathrm{N}(0,1)$. When b is large and/or n is small, this t statistic will have poor finite-sample properties. We can "improve" these properties by using the estimator

$$\hat{\theta}' = \theta_0 + n^{-1}b + n^{-1/2}(e + w), \tag{17}$$

where $w \sim \mathrm{N}(0, \sigma_w^2)$ and is uncorrelated with e, and σ_w^2 is $O(1)$. The corresponding t statistic is

$$\begin{aligned} t'_\theta = \frac{\hat{\theta}' - \theta_0}{n^{-1/2}(\sigma_e^2 + \sigma_w^2)^{1/2}} &= \frac{n^{-1}b + n^{-1/2}(e + w)}{n^{-1/2}(\sigma_e^2 + \sigma_w^2)^{1/2}} \\ &= n^{-1/2}\frac{b}{(\sigma_e^2 + \sigma_w^2)^{1/2}} + \varepsilon', \end{aligned} \tag{18}$$

where $\varepsilon' \sim \mathrm{N}(0,1)$. It is clear from (16) and (18) that t'_θ is better behaved under the null than t_θ, because its mean is closer to 0.

We have achieved this "improvement" simply by adding noise to $\hat{\theta}$. No sensible person would want to use $\hat{\theta}'$ instead of $\hat{\theta}$. But if we choose between these estimators on the basis of the performance of an asymptotic t test under the null, or, equivalently, on the basis of the coverage of the usual asymptotic confidence interval, that is exactly what we do.

Just why using $\hat{\theta}'$ instead of $\hat{\theta}$ is unwise becomes apparent when the null is false. If $\theta = \theta_1 \neq \theta_0$, then

$$t_\theta = n^{1/2} \frac{\theta_1 - \theta_0}{\sigma_e} + n^{-1/2} \frac{b}{\sigma_e} + \varepsilon, \tag{19}$$

and

$$t'_\theta = n^{1/2} \frac{\theta_1 - \theta_0}{(\sigma_e^2 + \sigma_w^2)^{1/2}} + n^{-1/2} \frac{b}{(\sigma_e^2 + \sigma_w^2)^{1/2}} + \varepsilon'. \tag{20}$$

The first term on the right-hand side of each of (19) and (20) is the noncentrality parameter. This is evidently smaller for t'_θ than for t_θ. The smaller noncentrality parameter implies that t'_θ rejects less often than t_θ. Thus using an inefficient estimator has improved performance under the null but reduced power. We could have obtained an exact test, but one with power everywhere equal to its size, by letting σ_w^2 tend to infinity.

This example, in our view, shows that it is misguided to use an estimator that is inefficient in finite samples simply because it leads to more accurate inferences, using conventional asymptotic procedures, than an efficient one. A much better approach would be to improve the inferences associated with the efficient estimator, perhaps by bootstrapping, perhaps by correcting finite-sample bias, or perhaps by using more accurate standard errors.

6. Conclusion

We have presented the results of a rather extensive set of Monte Carlo experiments on the finite-sample performance of three single-equation estimators for linear simultaneous equations models, namely, JIVE and the familiar 2SLS and LIML estimators. The results show clearly that, in most regions of the parameter space that we have studied, JIVE is inferior to LIML in all respects. It almost always has a greater median bias, it is more dispersed except when the instruments are extremely weak, and it usually leads to less reliable inferences. Thus, if an investigator wishes to use an estimator that has no moments, it is usually better to use LIML than JIVE.

Our results, however, do not provide unambiguous support for the use of LIML. It is always more dispersed than 2SLS, often dramatically so when the instruments are weak. This excessive dispersion appears to be associated with its lack of moments. Other estimators that have moments, such as Fuller's (1977) modification of LIML and the (genuine) jackknife 2SLS estimator proposed by Hahn, Hausman, and Kuersteiner (2002) outperform LIML in many circumstances, as the latter paper shows. It

would be of interest to investigate additional estimators using the experimental design of this paper. In order to make the case against JIVE as coherently as possible, however, we have not done so in this paper.

References

Angrist J. D., Imbens G. W., Krueger A. B. 1999. Jackknife instrumental variables estimation, *Journal of Applied Econometrics* **14**, 57–67.

Blomquist S., Dahlberg M. 1999. Small sample properties of LIML and jackknife IV estimators: Experiments with weak instruments, *Journal of Applied Econometrics* **14**, 69–88.

Davidson R., MacKinnon J. G. 2004. Moments of IV and JIVE estimators, mimeo, Queen's University.

Hahn J., Hausman J. A., Kuersteiner G. 2002. Estimation with weak instruments: Accuracy of higher order bias and MSE approximations, mimeo, MIT.

Fuller W. A. 1977. Some properties of a modification of the limited information estimator, *Econometrica* **45**, 939–953.

Nelson C. R., Startz R. 1990a. The distribution of the instrumental variables estimator and its t-ratio when the instrument is a poor one, *Journal of Business* **63**, S125–140.

Nelson C. R,, Startz R. 1990b. Some further results on the exact small sample properties of the instrumental variables estimator, *Econometrica* **58**, 967–976.

Phillips G. D. A., Hale C. 1977. The bias of instrumental variable estimators of simultaneous equation systems, *International Economic Review* **18**, 219–228.

Phillips, P. C. B. 1984. Exact small sample theory in the simultaneous equations model. In *Handbook of Econometrics*, Volume 1, Griliches Z, Intriligator M (eds.) Amsterdam: North-Holland.

Staiger D., Stock J. H. 1997. Instrumental variables regressions with weak instruments, *Econometrica* **65**: 557–586.

Stock J. H, Wright J. H, Yogo M. 2002. A survey of weak instruments and weak identification in generalized method of moments, *Journal of Business and Economic Statistics* **20**, 518–529.

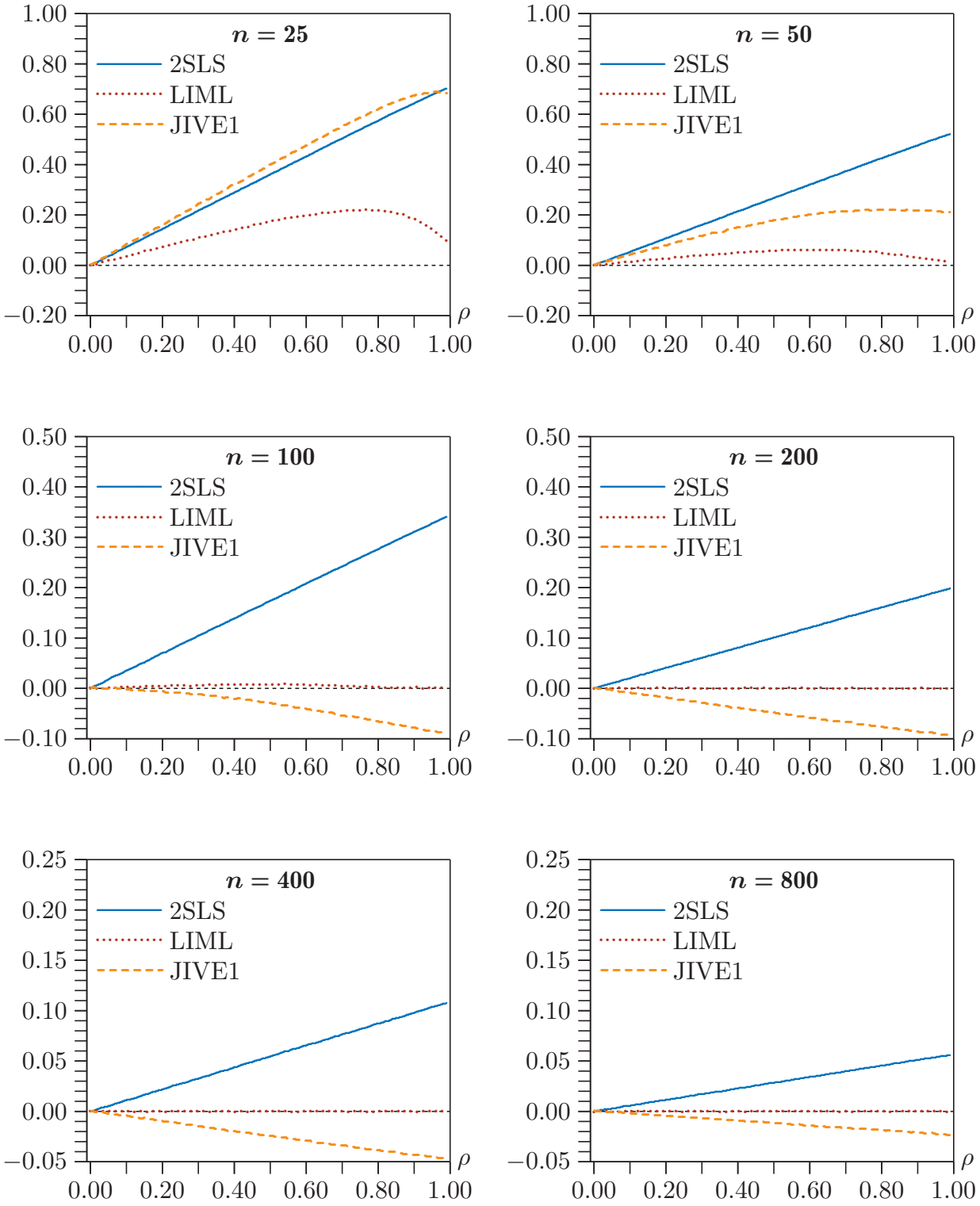

Figure 1. Median bias of three estimators, $r = 5$, $R^2_\infty = 0.1$

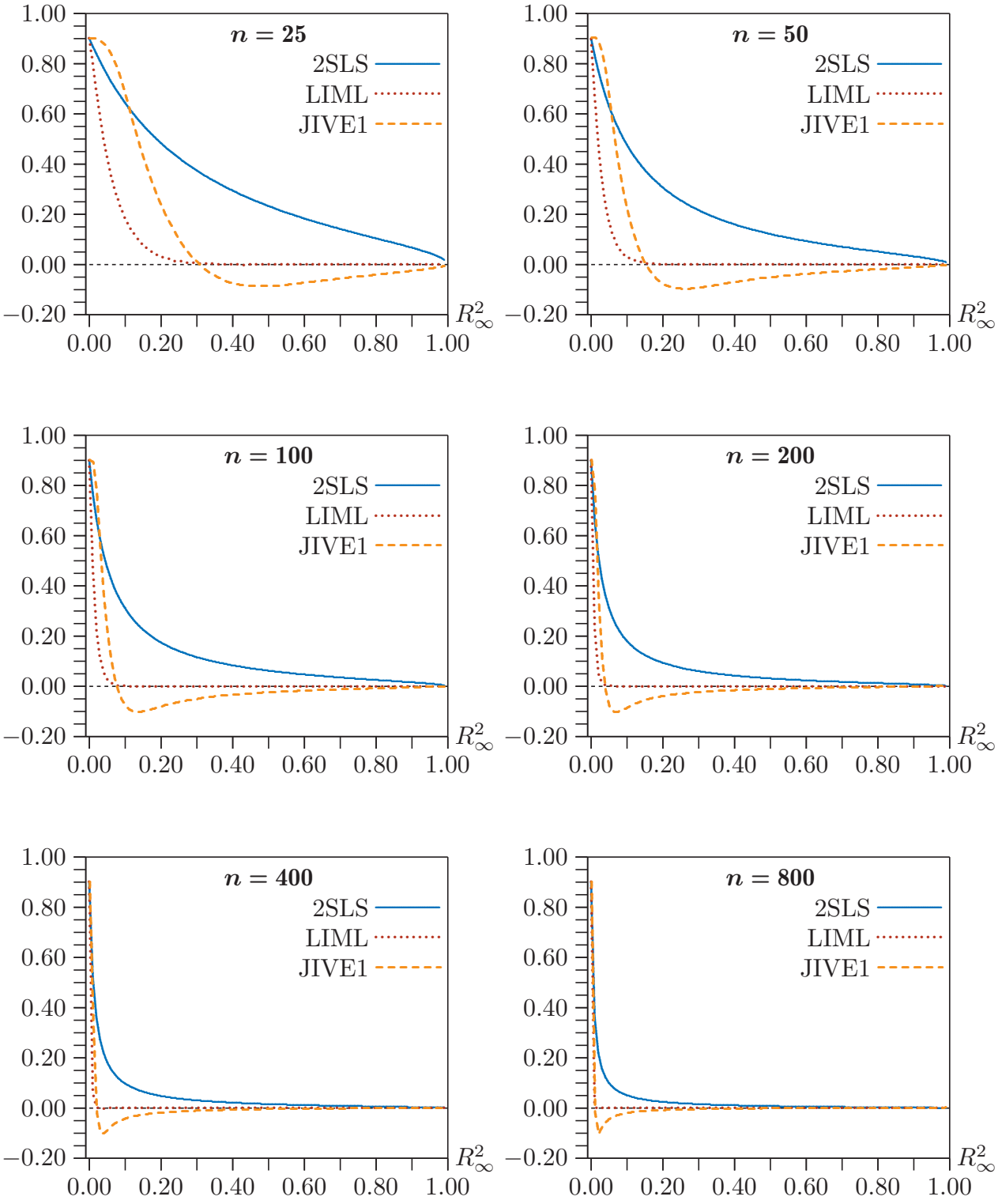

Figure 2. Median bias of three estimators, $r = 5$, $\rho = 0.9$

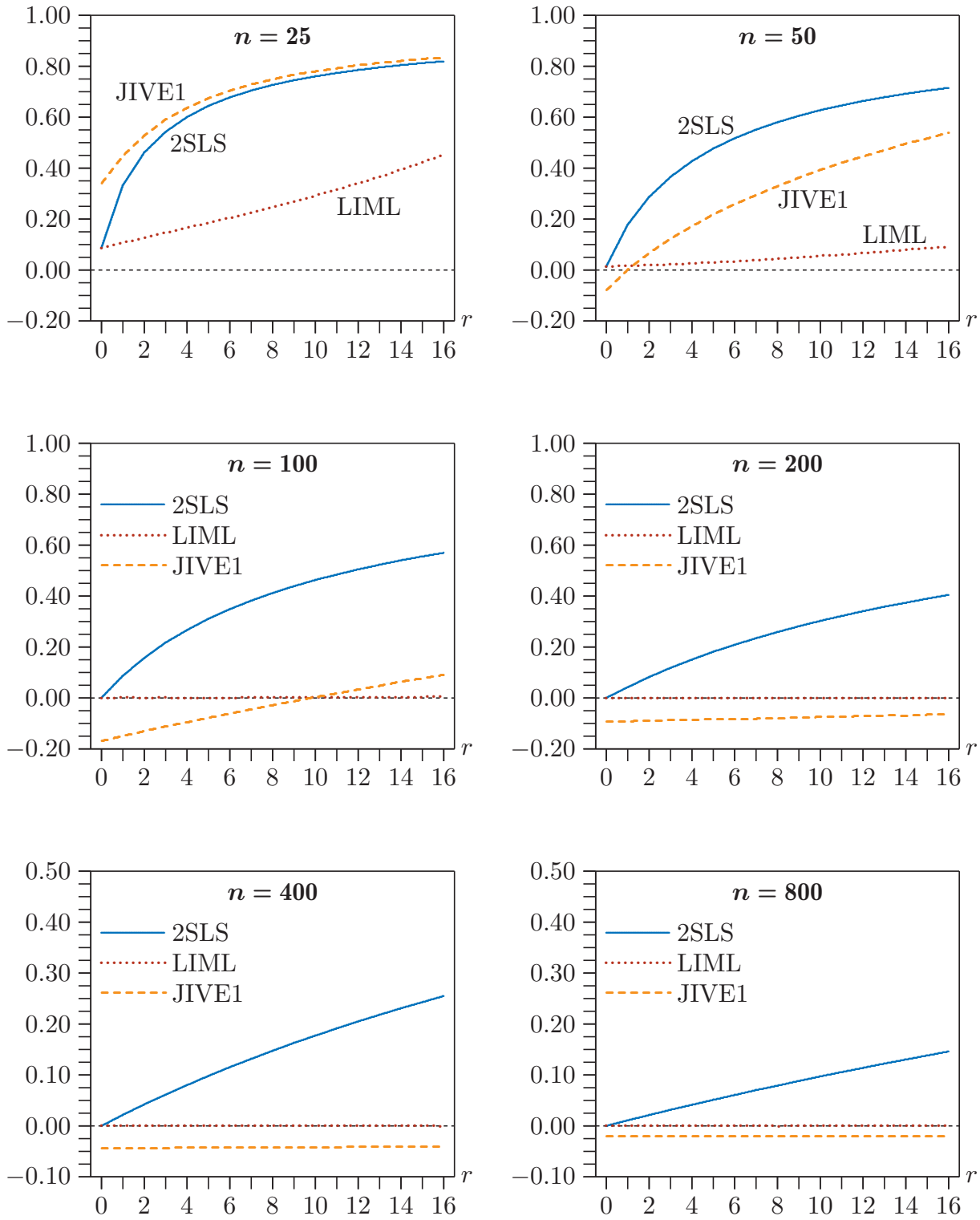

Figure 3. Median bias of three estimators, $R^2_\infty = 0.1$, $\rho = 0.9$

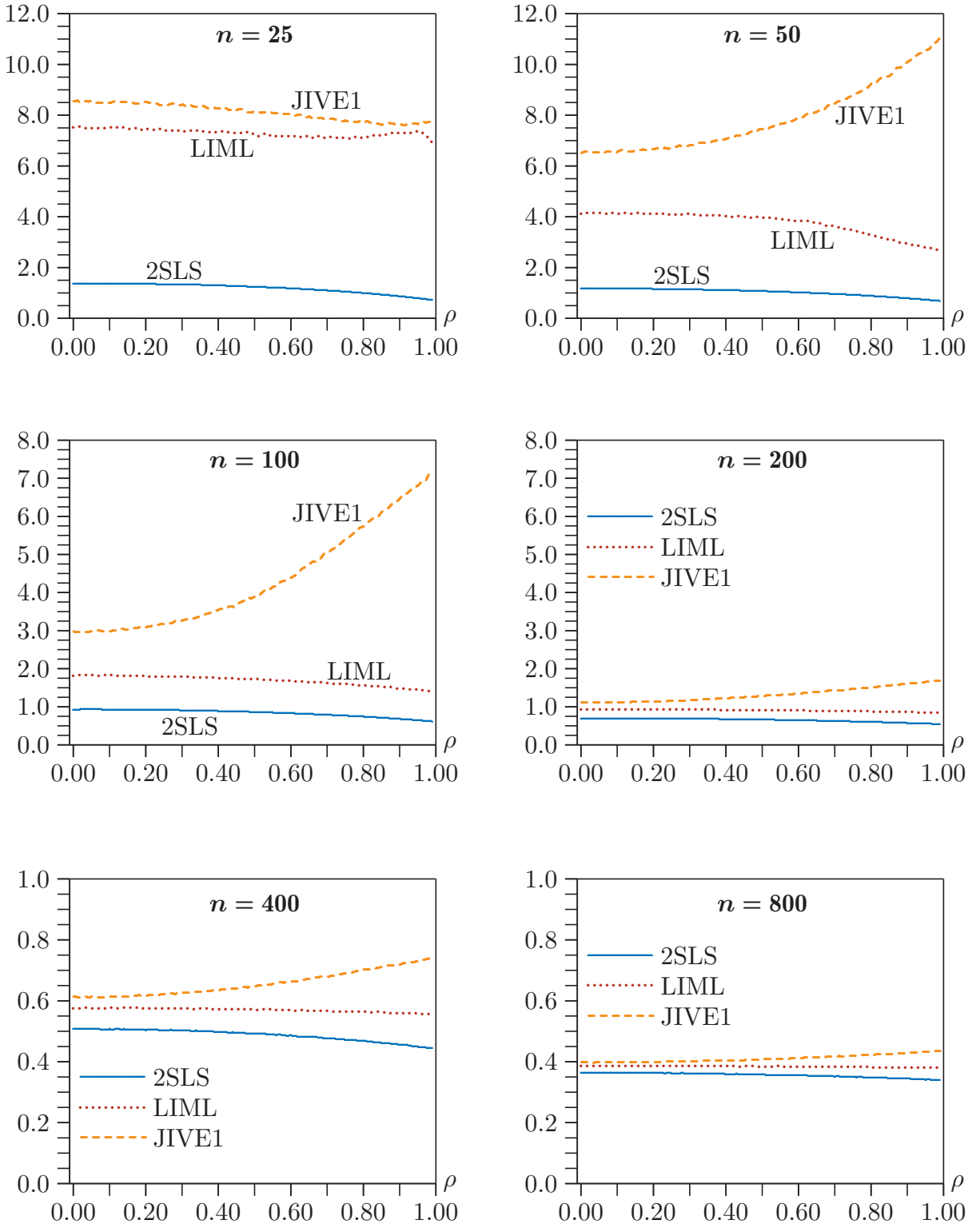

Figure 4. Nine decile range of three estimators, $r = 5$, $R^2_\infty = 0.1$

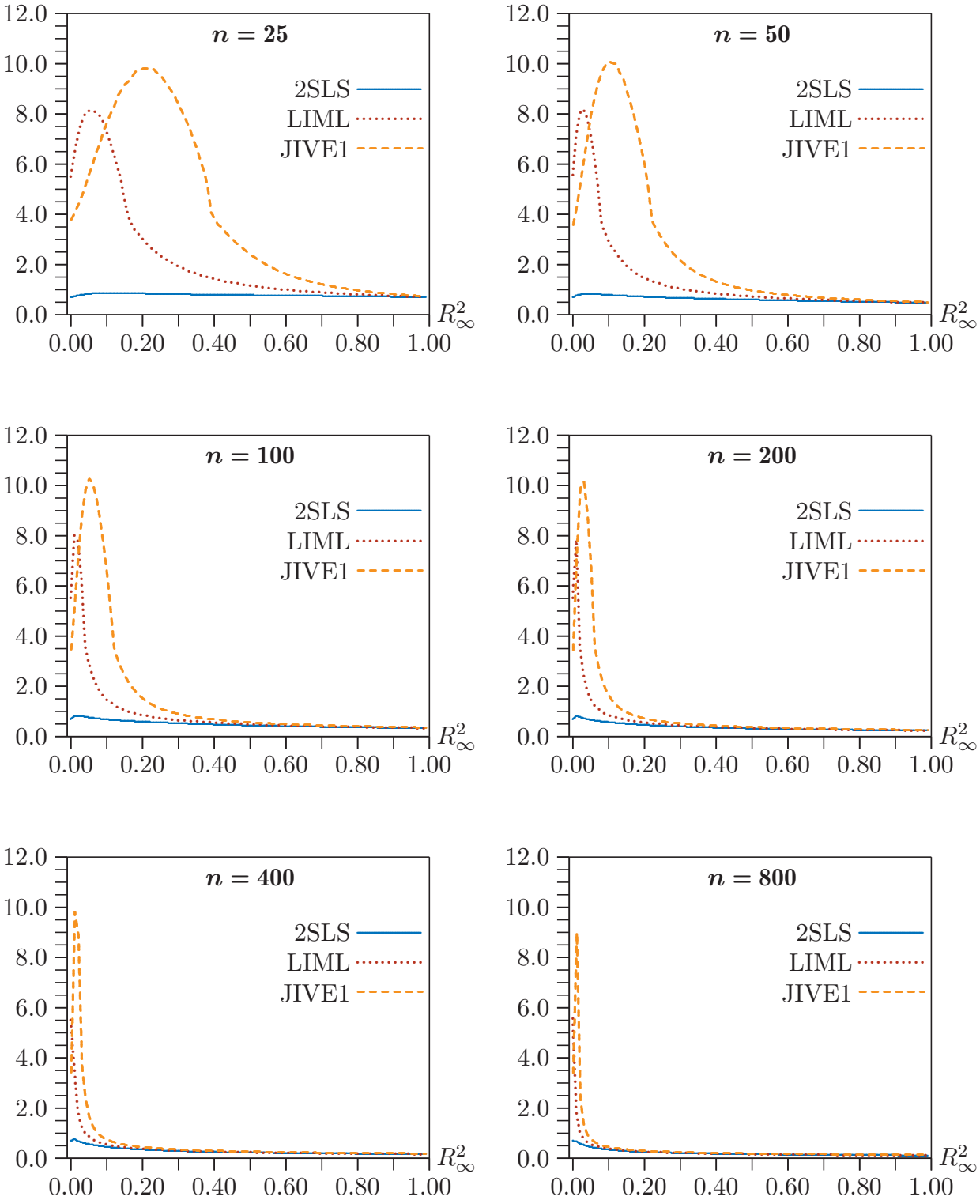

Figure 5. Nine decile range of three estimators, $r = 5$, $\rho = 0.9$

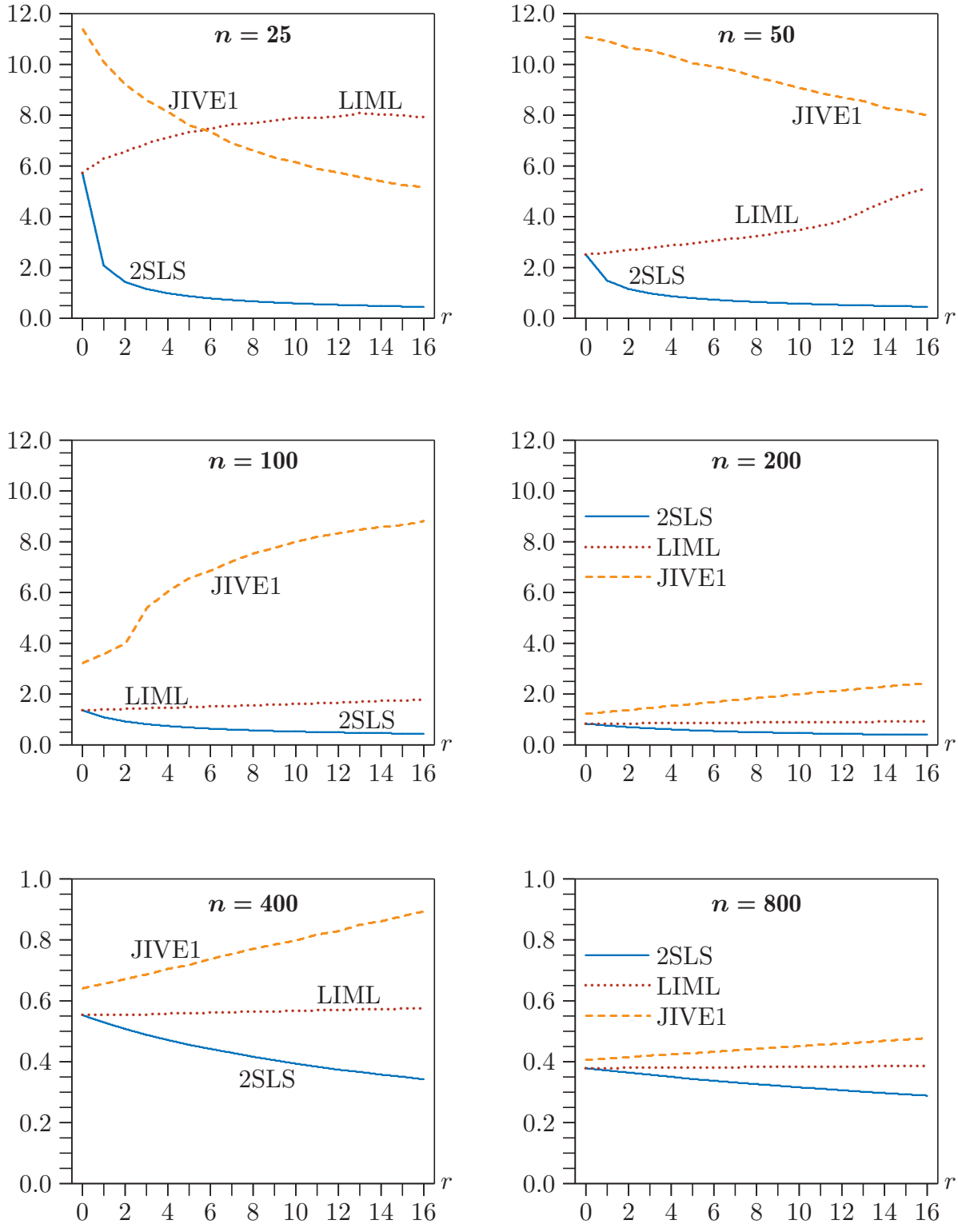

Figure 6. Nine decile range of three estimators, $R^2_\infty = 0.1$, $r = 0.9$

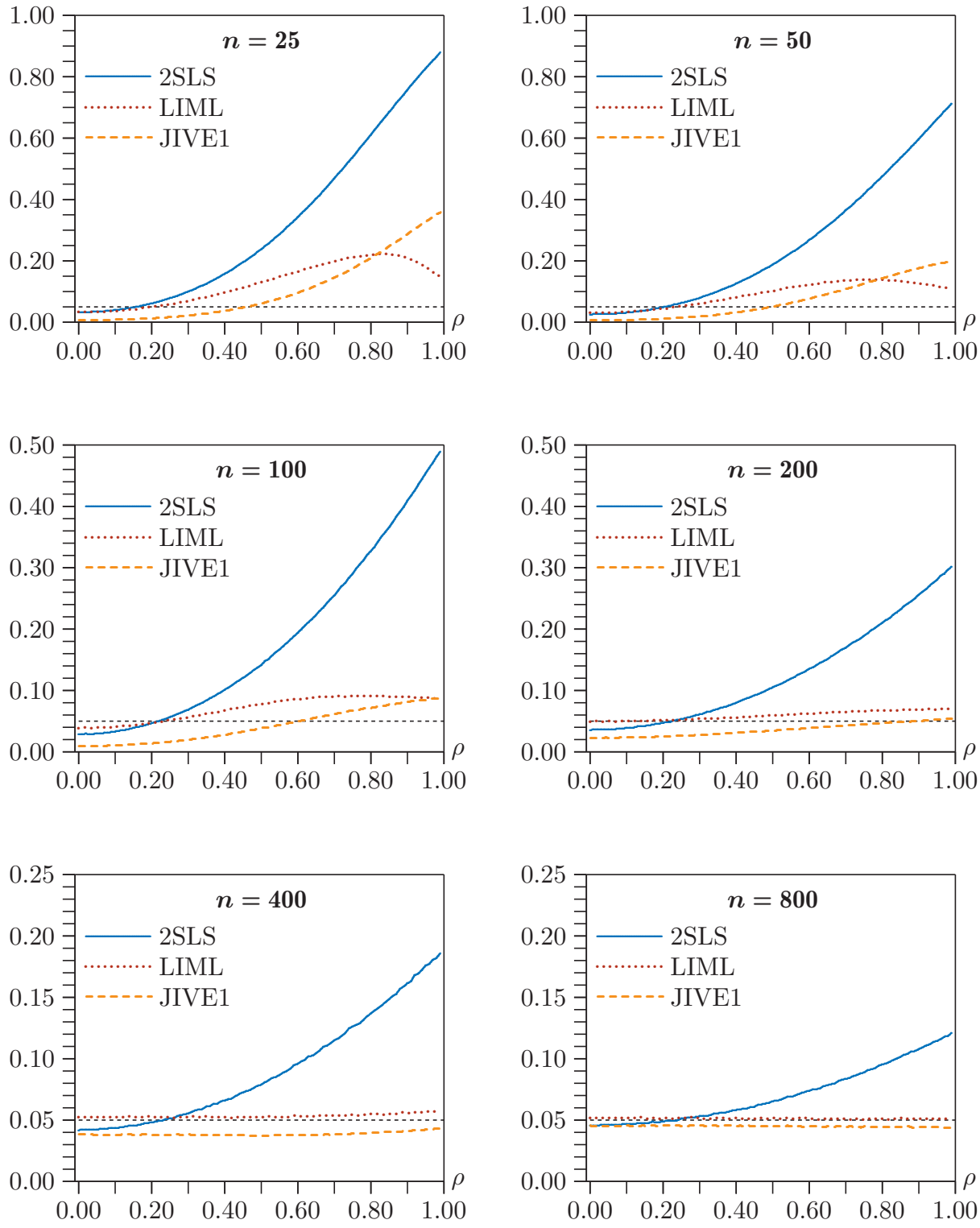

Figure 7. Rejection frequencies for asymptotic t tests at .05 level, $r = 5$, $R^2_\infty = 0.1$

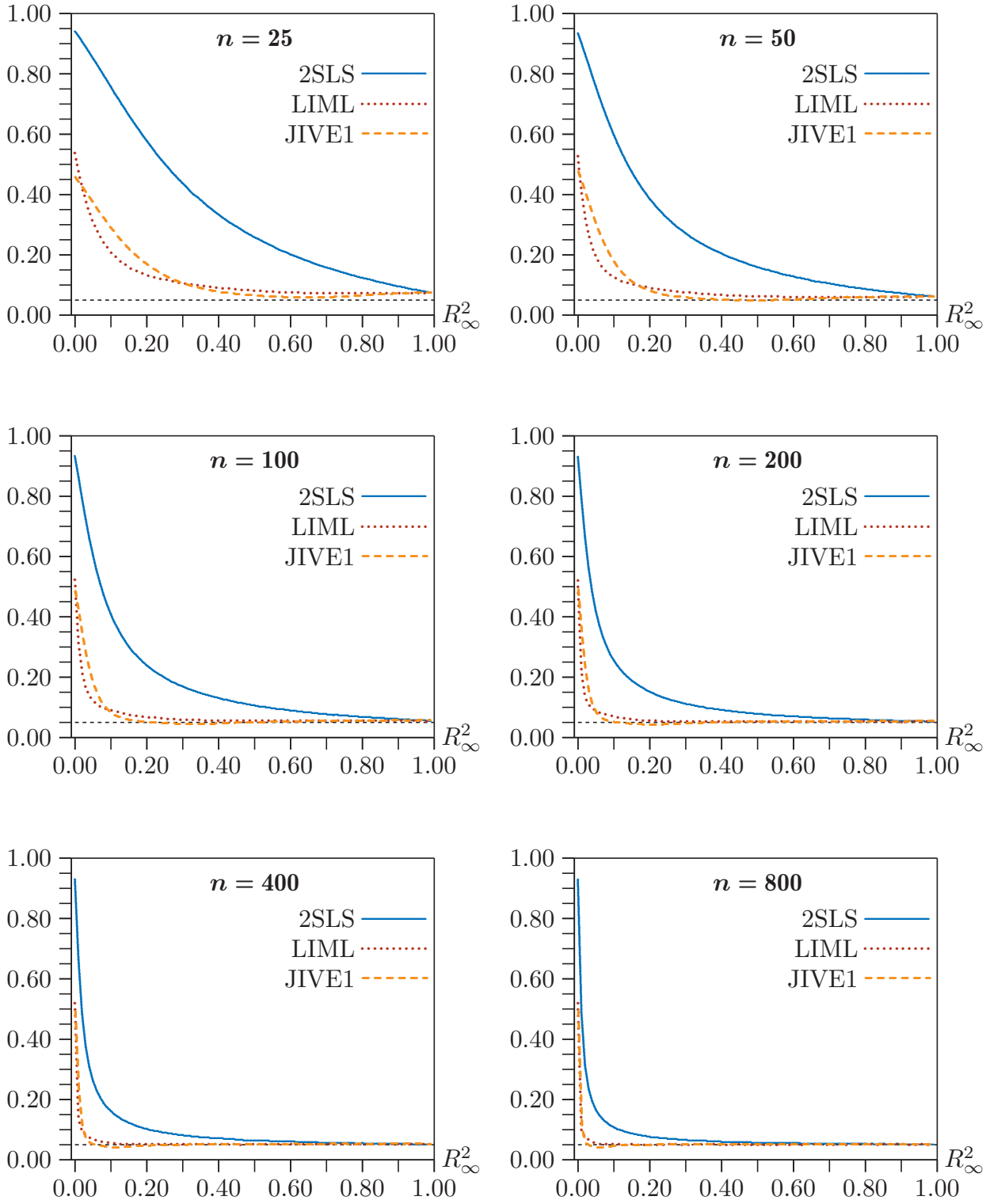

Figure 8. Rejection frequencies for asymptotic t tests at .05 level, $r = 5$, $\rho = 0.9$

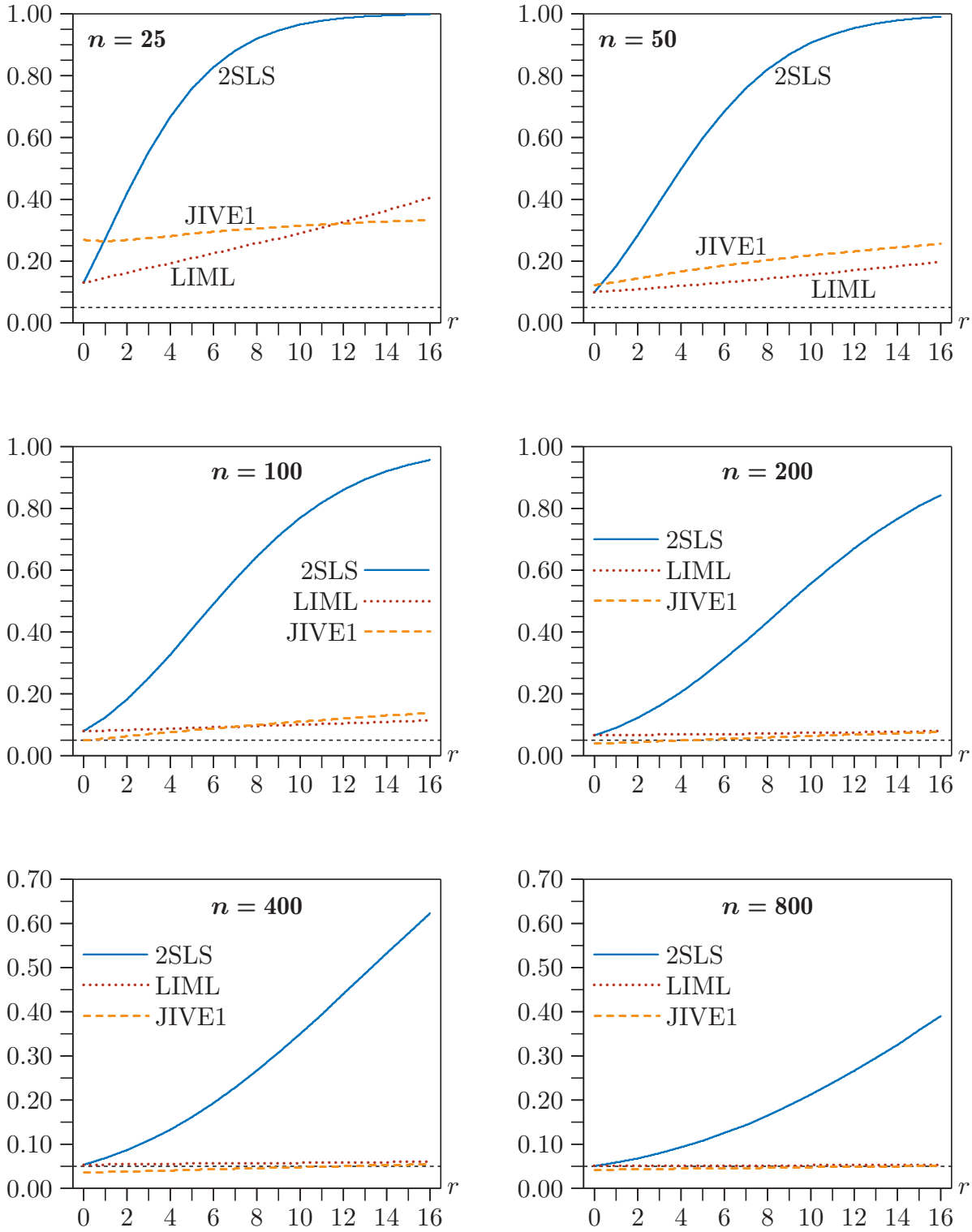

Figure 9. Rejection frequencies for asymptotic t tests at .05 level, $R^2_\infty = 0.1$, $\rho = 0.9$