Applications of the Fast Double Bootstrap

by

James G. MacKinnon

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

jgm@econ.queensu.ca

Abstract

The fast double bootstrap, or FDB, is a procedure for calculating bootstrap P values that is much more computationally efficient than the double bootstrap itself. It can be used to verify that the results of ordinary bootstrap tests are reliable, and it can provide more accurate results than they do. For the fast double bootstrap to be valid, the test statistic must be independent of the random parts of the bootstrap data generating process. This paper presents simulation evidence on the performance of fast double bootstrap tests in two cases of interest to econometricians. One of the cases involves both symmetric and equal-tail bootstrap tests, which can have quite different power properties.

This research was supported, in part, by grants from the Social Sciences and Humanities Research Council of Canada.

August, 2004

1. Introduction

The simplest, and usually the most informative, way to perform a bootstrap test is to compute a bootstrap P value. We first compute the test statistic itself in the usual way. Then we generate a number of bootstrap samples and use each of them to compute a bootstrap statistic. Finally, we calculate a bootstrap P value as the proportion of the bootstrap statistics that are more extreme than the actual test statistic. When this P value is sufficiently small, we reject the null hypothesis.

Bootstrap tests based on asymptotically pivotal test statistics should generally perform better in finite samples than tests based on asymptotic theory, in the sense that they will commit errors that are of lower order in the sample size n; see, among others, Hall (1992) and Davidson and MacKinnon (1999). However, this does not mean that bootstrap tests always perform acceptably well in finite samples. For an asymptotic test, one way to check whether it is reliable is simply to use the bootstrap. If the asymptotic and bootstrap P values associated with a given test statistic are similar, we can be fairly confident that the asymptotic one is reasonably accurate.

If the asymptotic and bootstrap P values are not close, however, it is not clear whether the latter is reliable. One possibility is to compute more than one bootstrap P value, perhaps based on different bootstrap data generating processes. There are cases where two or more genuinely different bootstrap DGPs are available, for example, ones based on the pairs bootstrap and the wild bootstrap for regression models with heteroskedastic errors. If two or more bootstrap P values based on different bootstrap DGPs are similar, then it may seem reasonable to trust both of them.

An alternative approach would be to compare the ordinary bootstrap P value with a double bootstrap P value, as discussed in the next section. However, the double bootstrap tends to be very computationally demanding. Davidson and MacKinnon (2001) therefore suggested what they called the **fast double bootstrap**, or **FDB**. It is very much less expensive to compute than the double bootstrap itself, but its validity requires stronger conditions.

There is, at present, only limited evidence on how useful the fast double bootstrap is likely to be in practice. Ideally, it would yield a P value similar to the ordinary bootstrap P value when the latter is reliable, and a P value more accurate than the ordinary bootstrap P value when the latter is unreliable. One objective of this paper is to see whether this is likely to be the case for some classes of econometric models.

In the next section, I discuss the mechanics of bootstrap testing and point out that there is more than one way to calculate bootstrap P values for two-tailed tests. Then, in Section 3, I discuss double bootstrap and fast double bootstrap tests and the relationship between them. In Sections 4 and 5, I present simulation results for some regression-based tests for serial correlation and for ARCH errors. These results suggest that using the fast double bootstrap can improve the performance of bootstrap tests, modestly in most cases, but quite substantially in some.

2. Bootstrap Tests

Let τ denote a test statistic, and let $\hat{\tau}$ denote the realized value of τ for a particular sample of size n. The statistic τ is assumed to be asymptotically pivotal, so that the bootstrap yields asymptotic refinements; see Beran (1988). A bootstrap DGP is used to generate B bootstrap samples, each of which is used to calculate a bootstrap test statistic τ_j^* for $j = 1, \ldots, B$.

There are several ways to calculate a bootstrap P value for $\hat{\tau}$. If τ is always positive, as it should be if it asymptotically follows a χ^2 or F distribution, then we normally want to reject when $\hat{\tau}$ is in the upper tail of the empirical distribution function (or EDF) of the τ_j^*. In this case, the bootstrap P value is

$$p^*(\hat{\tau}) = \frac{1}{B} \sum_{j=1}^{B} I(\tau_j^* > \hat{\tau}), \tag{1}$$

that is, the fraction of the bootstrap samples for which τ_j^* is larger than $\hat{\tau}$. If the level of the test is α, we reject the null hypothesis whenever $p^*(\hat{\tau}) < \alpha$.

When the τ_j^* follow precisely the same distribution as τ, this sort of test is called a Monte Carlo test. If B is chosen so that $\alpha(B+1)$ is an integer, Monte Carlo tests are exact; see Dufour and Khalaf (2001) for a review of the literature on Monte Carlo tests. In this paper, however, I will be concerned with the much more common situation in which the bootstrap distribution that the τ_j^* follow differs, in finite samples, from the distribution of τ.

When τ can take on either positive or negative values, for example, when it has the form of a t statistic, we often wish to perform a two-tailed test. In this case, there are two ways to proceed. The first is to assume that the distribution of τ is symmetric around zero in finite samples, just as it is asymptotically. This leads to the **symmetric bootstrap *P* value**

$$\hat{p}^*(\hat{\tau}) = \frac{1}{B} \sum_{j=1}^{B} I\big(|\tau_j^*| > |\hat{\tau}|\big). \tag{2}$$

Once again, we reject the null hypothesis when $p^*(\hat{\tau}) < \alpha$. There is evidently a close relationship between (1) and (2). Suppose that τ has the form of a t statistic, so that τ^2 is asymptotically $\chi^2(1)$. Then the P value for $\hat{\tau}$ based on (2) must be identical to the P value for $\hat{\tau}^2$ based on (1), when both are calculated using the same set of τ_j^*. Rejecting when $\hat{p}^*(\hat{\tau}) < \alpha$ is equivalent to rejecting when $|\hat{\tau}|$ is greater than the $1 - \alpha$ quantile of the EDF of the $|\tau_j^*|$.

The symmetry assumption may often be excessively strong. If we do not wish to assume symmetry, we can base a test on the **equal-tail bootstrap *P* value**

$$\hat{p}^*(\hat{\tau}) = 2 \min\left(\frac{1}{B} \sum_{j=1}^{B} I(\tau_j^* < \hat{\tau}),\ \frac{1}{B} \sum_{j=1}^{B} I(\tau_j^* > \hat{\tau}) \right). \tag{3}$$

Here we calculate P values for two one-tailed tests and reject if either of these P values is less than $\alpha/2$. In other words, we reject when $\hat{\tau}$ either falls below the $\alpha/2$ quantile or above the $1-\alpha/2$ quantile of the EDF of the τ_j^*. The factor of 2 is needed because it is twice as likely that $\hat{\tau}$ will be far out in either tail of the bootstrap distribution as that it will be far out in one specified tail. This procedure is equivalent to using the $\alpha/2$ and $1-\alpha/2$ quantiles of the τ_j^* as critical values. When τ has the form of a t statistic, the procedure is analogous to forming an equal-tail percentile t confidence interval. For a Monte Carlo test based on the equal-tail P value (3) to be exact, it is required that $(\alpha/2)(B+1)$ be an integer.

As we will see below, the power of tests based on symmetric and equal-tail bootstrap P values against certain alternatives may be quite different. Moreover, these tests may have different finite-sample properties when the distribution of τ is not symmetric around zero. There is reason to believe that the bootstrap may perform better for tests based on (2) than for tests based on (3), because the order of the bootstrap refinement is often higher for two-tailed than for one-tailed tests; see Hall (1992). Therefore, there may be more scope for the fast double bootstrap to improve the performance of tests based on (3).

3. Double Bootstrap and Fast Double Bootstrap Tests

The double bootstrap was proposed by Beran (1988). In various forms, it can be used for a variety of purposes, including the calculation of either confidence intervals or P values. Before discussing the fast double bootstrap, I describe how the double bootstrap itself can be used to calculate P values that should, at least in theory, be more accurate than ordinary bootstrap P values.

The first step is to generate B_1 first-level bootstrap samples that are used to compute bootstrap statistics τ_j^* for $j = 1, \ldots, B_1$. Then the ordinary bootstrap P value $p^*(\hat{\tau})$ is calculated using one of the formulae discussed in the previous section. For concreteness, let us assume that (1) is used. For each first-level bootstrap sample indexed by j, we obtain a second-level bootstrap DGP in essentially the same way as the first-level bootstrap DGP was obtained from the actual sample. Each second-level bootstrap DGP is then used to generate B_2 bootstrap samples that are used to compute test statistics τ_{jl}^{**} for $l = 1, \ldots, B_2$.

The next step is to compute the second-level bootstrap P value

$$p_j^{**} = \frac{1}{B_2} \sum_{l=1}^{B_2} I(\tau_{jl}^{**} > \tau_j^*) \tag{4}$$

for each first-level bootstrap sample j. This is the P value for the bootstrap statistic τ_j^* based on the EDF of the τ_{jl}^{**}. We then use the p_j^{**} to calculate the **double-bootstrap P value** as

$$p^{**} = \frac{1}{B_1} \sum_{j=1}^{B_1} I(p_j^{**} \leq p^*). \tag{5}$$

Thus p^{**} is equal to the proportion of the second-level bootstrap P values that are more extreme than the first-level bootstrap P value. The inequality in (5) is not strict, because, depending on the values of B_1 and B_2, there may well be cases for which $p_j^{**} = p^*$.

If $\hat{\tau}$, the τ_j^*, and the τ_{jl}^{**} all came from the same distribution, then the p_j^{**} should be uniformly distributed on the zero-one interval, and, as $B_1 \to \infty$, we should find that $p^{**} = p^*$. In this case, using the double bootstrap merely demonstrates that the ordinary bootstrap P value is valid in finite samples.

Suppose, instead, that the bootstrapping process causes the distribution of the τ_j^* to contain fewer extreme values than the distribution of τ itself. Therefore, the P values associated with moderately extreme values of $\hat{\tau}$ must be too small. But it is reasonable to expect that the distributions of the τ_{jl}^{**} contain even fewer extreme values than the distribution of the τ_j^*. Therefore, the p_j^{**} should tend to be too small, at least for small values of p_j^*. This implies that the double-bootstrap P value p^{**} will be larger than p^*, which is exactly what we want. By a similar argument, p^{**} will tend to be smaller than p^* when the distribution of the τ_j^* contains more extreme values than the distribution of τ itself.

A serious problem with the double bootstrap is that it is computationally very costly. For each of B_1 bootstrap samples, we need to compute $B_2 + 1$ test statistics. Thus the total number of test statistics that must be computed is $1 + B_1 + B_1 B_2$. For example, if $B_1 = 999$ and $B_2 = 399$, the double bootstrap will require the computation of no less than 399,601 test statistics.

The double bootstrap is costly because we need to generate B_2 second-level bootstrap samples for every first-level bootstrap sample. This is necessary because the distribution of the τ_{jl}^{**} may not be independent of τ_j^*. If we make the assumption that this distribution is independent of τ_j^*, we can dramatically reduce the cost of the procedure. This is the key assumption of the FDB procedure.

When we perform a fast double bootstrap test, only one second-level bootstrap statistic, τ_j^{**}, is calculated along with each τ_j^*. Let $\hat{Q}^{**}(1-p^*)$ denote the $1-p^*$ quantile of the τ_j^{**}. This quantile is defined implicitly by the equation

$$\frac{1}{B}\sum_{j=1}^{B} I\big(\tau_j^{**} > \hat{Q}^{**}(1-p^*)\big) = p^*. \tag{6}$$

Of course, for finite B, there will be a range of values of Q^* that satisfy (6), and we will need to choose one of them in a somewhat arbitrary manner. If $p^* = 0$, as may well be the case then the null hypothesis is false, then it seems natural to define $\hat{Q}^{**}(1-p^*) = \hat{Q}^{**}(1)$ as the largest observed value of the τ_j^{**}, although there are certainly other possibilities. Given these definitions, the **FDB P value** is

$$p_F^{**} = \frac{1}{B}\sum_{j=1}^{B} I\big(\tau_j^{*} > \hat{Q}^{**}(1-p^*)\big). \tag{7}$$

Thus, instead of seeing how often the bootstrap test statistics are more extreme than the actual test statistic, we see how often they are more extreme than the $1 - p^*$ quantile of the τ_j^{**}.

When the distribution of τ_{jl}^{**} does not depend on τ_j^*, there is a very close relationship between p_F^{**} defined in equation (7) and p^{**} defined in equation (5). In expectation, the FDB P value is simply

$$p_F^{**} = \Pr\big(\tau_j^* > Q^{**}(1 - p^*)\big), \tag{8}$$

where $p^* = \Pr(\tau_j^* > \hat{\tau})$, and $Q^{**}(1-p^*)$ is defined by the population analog of equation (6). The probability in (8) is being taken with respect to the distribution of the τ_j^*. In contrast, the double bootstrap P value is

$$\begin{aligned} p^{**} &= \Pr\big(\Pr(\tau_i^{**} > \tau_j^*) < p^*\big) \\ &= \Pr\Big(\Pr(\tau_i^{**} > \tau_j^*) < \Pr\big(\tau_i^{**} > Q^{**}(1 - p^*)\big)\Big), \end{aligned} \tag{9}$$

where the second equality in (9) uses the definition of $Q^{**}(1-p^*)$. The outer probability in the expression on the second line is being taken with respect to the distribution of the τ_j^*, and the inner ones with respect to the distribution of the τ_i^{**}. Since these two distributions are assumed to be independent, it is clear that

$$\Pr(\tau_i^{**} > \tau_j^*) < \Pr\big(\tau_i^{**} > Q^{**}(1 - p^*)\big)$$

if and only if

$$\tau_j^* > Q^{**}(1 - p^*).$$

Thus we see that, if B, B_1, and B_2 were infinite, the FDB P value (5) would be identical to the double bootstrap P value (7) when the distribution of the τ_{jl}^{**} does not depend on τ_j^*.

Of course, if we wished to reject when the test statistic was in the lower tail of the distribution, we would replace equation (7) by

$$p_F^{**} = \frac{1}{B} \sum_{j=1}^{B} I\big(\tau_j^* < \hat{Q}^{**}(1 - p^*)\big). \tag{10}$$

To obtain an equal-tail FDB P value, we would compute both (7) and (10) and then take twice the minimum of these two FDB P values, as in (3).

In most cases of interest to econometricians, the distribution of a test statistic is asymptotically independent of the distribution of the parameter estimates under the null hypothesis, and this also applies to many nonparametric and semiparametric bootstrap DGPs. Whether or not the fast double bootstrap will perform well depends on how much this asymptotic independence carries over to finite samples. Davidson and

MacKinnon (2001, 2002) studied several cases in which it apparently does. In particular, for the "J test" for nonnested linear regression models (Davidson and MacKinnon, 1981), which is treated as a one-tailed test, the FDB yields substantially more accurate results than ordinary bootstrap tests in cases where the latter are somewhat inaccurate. There have also been simulations by Lamarche (2004) in the context of testing for unknown structural breaks and by Omtzigt and Fachin (2002) in the context of cointegrated VARs.

It is not entirely clear how best to estimate $Q^{**}(1-p^*)$. Since p^*B must be an integer, by the way p^* is calculated, it seems natural to use number $(1-p^*)B$ in the sorted list of the τ_j^{**}. But this does not work when $p^* = 1$, and we then use the smallest of the τ_j^{**}. It is quite possible that different methods of estimating quantiles may affect the performance of FDB tests when B is not large.

It is of interest to see how well FDB tests work under ideal conditions, when τ, the τ_j^*, and the τ_j^{**} all come from the same distribution. To investigate this question, I generated all three statistics from the standard normal distribution for various values of B between 99 and 3999 and then calculated ordinary and FDB bootstrap P values. Ordinary bootstrap tests always rejected just about 5% of the time at the .05 level. So did the FDB tests, but only when B was sufficiently large. There was a very noticeable tendency for the FDB tests to reject too often when B was not large, and the overrejection was much more severe for equal-tail FDB tests than for symmetric or one-tailed tests.

To quantify this tendency, I regressed the difference between the FDB rejection frequency and the ordinary bootstrap rejection frequency on $1/(B+1)$ and $1(B+1)^2$, with no constant term. Table 1 presents the results of running these regressions for tests at the .05 level. These regressions fit very well, and there was no evidence that either a constant or higher-order terms were needed. There were 40 experiments for the one-tailed and symmetric tests and 32 experiments for the equal-tail tests. There were fewer experiments for the latter because values of B for which $.05(B+1)$ was an integer but $.025(B+1)$ was not (namely, 99, 299, and 499) had to be removed. Since each experiment used 500,000 replications, experimental error should be very small.

Table 1. Regressions for overrejection at .05 level when B is small

Test	$1/(B+1)$	$1/(B+1)^2$	$B = 99$	$B = 999$	$B = 1999$
Symmetric	0.3766 (.0214)	−10.04 (2.54)	0.002762	0.000367	0.000186
One-tailed	0.3658 (.0145)	−8.46 (1.72)	0.002812	0.000357	0.000181
Equal-tail	1.8020 (.0370)	−58.97 (8.20)	0.007536	0.001743	0.000886

In addition to coefficients and standard errors, Table 1 shows the fitted values from each of the regressions for $B = 99$, $B = 999$, and $B = 1999$. It can be seen that all the FDB tests, especially the equal-tail ones, tend to overreject when B is small. Precisely why this is happening is not clear, although it is probably related to the way that quantiles are estimated. In practice, however, overrejection should not be a problem, because any sensible investigator will use a large value of B whenever the bootstrap P value is not well above, or well below, the level of the test; see Davidson and MacKinnon (2000) for a discussion of how to choose B sequentially when it is expensive to calculate bootstrap test statistics.

4. Tests for Serial Correlation

Tests for serial correlation are widely used, but commonly-used tests are not exact in models with lagged dependent variables or nonnormal errors. Consider the linear regression model

$$y_t = \boldsymbol{X}_t\boldsymbol{\beta} + \gamma y_{t-1} + u_t, \quad u_t = \rho u_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \text{IID}(0, \sigma_\varepsilon^2), \tag{11}$$

where there are n observations, and $\boldsymbol{X}_t$ is a $1 \times k$ vector of observations on exogenous variables. The null hypothesis is that $\rho = 0$. A simple and widely-used test statistic for serial correlation in this model is the t statistic on $\hat{u}_{t-1}$ in a regression of y_t on $\boldsymbol{X}_t$, y_{t-1}, and $\hat{u}_{t-1}$. This procedure was proposed by Durbin (1970) and Godfrey (1978). The test statistic is asymptotically distributed as $\text{N}(0, 1)$ under the null hypothesis. Since this test can either overreject or underreject in finite samples, it is natural to use the bootstrap in an effort to improve its finite-sample properties.

In order to bootstrap the Durbin-Godfrey test under weak assumptions about the error terms, we first estimate the regression in (11) by ordinary least squares. This yields $\hat{\boldsymbol{\beta}}$, $\hat{\gamma}$, and a vector of residuals with typical element $\hat{u}_t$. We can then generate bootstrap data using the semiparametric bootstrap DGP

$$y_t^* = \boldsymbol{X}_t\hat{\boldsymbol{\beta}} + \hat{\gamma} y_{t-1}^* + u_t^*, \tag{12}$$

where the u_t^* are obtained by resampling the vector of rescaled residuals with typical element $(n/(n-k-1))^{1/2}\hat{u}_t$. The initial value y_0^* is set equal to the actual pre-sample value y_0. The bootstrap DGP (12) imposes the IID assumption on the error terms without imposing any additional distributional assumptions.

I performed a large number of experiments on the finite-sample properties of bootstrap versions of the Durbin-Godfrey test. In most of the experiments, the error terms were normally distributed, the first column of the $\boldsymbol{X}$ matrix was a constant, and the remaining columns were generated from independent, stationary AR(1) processes with parameter ρ_x. Both asymptotic and bootstrap rejection frequencies were found to depend strongly on k, ρ_x, σ_ε, and γ, as well as the sample size n. Since the performance of the asymptotic test improved rapidly as n increased, I chose to use $n = 20$ for most of the experiments. Asymptotic results are based on 200,000 replications for values of

γ between -0.99 and 0.99 at intervals of 0.01. Bootstrap results are based on 100,000 replications for values of γ between -0.9 and 0.9 at intervals of 0.1 using 1999 bootstrap samples. This is an unusually large number to use in a Monte Carlo experiment. It was used because the results in Table 1 suggest that the equal-tail FDB tests will tend to overreject noticeably if B is not quite large.

Results under the null

Figure 1 shows three sets of rejection frequencies for the performance of asymptotic and bootstrap tests under the null hypothesis when $n = 20$. These are representative of the results for a much larger number of similar experiments. Rejection frequencies for tests at the .05 level are shown on the vertical axis, and γ is shown on the horizontal axis. Each row concerns the same set of experiments. Results for the asymptotic test are shown in both panels. The left-hand panel shows rejection frequencies for symmetric bootstrap and FDB tests, and the right-hand panel shows rejection frequencies for equal-tail bootstrap and FDB tests.

The first row of the figure contains results for a case in which all the bootstrap tests work very well. In the left-hand panel, we see that there is very little difference between the rejection frequencies for the symmetric bootstrap test, based on (2), and for its FDB variant. This is not merely true on average, but for every replication: The correlation between the two P values was 0.999 for every value of γ. Thus an investigator who performed both tests would obtain extremely similar results and would probably conclude, quite justifiably, that the bootstrap P value was very reliable.

In the right-hand panel of the first row of the figure, we see that the rejection frequencies for the equal-tail bootstrap test are generally not quite as reliable as for the symmetric bootstrap test. Moreover, the FDB procedure yields noticeably different rejection frequencies which are, in most cases, closer to the nominal level of .05. The correlation between the two P values is still very high at approximately 0.996 for all values of γ.

The second and third rows of the figure show results for cases in which, on average, the bootstrap tests do not work as well. In both cases, $\sigma = 10$, which is ten times larger than for the case in the first row, and $k = 6$, which is twice as large. Thus the bootstrap DGP depends on more parameters, and they are estimated less precisely. The only difference between the two cases is that $\rho_x = 0.8$ in the second row and $\rho_x = -0.8$ in the third row.

Several interesting results are evident in the second and third rows of the figure. All four bootstrap tests generally work much better than the asymptotic test on which they are based. It is apparent that a symmetric bootstrap test can overreject when an equal-tail test underrejects, and *vice versa*. However, the equal-tail tests seem to be a bit more prone to overreject than the symmetric tests. The FDB tests generally work better than the ordinary bootstrap tests, especially when the latter are least reliable. Nevertheless, the correlations between the ordinary bootstrap and FDB tests remain quite high. They are never less than 0.976 for the equal-tail tests and 0.998 for the symmetric ones.

It is of interest to see how fast the performance of the ordinary bootstrap and FDB tests improves as the sample size increases. Figure 2 contains six panels, which are comparable to the six panels in Figure 1. In each of these experiments, γ is fixed at a value associated with relatively poor performance of at least one of the tests for $n = 20$, and n takes on the values 10, 14, 20, 28, 40, 56, 80, 113, 160, 226, and 320. As before, there were 100,000 replications, and $B = 1999$.

The left-hand panel of the first row shows that the symmetric bootstrap and FDB tests work extremely well for all sample sizes when $k = 3$ and $\sigma_\varepsilon = 1$. There is essentially nothing to choose between them. However, as can be seen from the right-hand panel, in this case the equal-tail tests tend to underreject for very small values of n, with the FDB tests underrejecting less severely than the ordinary bootstrap tests.

The next two rows of the figure are more interesting. We see both noticeable overrejection and noticeable underrejection by the ordinary bootstrap tests. With a few exceptions, the FDB tests perform substantially better than the ordinary bootstrap tests when the latter perform badly. The results in the right-hand panel of the second row and the left-hand panel of the third row are particularly dramatic. In these cases, the gain from using the FDB procedure is quite substantial.

It appears that the equal-tail FDB tests overreject slightly for large values of n. This appears to be a manifestation of the phenomenon that we saw in Table 1. Since the magnitude of the overrejection is just about what we would expect from the results in Figure 1, allowing for a certain amount of experimental error, it would surely be even smaller if B were larger than 1999.

Results under the alternative

Figure 3 shows power functions for six sets of experiments. The value of ρ is on the horizontal axis, and the rejection frequency is on the vertical axis. Asymptotic results are based on 200,000 replications for 199 values of ρ between -0.99 and 0.99, and bootstrap results are based on 100,000 replications for 19 values of ρ between -0.9 and 0.9. It is evident from every panel of the figure that using an FDB test rather than an ordinary bootstrap test has essentially no impact on power.

In the first two rows of the figure, $n = 20$. In the four panels in these rows, the shapes of the asymptotic power functions differ dramatically from the inverted bell shape that they must have asymptotically. The power functions for the symmetric bootstrap tests always have essentially the same shape as those for the asymptotic tests, although with a vertical displacement that is quite large in the case of the left-hand panel in the second row. This vertical displacement arises because the asymptotic test overrejects quite severely under the null hypothesis. The symmetric bootstrap test, which does not overreject, inevitably has noticeably less power against all alternatives.

In contrast, the shapes of the power functions for the equal-tail bootstrap tests are dramatically different from the shapes of the power functions for the symmetric bootstrap tests. The former have somewhat less power in whichever direction the asymptotic tests have high power, but they have much more power in the other direction. Specifically, when ρ_x and γ are both positive, the equal-tail tests always have more

power against positive values of ρ than the symmetric tests, and the differences are often dramatic. Since this is a case that we might expect to encounter quite frequently, this is an important result.

In the third row of the figure, $n = 40$. Increasing the value of n brings the shape of the asymptotic power functions much closer to the inverted bell shape that they should have, as can be seen by comparing the left-hand panel in the top row with the left-hand panel in the bottom row and the left-hand panel in the middle row with the right-hand panel in the bottom row. However, it does not change the results about the power of the symmetric and equal-tail bootstrap tests. The equal-tail tests have somewhat less power against negative values of ρ and a great deal more power against positive values than do the symmetric tests, because the power functions of the former are much closer to being symmetric about $\rho = 0$.

Notice that, in several panels of Figure 3, the asymptotic tests are reasonably reliable under the null. Nevertheless, there are very substantial gains in power to be had by using equal-tail bootstrap tests instead of asymptotic tests. This suggests that equal-tail bootstrap tests for serial correlation should be used routinely, even when there is no reason to believe that asymptotic tests are unreliable.

5. Tests for ARCH Errors

Since the seminal work of Engle (1982), it has been recognized that serial dependence in the variance of the error terms of regression models using times-series data is a very common phenomenon. In the case of financial data at high or moderate frequencies, there is not much point simply testing for ARCH errors, because we know that we will find strong evidence of them, whether or not ARCH is actually the best way to model the properties of the error terms. However, in the case of low-frequency financial data, or non-financial macroeconomic data, the hypothesis of serial independence is not unreasonable, and it may therefore make sense to test for ARCH errors.

Consider the linear regression model

$$y_t = \boldsymbol{X}_t\boldsymbol{\beta} + u_t, \quad u_t = \sigma_t\varepsilon_t, \quad \sigma_t^2 = \alpha_0 + \alpha_1 u_{t-1}^2 + \delta_1\sigma_{t-1}^2, \quad \varepsilon_t \sim \text{IID}(0,1). \tag{13}$$

The error terms of this model follow the GARCH(1,1) process introduced by Bollerslev (1986). It is easy to generalize this process to have more lags of u_t^2, more lags of σ_t^2, or both. In this paper, however, attention is restricted to the GARCH(1,1) process, partly for simplicity, and partly because this process generally works extraordinarily well in practice.

The easiest way to test the null hypothesis that the error terms are IID in the model (13) is to run the regression

$$\hat{u}_t^2 = b_0 + b_1\hat{u}_{t-1}^2 + \text{residual}, \tag{14}$$

where $\hat{u}_t$ is the t^{th} residual from an OLS regression of y_t on $\boldsymbol{X}_t$. The null hypothesis that $\alpha_1 = \delta_1 = 0$ can be tested by testing the hypothesis that $b_1 = 0$. For a simple

derivation of the test regression (14), and an explanation of why it has just two coefficients even though the GARCH(1,1) model has three, see Davidson and MacKinnon (2004, Section 13.6).

There are several valid test statistics based on regression (14). These include the ordinary t statistic for $b_1 = 0$, which is asymptotically distributed as $\mathrm{N}(0,1)$, and n times the centered R^2, which is asymptotically distributed as $\chi^2(1)$. Results are reported only for the second of these statistics, partly because it seems to be the most widely used test for ARCH errors, and partly because it generalizes easily to tests for higher-order ARCH and GARCH processes, in which there are more lags of $\hat{u}_t^2$ in the test regression. It would be interesting to compare the finite-sample performance of alternative tests, but that would involve a substantial digression.

The simulation experiments focused on the effects of the sample size and the distribution of the ε_t. In all the reported experiments, $\boldsymbol{X}_t$ consisted of a constant and two independent, standard normal random variates. The sample size took on the values 10, 14, 20, 28, 40, 56, 80, 113, 160, 226, or 320. The error terms were either standard normal, Student's t with 4 degrees of freedom, or centered $\chi^2(2)$. The first of these distributions is in some sense the base case, the second involves severe leptokurtosis, and the third involves severe skewness. Since the test statistics can easily be shown to be invariant to the variance of the error terms under the null hypothesis, that aspect of the experimental design was not varied. Changing the number of regressors had only a modest effect on the finite-sample behavior of the tests, and so all reported results are for a case in which there is a constant and two other regressors.

The top left-hand panel of Figure 4 shows rejection frequencies for the asymptotic test as a function of the sample size for the three different error distributions. These results are based on 100,000 replications for each value of n. The test underrejects severely in all cases, especially when the error terms are nonnormal. As we would expect, performance improves with the sample size, but the rate of improvement is fairly slow, especially when the errors are $t(4)$.

There are several ways to bootstrap this test. One possibility is to use a parametric bootstrap, drawing the error terms from the normal distribution. It is easy to see that this will lead to an exact test when the errors actually are normally distributed. The test statistic depends solely on the $\boldsymbol{X}$ matrix and the vector of innovations $\boldsymbol{\varepsilon}$. The former is known. If the distribution of the latter is known, then the test statistic does not depend on any unknown features of the DGP. It then follows by standard arguments for Monte Carlo tests that, when B is chosen so that $\alpha(B+1)$ is an integer, the parametric bootstrap test is exact; see Dufour *et al.* (2004).

The top right-hand panel of Figure 4 shows rejection frequencies for parametric bootstrap tests, with $B = 1999$. As expected, these tests work perfectly when the errors are actually normally distributed. The very small deviations from a frequency of 0.05 are well within the margins of experimental error. However, the tests are evidently not exact when the error terms are not normally distributed. For the largest sample sizes, they are no better than the corresponding asymptotic tests. Since the FDB

tests performed almost identically to the parametric bootstrap tests on which they are based, results for parametric FDB tests are not shown.

Of course, we would not expect parametric bootstrap tests to perform well when they are based on an incorrect distributional assumption, and we would not expect the FDB procedure to help. It therefore seems natural to use a semiparametric bootstrap DGP, like the one in equation (12). Results for this procedure are shown in the bottom left-hand panel of Figure 4 and in the two left-hand panels of Figure 5. For the normal distribution, the semiparametric bootstrap test underrejects, quite noticeably so for the smaller sample sizes. Interestingly, except for the very smallest sample sizes, the FDB version performs considerably better. It appears to be essentially exact for $n \geq 80$, whereas the ordinary semiparametric bootstrap test always underrejects to some extent.

The results are more interesting when the error terms are nonnormal. When they are $t(4)$, the semiparametric bootstrap test underrejects quite severely for small sample sizes. However, its performance gradually improves as n increases. More interestingly, there is a noticeable gain from using the FDB procedure, except when n is very small. When the error terms are centered $\chi^2(2)$, the underrejection is even more severe for small sample sizes, but the rate of improvement as n increases is much more rapid. Once again, there is generally a noticeable gain from using the FDB procedure.

The errors committed by the semiparametric bootstrap test must arise from the fact that the empirical distribution of the residuals provides an inadequate approximation to the distribution of the error terms. One way to improve this approximation is to smooth the bootstrap errors. This can be done by using a kernel estimator. The kernel estimator of the CDF of u at the point u', using a sample of n residuals $\hat{u}_t$, is given by

$$\hat{F}_h(u) = \frac{1}{n} \sum_{t=1}^{n} K(\hat{u}_t, u', h), \tag{15}$$

where $K(\hat{u}_t, u', h)$ is a cumulative kernel, such as the standard normal CDF, called the Gaussian kernel, and h is the bandwidth; see Azzalini (1981) and Reiss (1981). A reasonable choice for h is $1.587\hat{\sigma}n^{-1/3}$, where $\hat{\sigma}$ is the standard deviation of the (possibly rescaled) residuals.

To draw bootstrap errors from (15), we simply resample from the residuals $\hat{u}_t$ and then add independent normal random variables with variance h^2. The resulting bootstrap errors have too much variance, but they can easily be rescaled. However, in the context of tests for ARCH errors, this rescaling is not needed, because the test statistics are invariant to the variance of the error terms.

The bottom right-hand panel of Figure 4 and the two right-hand panels of Figure 5 show the effects of using bootstrap errors that were smoothed in this way, where the Gaussian kernel with the bandwidth given above was used. When the error terms are actually normal, resampling smoothed residuals works substantially better than resampling ordinary residuals for small sample sizes. However, there appears to be no appreciable gain from smoothing when the errors are $t(4)$ or centered $\chi^2(2)$. Whether

or not smoothing is used, the FDB procedure always brings the rejection frequency noticeably closer to 0.05, except when the sample size is very small.

Because Figures 4 and 5 deal only with tests at the 0.05 level, they do not tell the whole story. To show the effect of the level of the test, Figure 6 plots the difference between the rejection frequency and the level of the test for all levels between 0.005 and 0.25 for two sample sizes, 40 and 160. The nominal level is on the horizontal axis, and the "rejection frequency discrepancy" is on the vertical axis. Several interesting facts emerge from this figure. First, the asymptotic test can actually overreject for small levels. Second, for nonnormal errors, the distortion of the asymptotic test becomes steadily worse as the level increases. Finally, and of most interest for this paper, the improvement from using the FDB rather than the ordinary bootstrap becomes larger as the level of the test increases. Moreover, the extent of the improvement is greater for $n = 160$ than for $n = 40$, especially in relative terms. To see this, compare the left-hand and right-hand panels in the second and third rows of the figure.

6. Conclusion

In this paper, I have provided evidence on the performance of fast double bootstrap tests for two cases of interest to econometricians, namely, tests for serial correlation and tests for ARCH errors. The fast double bootstrap is certainly not a silver bullet. In the experiments reported here, it never transforms a bootstrap test that performs poorly into a test that performs perfectly across a wide range of parameter values. However, it never worsens the performance of a bootstrap test more than marginally, and it often improves performance quite noticeably. In many cases, the improvement is as great as what could otherwise be achieved by doubling or even quadrupling the sample size. Moreover, when an ordinary bootstrap test works well, the FDB procedure usually yields a very similar P value, and it thus provides some reassurance that the bootstrap test is reliable.

It is important that B be reasonably large when using the fast double bootstrap. As we saw in Section 3, the procedure tends to overreject when B is small. This affects simulation experiments like the ones reported here more than it does practitioners who actually use the FDB. For most bootstrap tests, one can safely use a value of B like 99 or 199 when investigating test performance under the null by simulation, but that is not true for FDB tests.

One interesting result of the experiments discussed in Section 4 is that equal-tail bootstrap tests can be much more powerful than either asymptotic tests or symmetric bootstrap tests, even when the asymptotic tests are well-behaved under the null. Although this result is not specifically related to FDB tests, it does suggest that equal-tail bootstrap tests deserve closer investigation for a variety of problems where two-tailed tests are commonly used.

References

Azzalini, A. (1981). "A note on the estimation of a distribution function and quantiles by a kernel method," *Biometrika*, **68**, 326–328.

Beran, R. (1988). "Prepivoting test statistics: a bootstrap view of asymptotic refinements," *Journal of the American Statistical Association*, **83**, 687–697.

Bollerslev, T. (1986). "Generalized autoregressive conditional heteroskedasticity," *Journal of Econometrics*, **31**, 307–27.

Davidson, R., and J. G. MacKinnon (1981). "Several tests for model specification in the presence of alternative hypotheses," *Econometrica*, **49**, 781–793.

Davidson, R., and J. G. MacKinnon (1999). "The size distortion of bootstrap tests," *Econometric Theory*, **15**, 361–376.

Davidson, R., and J. G. MacKinnon (2000). "Bootstrap tests: How many bootstraps?" *Econometric Reviews*, **19**, 55–68

Davidson, R., and J. G. MacKinnon (2001). "Improving the reliability of bootstrap tests," Queen's Institute for Economic Research Discussion Paper No. 995, revised.

Davidson, R., and J. G. MacKinnon (2002). "Fast double bootstrap tests of nonnested linear regression models," *Econometric Reviews*, **21**, 417–427.

Davidson, R., and J. G. MacKinnon (2004). *Econometric Theory and Methods*, New York, Oxford University Press.

Dufour, J.-M., and L. Khalaf (2001). "Monte Carlo test methods in econometrics," Ch. 23 in *A Companion to Econometric Theory*, ed. B. Baltagi, Oxford, Blackwell Publishers, 494–519.

Dufour, J.-M., L. Khalaf, J.-T. Bernard, and I. Genest (2004). "Simulation-based finite-sample tests for heteroskedasticity and ARCH effects," *Journal of Econometrics*, forthcoming.

Durbin, J. (1970). "Testing for serial correlation in least-squares regression when some of the regressors are lagged dependent variables," *Econometrica*, **38**, 410–421.

Engle, R. F. (1982). "Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation," *Econometrica*, **50**, 987–1007.

Godfrey, L. G. (1978). "Testing against general autoregressive and moving average error models when the regressors include lagged dependent variables," *Econometrica*, **46**, 1293–1301.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. New York: Springer-Verlag.

Lamarche, J.-F. (2004). "The numerical performance of fast bootstrap procedures," *Computational Economics*, **23**, 379–389.

Omtzigt, P., and S. Fachin (2002). "Bootstrapping and Bartlett corrections in the cointegrated VAR model," University of Amsterdam Discussion Paper No. 2002/15.

Reiss, R. D. (1981). "Nonparametric estimation of smooth distribution functions, *Scandinavian Journal of Statistics*, **9**, 65–78.

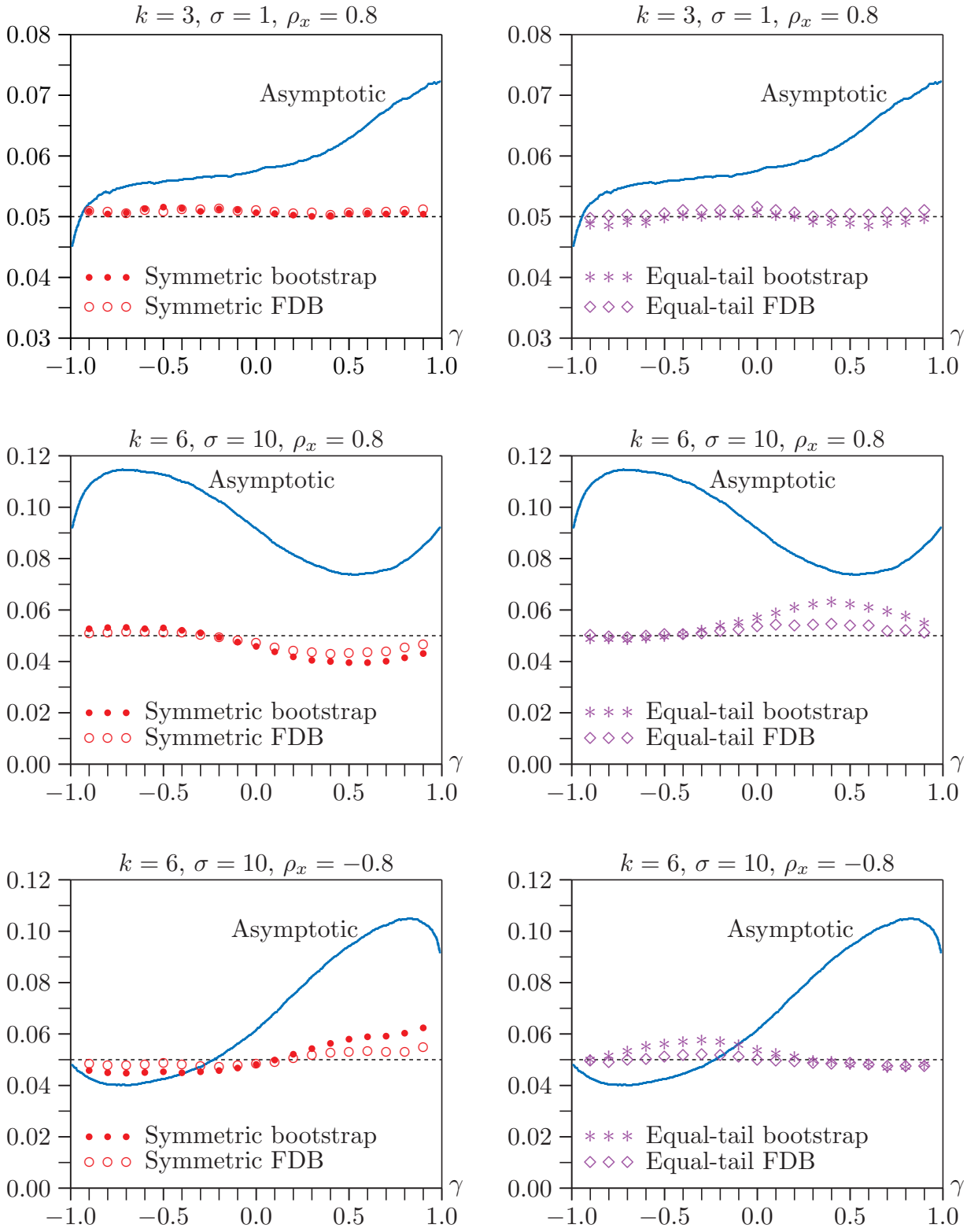

Figure 1. Durbin-Godfrey rejection frequencies at .05 level under the null, $n = 20$

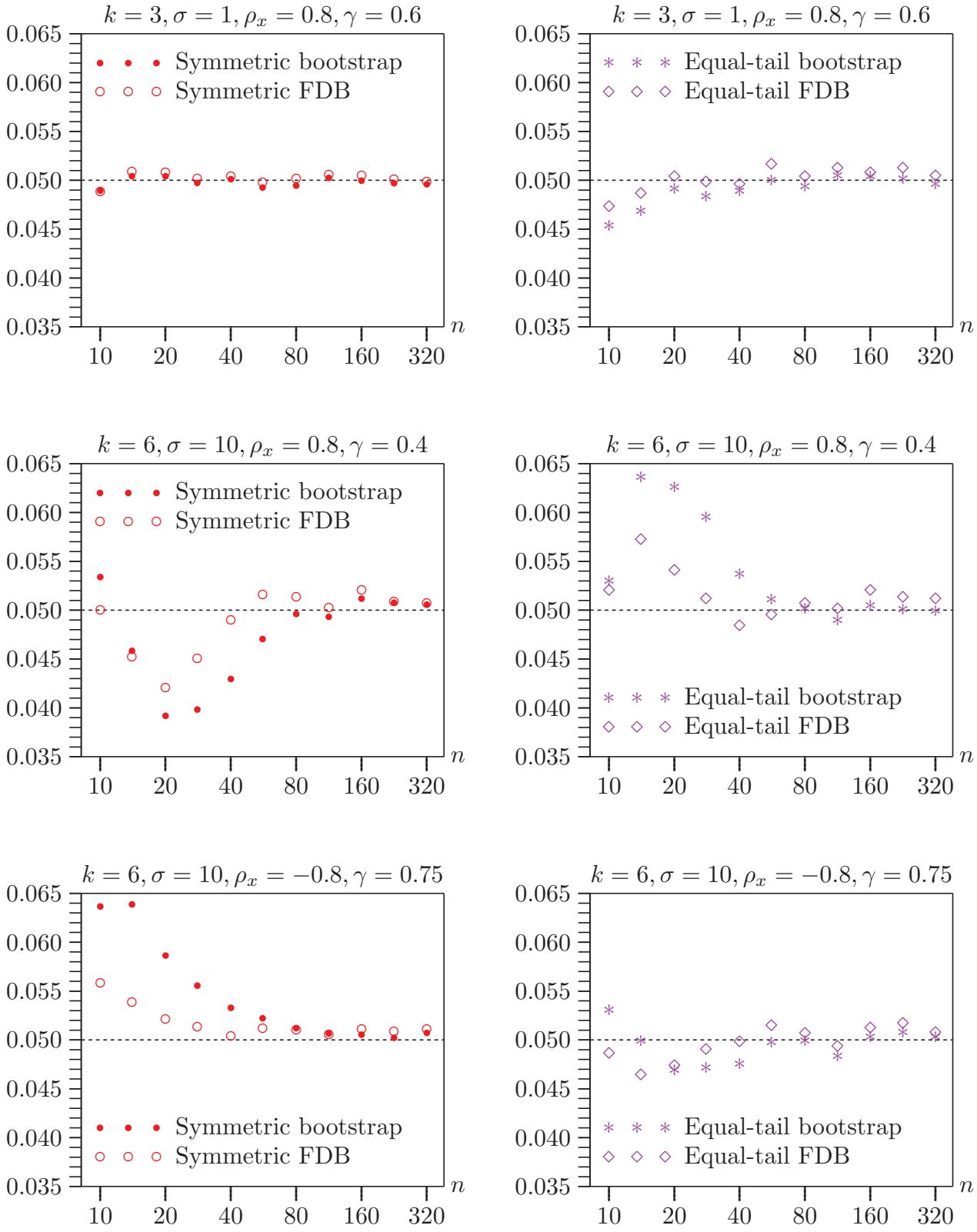

Figure 2. Durbin-Godfrey rejection frequencies at .05 level under the null

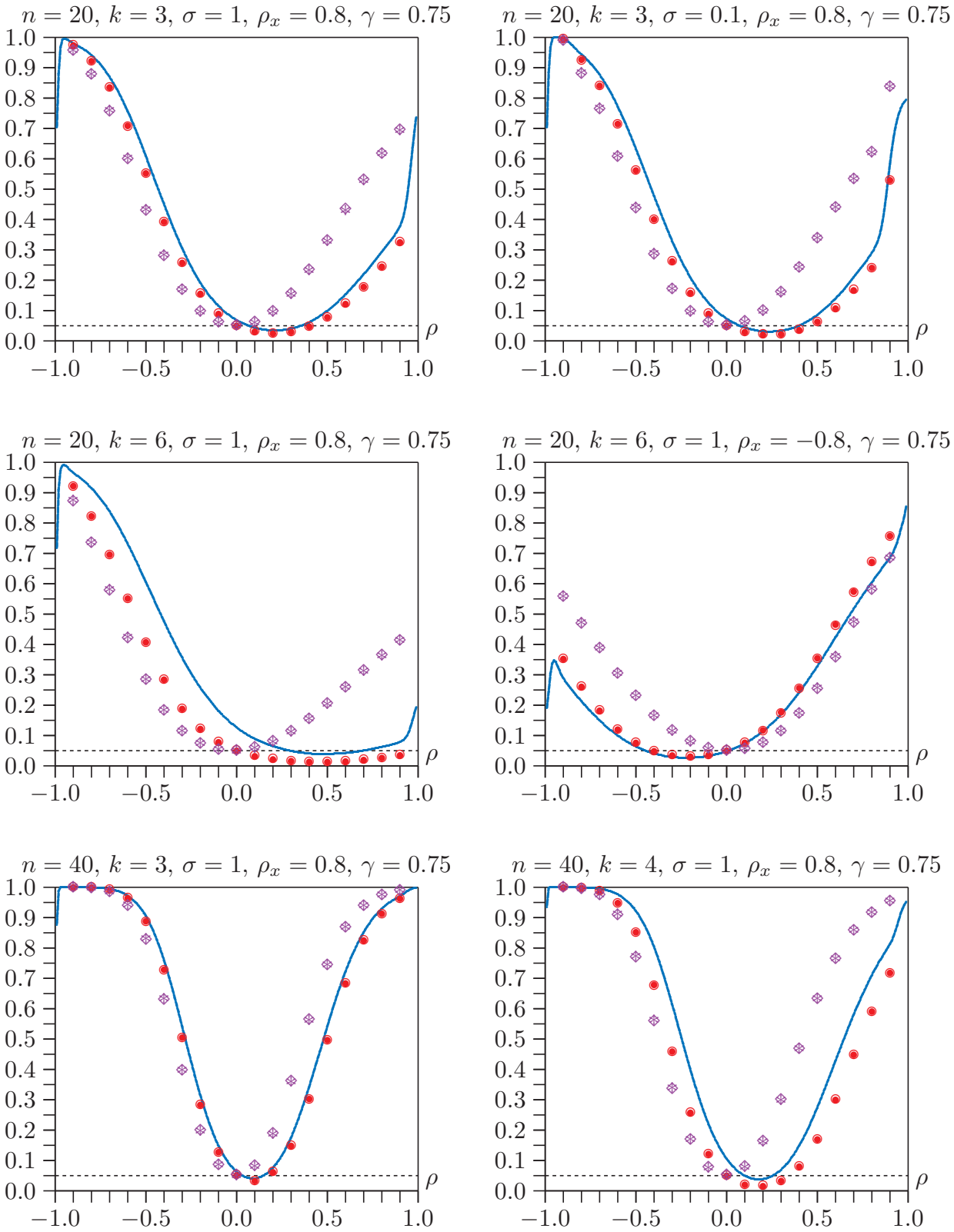

Figure 3. Power of Durbin-Godfrey tests at .05 level

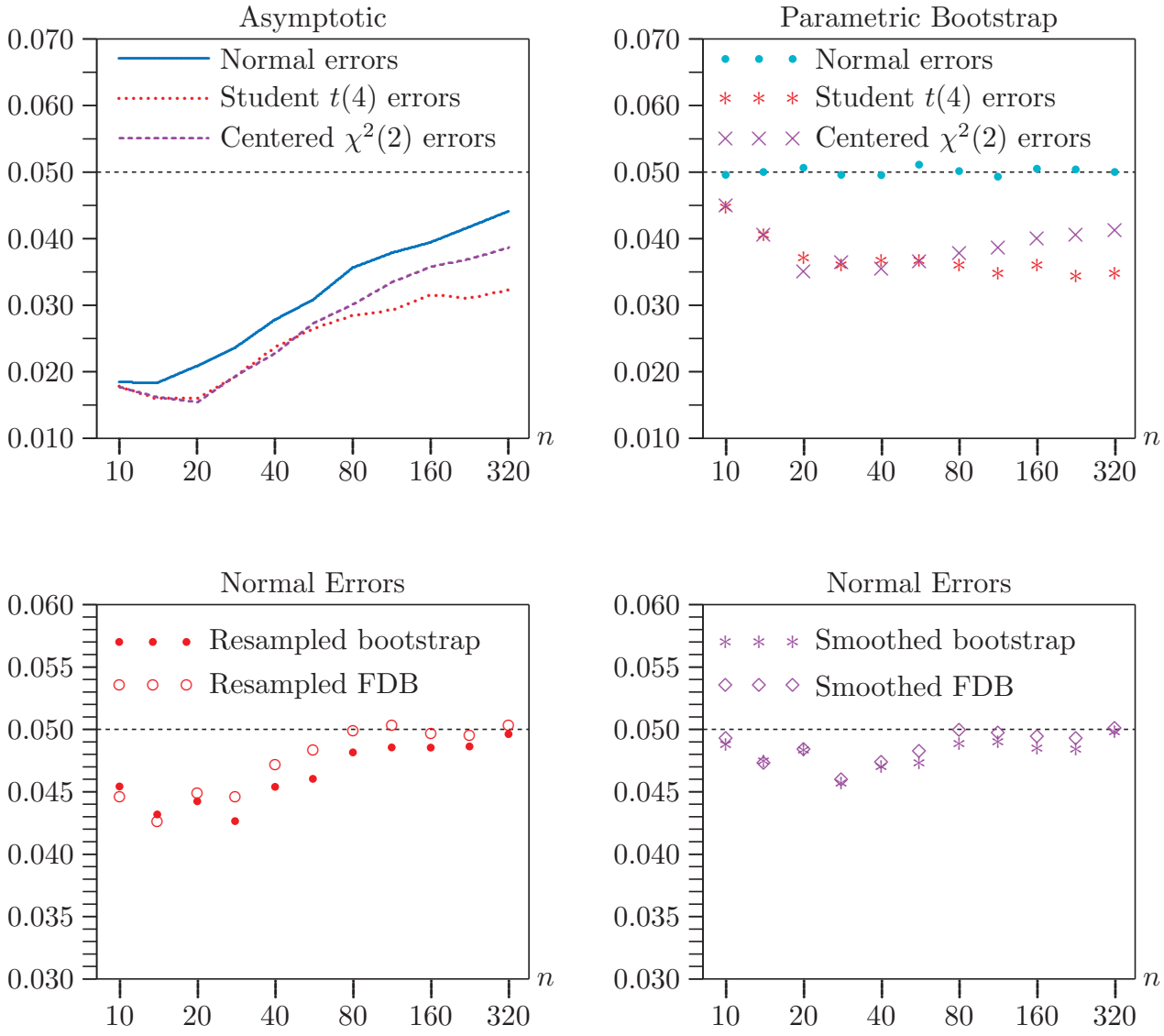

Figure 4. ARCH test rejection frequencies at .05 level under the null

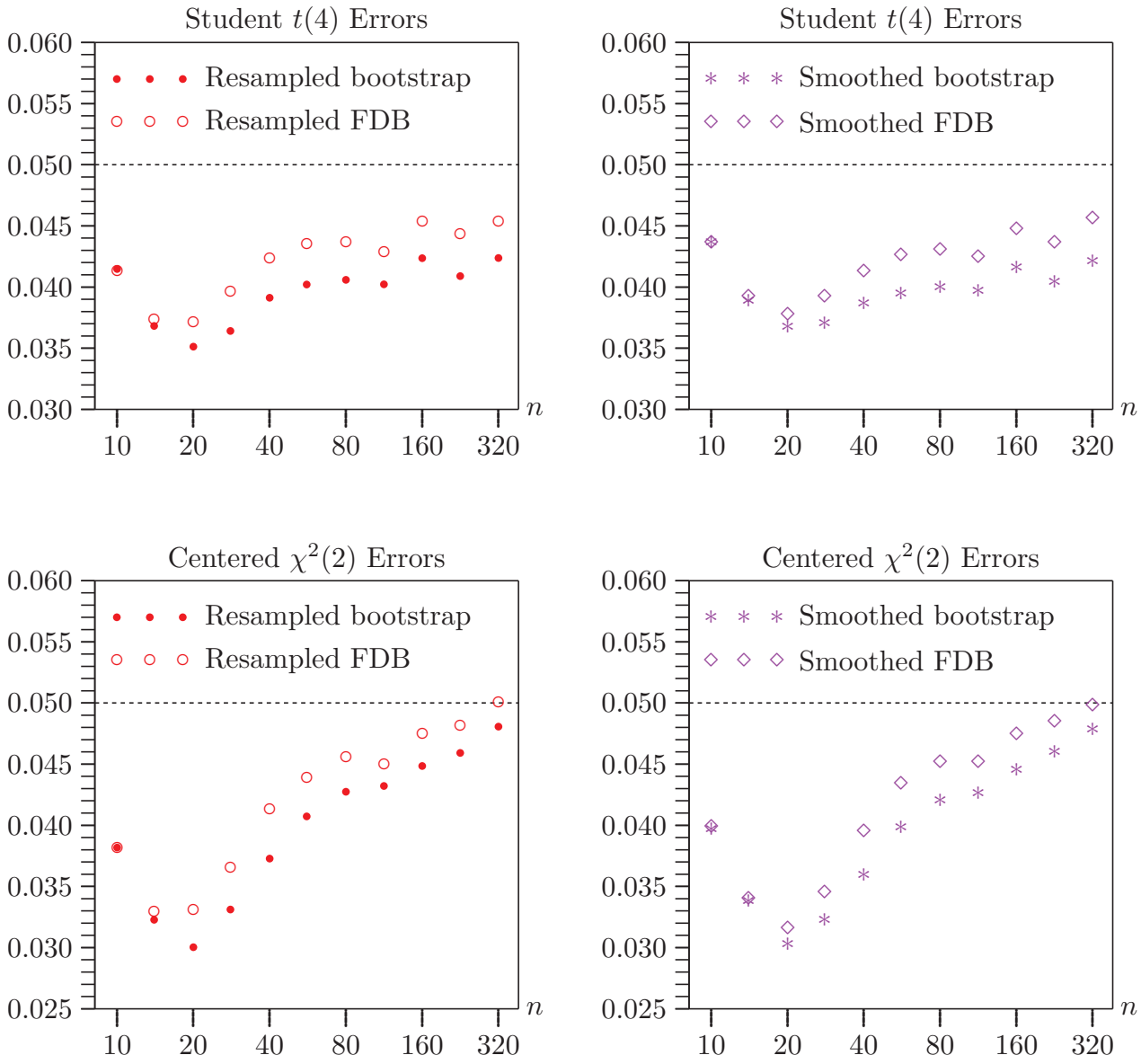

Figure 5. ARCH test rejection frequencies at .05 level under the null

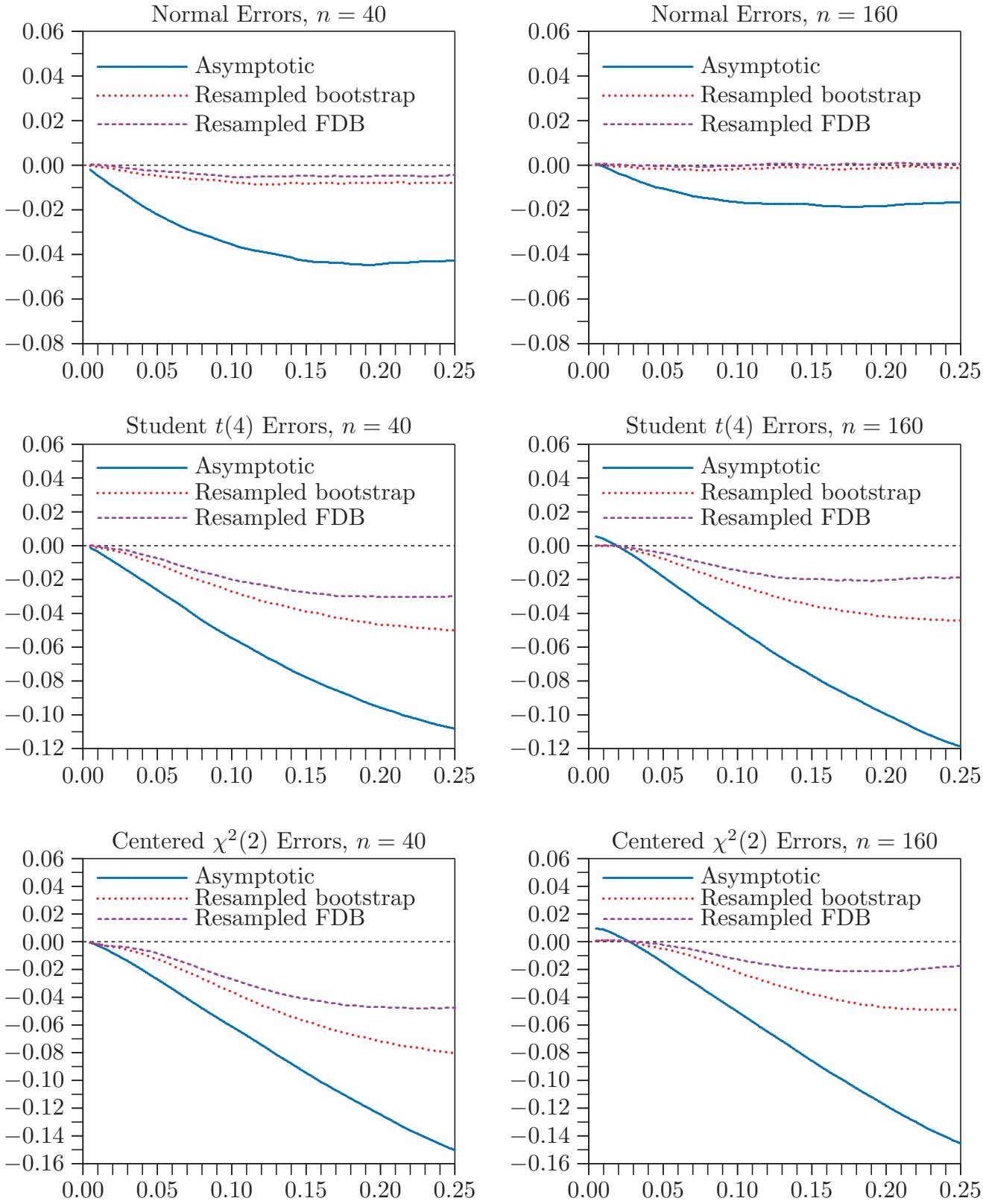

Figure 6. Rejection frequency discrepancy plots for ARCH tests