Improving Fractional Integration Tests with Bootstrap Distributions

Michael K. Andersson
National Institute of Economic Research, Sweden
and

Mikael P. Gredenhoff
AGL Structure Finance, Sweden

Working Paper No. 74
National Institute of Economic Research

May 23, 2000

Abstract

Asymptotic tests for fractional integration are usually badly sized in small samples, even for normally distributed processes. Furthermore, tests that are well-sized under normality may be severely distorted by non-normalities and ARCH errors. This paper demonstrates how the bootstrap can be implemented to correct for such size distortions.

Key words: Long memory; Resampling; Skewness and kurtosis; ARCH; Size correction; Power.
JEL-classification: C12; C15; C22; C52.

1 Introduction

Many financial time series display observations that are non-normally distributed (i.e. with excess skewness and kurtosis), conditionally heteroskedastic and ruled by long memory. For instance Ding, Granger and Engle (1993) report evidence of autocorrelations between distant lags in the absolute returns of the Standard and Poor 500, S&P500, composite stock index. Furthermore, Granger and Ding (1995) show that the absolute value of the rate of return for a variety of stock prices, commodity prices and exchange rates exhibit excess skewness and kurtosis.

Long memory is usually described by a fractionally integrated specification, hence testing for long memory may be performed via a test for a fractional differencing power. The asymptotic tests often exhibit non-negligible size distortions in small samples. To improve inference, this paper suggests parametric bootstrap methods to adjust the critical values. We seek tests that are robust to short-term dependencies, non-normalities and ARCH effects in data. The performance of a bootstrap test depends to some extent on the chosen resampling algorithm, it has better size properties than the corresponding asymptotic test.

The paper unfolds as follows. Section 2 introduces the fractional integration bootstrap testing procedure and Sections 3 and 4 presents the Monte Carlo setup and the size and power of the tests. Section 5 concludes the paper.

2 The Fractional Integration Bootstrap Test

2.1 Testing for Fractional Integration

A fractionally integrated autoregressive moving average (ARFIMA) time series process $\{x_t\}$ is described by the equation

$$\phi(B)(1-B)^d x_t = \theta(B) a_t, \qquad t = 1, ..., T \tag{1}$$

where the roots of $\phi(B)$ and $\theta(B)$ are all outside the unit circle and a_t is *iid* with mean zero and finite variance σ_a^2. The differencing parameter d is allowed to take any real number, but if d is restricted to the set of integers the specification (1) reduces to an ARIMA process. The sample autocorrelation function of a long memory process may be approximated by a fractionally integrated model, hence testing for long-memory can be done by a test on d. Such tests are applicable on stationary and invertible[1] series, and the series should be differenced or summed until this is satisfied. Thus, $d = 0$ is a natural null-hypothesis when testing for fractional integration.

We consider three tests, namely, the log periodogram regression of Geweke and Porter-Hudak (GPH, 1983), the modified rescaled range (MRR) statistic, Lo (1991), and the LM test, denoted REG, of Agiakloglou and Newbold (1994). Cheung (1993) reports evidence of serious size distortions for the tests. Our idea is to correct for the distortions using bootstrap methods.

2.2 The Bootstrap Test

The bootstrap, described by Efron and Tibshirani (1993), provides a feasible method for estimation of the small-sample distribution of a statistic. If the exercise is bootstrap hypothesis testing the bootstrap samples must obey the null hypothesis and, as far as possible, resemble the real sample.

Asymptotic theory is only exact if the p-value is independent of the actual data generating process and sample size, which is usually not the case. A small sample solution is to replace the p-value by the bootstrap counterpart, which can be estimated as

$$\hat{p}^* \left(\hat{\tau} \right) = \frac{1}{R} \sum_{r=1}^{R} I \left(\left| \tau_r^* \right| \geq \left| \hat{\tau} \right| \right), \tag{2}$$

where R is the number of bootstrap replicates, $I(\cdot)$ the usual zero/one indicator function, $\hat{\tau}$ a realized value of the test statistic τ based on a sample $\mathbf{x} = \{x_1, ..., x_T\}$ and τ_r^* the value of the same test statistic, based on the bootstrap resample $\mathbf{x}_r^* = \{x_{r1}^*, ..., x_{rT}^*\}$.

The theory of bootstrap testing is given by Davidson and MacKinnon $(1996, 1999a)$. They show that if the test statistic is (asymptotically) pivotal, that is independent of nuisance parameters, the size-distortion refinement is of order $T^{-1/2}$ when using the bootstrap p-value compared to the corresponding asymptotic. A further refinement, usually also of order $T^{-1/2}$ is obtained whenever the test statistic is independent of the bootstrap DGP. Moreover, the power of a bootstrap test, based on a pivotal statistic, is generally close

[1] Stationarity and invertibility require that $d < |1/2|$. The ARFIMA model is presented in greater detail by Granger and Joyeux (1980) and Hosking (1981).

to the size-adjusted asymptotic test. Even if the statistic is only close to pivotal this is true in most cases.

Davidson and MacKinnon (1999*b*) demonstrate that a small number of bootstrap replicates imply a loss in power and that at least $R = 399$ is required to guarantee a power loss of no more than 1% at the 0.05 level. The size of a bootstrap test is less sensitive to the number of replicates. We use 399 replicates, but we have also compared the results with $R = 999$ for a selection of the parameter values considered. The size and power are altered only slightly when using larger number of bootstrap samples.

2.3 Construction of the Bootstrap Samples

For the construction of the bootstrap samples we use a model-based approach, which is natural since a well-defined model constitutes the null hypothesis. The bootstrap B_A, which is constructed to preserve ARCH(1) dependence in the residuals, is conducted as follows:

1. Estimate the AR(p)-ARCH(1) model

$$\begin{aligned}\left(1-\phi_0-\phi_1 B-...-\phi_p B^p\right)(x_t) &= a_t, \quad a_t\left|\mathcal{I}_{t-1}\sim N\left(0,\omega_t\right)\right. \\ \omega_t &= \beta_0+\beta_1 a_{t-1}^2\end{aligned} \tag{3}$$

   which clearly obeys the null-hypothesis, which is crucial. The autoregressive order p is determined by the BIC and the parameters are estimated through maximization of the log-likelihood function according to the Davidon-Fletcher-Powell algorithm, see Press *et al* (1992).

2. Due to the assumed normality of the disturbances a_t in (1), the bootstrap residuals $\{a_t^*\}$ are constructed accordingly; let ε_t^* be an independent draw from a $N(0,1)$ distribution, and compute

$$\begin{aligned}\hat{\omega}_t &= \hat{\beta}_0+\hat{\beta}_1 a_{t-1}^{*2} \\ a_t^* &= \varepsilon_t^*\sqrt{\hat{\omega}_t}.\end{aligned}$$

3. Finally, the bootstrap samples $\mathbf{x}_r^*$, $r = 1,...,399$, are created by the recursion

$$x_{r,t}^* = \hat{\phi}(B)^{-1} a_t^*, \tag{4}$$

   where $\hat{\phi}(B)$ is the estimated polynomial of (3).

Of course, the procedure is not limited to ARCH(1) errors, it can easily be extended to unknown lag-orders and GARCH processes. The reason for only considering ARCH(1) processes and a pre-specified lag-order is purely time saving, bootstrap-Monte Carlo studies are computationally demanding.

For the sake of comparison, a simple bootstrap version, which ignores the ARCH, is also included. This resampling, denoted b_S, draws residuals a_t^* independently and directly from a normal distribution with mean zero and variance $s_{\hat{a}}^2$ and the resamples are created using eq. (4).

Moreover, beside the two resampling schemes above we have also tried two nonparametric ones: one resampling where ARCH in errors are incorporated and one simple nonparametric.

3 Monte Carlo Design

The experiment covers first-order autoregressions and fractional noise series of lengths $T = 50$, 100, 200, 300 and 500. We generate $T + 100$ error observations by the IMSL routine (D)RNNOA, and discard the first 100 observations to reduce the initial value effect. The AR series are then constructed recursively and the fractional noise series are generated by Cholesky factorization of the ARFIMA error covariance matrix.

The Monte Carlo study is programmed in FORTRAN (using the Digital Visual Fortran 5.0 compiler) and involves 10,000 trials (series). Each series is tested for fractional integration using the tests in Sections 2. Estimated size and power of the different processes are computed as the rejection frequencies of the non-fractional null hypothesis.

The size is examined for the autoregressive datagenerating process (DGP)

$$x_t = \phi x_{t-1} + a_t \tag{5}$$

and the power is studied using data constructed by

$$(1 - B)^d x_t = e_t. \tag{6}$$

The members of $\{a_t\}$ and $\{e_t\}$ are iid $N(0, 1)$. We report results for ϕ equals $0.0, 0.5$ and 0.9 and d equal to ± 0.05, ± 0.25 and ± 0.45.

Besides normality of the disturbances, we also construct data which display non-normality (excess skewness and kurtosis) and ARCH errors.

In the non-normal case, the disturbances a_t (and e_t) are distributed with skewness and kurtosis equal to γ_s and γ_k respectively, by the Fleichmann (1978) transformation. γ_s and γ_k are chosen to generate series x_t with a skewness and kurtosis of 2 and 9.

For the final set of processes the disturbances are conditionally distributed as $a_{t|t-1} \sim N(0, \omega_t)$, where $\omega_t = 1 - \beta + \beta a_{t-1}^2$ and β is selected as 0.5 and 0.9. e_t is constructed equvivalently.

4 Results

Table 1 presents the sensitivity, at a nominal 5% level of significance, of the empirical size of the tests. The power of the bootstrap tests are presented in Figure 1. The reason for not including the power of the asymptotic tests is that we only compare tests that are (approximately) well sized.

The *estimated size* of the asymptotic *Geweke and Porter-Hudak, GPH, test* differ from the nominal, when the autoregressive parameter assumes a large positive value. Furtermore, the test is robust to excess skewness and kurtosis, and conditionally heteroskedastic errors, in the sense that the results are similar to the normal case.

The size problems of the GPH test are adjusted by both bootstrap procedures, and the bootstraps do not impose distortions where the asymptotic test is correct in size. The robustness of the asymptotic test can be detected in the bootstrap procedures since the simple bootstrap works as well as the ARCH resampling for all investigated combinations of ϕ and β.

For short series, the bootstrap GPH tests have a very poor ability to detect a fractional difference no matter of the size of the differencing parameter. The *power* increases steadily with the sample size and at $T = 300$ we are likely to track down (in particular positive) fractional integration. Throughout, we notice that the ARCH bootstrap has a slightly lower power than the simple one, thus we conclude that the GPH test shall be combined with the simple bootstrap.

The *size* of the asymptotic *Modified Rescaled Range test* also differs from the nominal, notably, conservatively for large positive parameters. The size distortion decreases as T grows, but the rate of convergence appears to be slow. The MRR test is fairly robust to excess skewness and kurtosis, and compared with the case of no heteroskedasticity in errors the test tends to be more conservative as the ARCH parameter increases.

In general, the bootstrap MRR testing procedure is able to improve the asymptotic test; every bootstrap test has better size properties than any of the original. In more detail, the bootstrap MRR test is, for the smallest sample size, still conservative when $\phi = 0.9$. However, the estimated size approaches the nominal as the serial length increases. The simple resamplings work satisfactorily for normal and non-normal processes, but exhibit empirical sizes a bit below five percent given heteroskedasticity. Under ARCH, the b_A bootstrap is close to exact, and is also works well when the error does not follow the ARCH(1) specification.

The bootstrap MRR tests have well-behaved *power curves*, and appears quite powerful when testing for negative fractional integration. For normal and skewed errors we see that b_A generally exhibits a lower power than b_S. In the ARCH case, there is no most powerful bootstrap test and the test is almost useless when testing for negative differences, that is $d < 0$.

The *estimated size* of the asymptotic *Lagrange Multiplier REG test* is very close to the nominal five percent when the observations are normally distributed. For skewed data with excess kurtosis, the empirical size is greater than the nominal, but the difference is reduced as the sample size increases. The REG test is also very sensitive to ARCH effects and exhibits a seriously distorted size for $\beta = 0.5$ and in particular for $\beta = 0.9$, and the distortion grows with T.

Exactly as the asymptotic test, the bootstrap REG test based on the simple parametric resampling is well-sized for normal processes. The resamplings that correct for the (non-existing) ARCH have reasonable, but conservative, sizes. This is also the case for non-normal data. Given ARCH, the bootstrap b_A is not only superior to the original test, but also much better than b_S.

The bootstrap REG test reduces in power when the true d is close to 0.5 compared to a slightly lower d value. When specifying the test, a large fractional differencing power is interpreted as a large autoregressive (or moving average) order, yielding decreased rejection frequencies. For normally distributed fractional noise, the ARCH bootstraps have a lower power than the simple parametric b_1. This is also the case for fractionally integrated white noise with ARCH disturbances, but then the simple algorithm is not advisable due to the size distortion.

5 Conclusions

The concept of bootstrap testing for fractional integration works extraordinarily well. If the significance level is calculated by a bootstrap procedure a well-sized test is almost always the result. However, the choice of resampling algorithm may affect the degree of size adjustment. For instance, if the original test is sensitive to distributional assumptions, in particular ARCH effects, this should be accounted for when specifying the resampling model. If the test is robust to ARCH errors, the choice of resampling is not very important for the size properties of that test.

Since economic and financial data are often heteroskedastic we recommend the use of an ARCH resampling scheme for the REG test. On the other hand, if prior information suggests that the investigated series does not have ARCH effects, the simple parametric bootstrap has better size and power properties.

The MRR and GPH tests, which are robust to deviations from the iid normality of the disturbances, have nice size properties for all bootstrap procedures. Due to the simplicity and the slightly higher power of the simple algorithms, they are preferred when bootstrapping the MRR and GPH tests.

The main conclusions are that the bootstrap tests are remarkably well-sized (whereas the asymptotic tests are not) and robust to non-normalities and ARCH effects, and that reliable testing for fractional integration in many cases requires a bootstrap test.

Acknowledgements

We are greatly indebted to Tor Jacobson, Sune Karlsson, James G. MacKinnon and Paul Newbold for useful discussions and comments. The usual disclaimer applies.

References

Agiakloglou, C. and P. Newbold, 1994, Lagrange multiplier tests for fractional difference, Journal of Time Series Analysis 15, 253-262.

Cheung, Y.-W., 1993, Tests for fractional integration: a Monte Carlo investigation, Journal of Time Series Analysis 14, 331-345.

Davidson, R. and J.G. MacKinnon, 1996, The power of bootstrap tests, Queens Institute for Economic Research Discussion Paper No. 937. (Department of Economics, Queens University, Kingston, Canada).

Davidson, R. and J.G. MacKinnon, 1999*a*, The size distortion of bootstrap tests, *Econometric Theory*, forthcoming.

Davidson, R. and J.G. MacKinnon, 1999*b*, Bootstrap Tests: How many Bootstraps?, *Econometric Theory*, forthcoming.

Ding, Z., C.W.J. Granger, and R.F. Engle, 1993, A long memory property of stock market returns and a new model, Journal of Empirical Finance 1, 83-106.

Efron, B. and R.J. Tibshirani, 1993, An introduction to the bootstrap (Chapman and Hall, New York).

Fleichmann, A.I., 1978, A method for simulating non-normal distributions, Psychometrika 43, 521-532.

Geweke, J. and S. Porter-Hudak, 1983, The estimation and application of long memory time series models, Journal of Time Series Analysis 4, 221-238.

Granger, C.W.J. and Z. Ding, 1995, Stylized facts on the temporal and distributional properties of daily data from speculative markets, Unpublished manuscript, Department of Economics, University of California, San Diego.

Granger, C.W.J. and R. Joyeux, 1980, An introduction to long-memory time series models and fractional integration, Journal of Time Series Analysis 1, 15-29.

Hosking, J.R.M., 1981, Fractional differencing, Biometrika 68, 165-176.

Lo, A.W., 1991, Long term memory in stock market prices, Econometrica 59, 1279-1313.

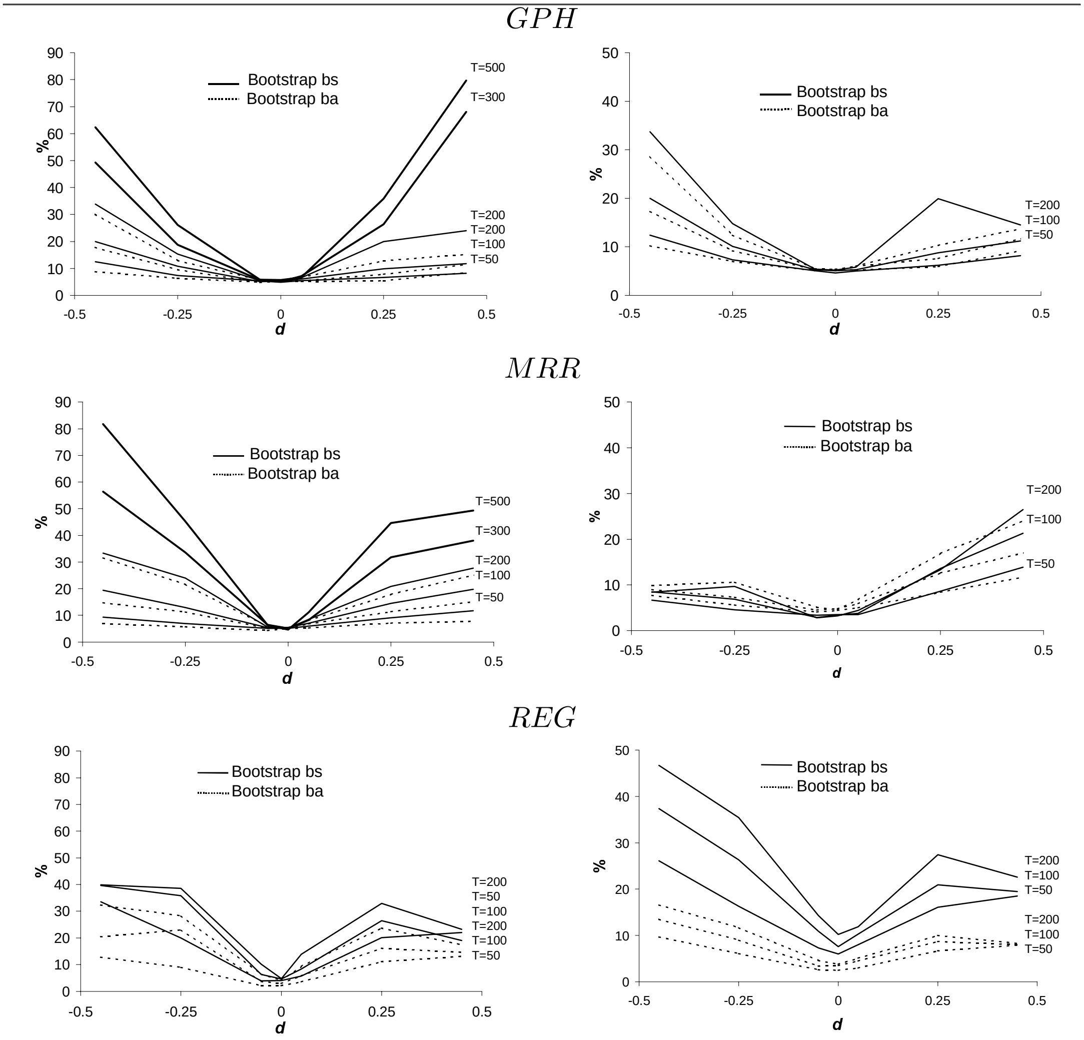
Press, H., Teukolisky, S., Vetterling, W. and B. Flannery, 1992, Numerical Recipes in Fortran: The Art of Scientific Computing, second edition (Cambridge University Press, New York, USA).

Table 1: Rejection percentage of the fractional integration tests when the data follow an AR(1) process.

		ϕ								
		$T = 50$			$T = 100$			$T = 200$		
		0.0	0.5	0.9	0.0	0.5	0.9	0.0	0.5	0.9
					Normal Errors					
G	As	5.3	8.0	61.0	5.2	6.7	70.4	5.9	6.1	71.1
	b_S	4.9	4.1	4.2	5.2	4.8	3.3	5.5	5.5	3.0
	b_A	4.6	3.6	4.0	5.0	4.5	3.2	5.5	5.2	3.0
M	As	6.9	2.3	0.7	6.7	2.5	1.0	6.5	3.8	1.6
	b_S	5.1	4.1	3.5	5.5	5.0	4.0	5.2	5.3	4.4
	b_A	5.1	4.2	3.2	5.3	4.9	3.7	5.3	5.4	4.5
R	As	5.0	5.8	5.2	5.0	6.0	5.2	4.7	5.8	5.0
	b_S	4.0	4.5	4.4	4.6	5.0	4.8	4.7	5.0	4.6
	b_A	2.1	3.9	2.3	2.9	3.6	2.7	4.1	3.6	3.3
					Non-normal Errors					
R	As	6.8	8.5	11.5	5.3	6.3	8.5	5.0	5.9	6.8
	b_S	4.5	4.0	3.6	4.4	4.2	3.9	4.5	4.8	3.9
	b_A	2.1	3.1	1.5	2.8	3.2	2.3	3.6	4.1	3.3
					ARCH Errors, $\beta = 0.9$					
G	As	5.2	7.8	59.5	5.3	6.7	69.8	5.1	5.5	70.5
	b_S	4.6	3.9	3.8	5.1	4.8	3.8	5.2	5.1	5.0
	b_A	5.4	4.7	4.7	5.1	4.7	4.4	5.3	5.3	4.4
M	As	3.1	1.5	0.5	3.1	1.8	0.9	3.2	2.1	1.0
	b_S	3.5	3.8	4.3	3.4	3.4	4.2	3.3	3.2	4.1
	b_A	4.4	4.9	4.8	4.8	5.0	4.9	4.6	5.3	4.8
R	As	27.2	26.6	30.2	26.9	26.2	32.1	30.0	28.6	38.4
	b_S	6.0	5.3	6.7	7.6	6.6	8.5	10.2	7.9	10.3
	b_A	2.5	3.3	3.2	3.5	3.9	3.5	3.8	4.4	4.4

The entries are rejection percentages of the two-sided nominal 5% test. As denotes the asymptotic test and b_S and b_A the bootstraps. G, M and R denote the GPH, MRR and REG tests. Under non-normality, the skewness and kurtosis are 2 and 9 (respectively) for all processes. β denotes the ARCH parameter. The results are based on 10,000 trials.

Figure 1: Rejection percentage of the nominal 5 percent GPH test when the data follow an ARFIMA(0,d,0) process.



Normal Errors ARCH Errors

GPH

MRR

REG

See note to Table 1.