Can portmanteau nonlinearity tests serve as general mis-specification tests?
Evidence from symmetric and asymmetric GARCH models

Chris Brooks\textsuperscript{a,}\*, \Ólan T. Henry\textsuperscript{b}

\textsuperscript{a}ISMA Centre, Department of Economics, The University of Reading, Reading RG6 6BA, UK
\textsuperscript{b}Department of Economics, The University of Melbourne, Parkville, Victoria 3052, Australia

Received 14 July 1999; accepted 16 December 1999

Abstract

A number of recent papers have employed the BDS test as a general test for mis-specification for linear and nonlinear models. We show that for a particular class of conditionally heteroscedastic models, the BDS test is unable to detect a common mis-specification. Our results also demonstrate that specific rather than portmanteau diagnostics are required to detect neglected asymmetry in volatility. However for both classes of tests reasonable power is only obtained using very large sample sizes.

\textsuperscript{\ddagger}This paper was written while the second author was on study leave at the ISMA Centre, University of Reading.
\textsuperscript{*}Corresponding author. Tel.: +44-118-931-6768; fax: +44-118-931-4741.
E-mail address: C.Brooks@ISMACentre.Reading.ac.uk (C. Brooks)

Keywords: GARCH models; Asymmetries; Mis-specification; Engle and Ng tests; BDS test

\textit{JEL classification:} C22; C52

1. Introduction

Testing for nonlinearity in financial and economic time series has become a highly active area of research over the past decade. Interest in nonlinear models has developed in parallel with an expansion in the number of and understanding of the properties of tools for nonlinear data analysis. By far the most widely adopted test for non-linear structure has been the BDS test due to Brock et al. (1987, revised in 1996). Evidence of nonlinearity per se does not provide an insight into the sources of the nonlinearity, or more importantly, the appropriate functional form for the resultant nonlinear model. The BDS test has reasonable power against the GARCH family of models. However it is often
difficult to disentangle the nonlinearity generated by this form of dependence in the second moment from nonlinearities arising as a result of other causes (see Brooks, 1996, for a more detailed discussion of this issue). One solution to this potential problem is to estimate some form of GARCH model for the series \( x_t \), such as

\[
\begin{align*}
x_t &= \mu + u_t, \quad u_t \sim N(0, h_t) \\
\log(h_t) &= \gamma + \alpha_1 u_{t-1}^2 + \beta_1 h_{t-1}
\end{align*}
\]  

(1)

The standardised residual, \( u_t h_t^{-1/2} \), may be subjected to the BDS test and the null hypothesis then becomes one that the specified GARCH model is sufficient to model the nonlinear structure in the data against an unspecified alternative that it is not.

This procedure has been followed a number of times in the literature, the conclusion being that if the BDS test cannot reject the iid null using appropriate critical values derived from simulation, then the model estimated is assumed to be an adequate characterisation of the data. In other words, it has been suggested (for example, Brock et al. (1991, p. 19 or p. 69) that the BDS test can be used as a general test of model mis-specification. This procedure has been followed by various authors (e.g. Hsieh, 1993, for autoregressive volatility models; Hsieh, 1989 or Abhyankar et al., 1995, for GARCH(1,1)).

In this paper, we extend and generalise the recent Monte Carlo studies described above to consider the effectiveness of the BDS test in detecting neglected asymmetries in volatility. Engle and Ng (1993) and Henry (1998) discuss the difficulty in selecting between symmetric and asymmetric GARCH models. Two asymmetric models are considered in this paper. The first model is the so-called GJR (after Glosten et al. (1993)) or threshold GARCH written as

\[
\begin{align*}
h_t &= \alpha_0 + \alpha_1 u_{t-1}^2 + \beta_1 h_{t-1} + \delta \xi_{t-1}^2
\end{align*}
\]  

(2)

where \( \xi_{t-1} = \min(0, u_{t-1}) \). The \( \xi_{t-1} \) term allows us to distinguish between positive and negative innovations. Unlike the GARCH and GJR models the EGARCH or exponential GARCH model of Nelson (1991) is non-linear in the parameters of the conditional variance equation. The EGARCH may be written as

\[
\begin{align*}
\log(h_t) &= \alpha_0 + \beta_1 \log(h_{t-1}) + \delta \left( \frac{u_{t-1}}{\sqrt{h_{t-1}}} \right) + \alpha_1 \left[ \frac{|u_{t-1}|}{\sqrt{h_{t-1}}} \right] - \frac{2}{\pi}
\end{align*}
\]  

(3)

The BDS test is compared to the set of diagnostic tests designed to detect asymmetry in volatility proposed by Engle and Ng (1993). The results suggest that this line of enquiry is highly relevant in the light of many recent studies that have followed the procedure of linear and GARCH filtering and then a BDS application to ‘see what is left’. The performance of the BDS test in detecting neglected asymmetry is highly questionable.

The remainder of the paper develops as follows. The next section provides a brief description of the BDS and Engle–Ng tests. Section 3 outlines the Monte Carlo experiments and reports the results obtained therein. The final section offers some concluding remarks.
2. The BDS and size/sign bias tests

The null hypothesis for the BDS test is that the data are independently and identically distributed (iid), and any departure from iid should lead to rejection of this null in favour of an unspecified alternative. Hence the test can be considered a broad portmanteau test which has been shown to have reasonable power against a variety of nonlinear data generating processes (see Brock et al., 1991 for an extensive Monte Carlo study). The BDS test statistic is calculated as follows. First, the ‘m-histories’ of the data \( x_t^m = (x_t, x_{t+1}, \ldots, x_{t-m+1}) \) are calculated for \( t = 1, 2, \ldots, T - m \) for some integer embedding dimension \( m \geq 2 \). The correlation integral is then computed, which counts the proportion of points in \( m \)-dimensional hyperspace that are within a distance \( \epsilon \) of each other:

\[
C_{m,T}(\epsilon) = \frac{2}{(T-m+1)(T-m)} \sum_{t < s} I_s(x_t^m, x_s^m)
\]

(4)

where \( I_s \) is an indicator function that equals one if \( |x_t^m - x_s^m| < \epsilon \) and zero otherwise, and \( \| \cdot \| \) denotes the sup. norm. BDS show that under the null hypothesis that the observed \( x_t \) are iid, then \( C_{m,T}(\epsilon) \to C_1(\epsilon)^m \) with probability one as the sample size tends to infinity and \( \epsilon \) tends to zero. The BDS test statistic, which has a limiting standard normal distribution, then follows as:

\[
W_{m,T}(\epsilon) = T^{1/2} \frac{C_{m,T}(\epsilon) - C_1(\epsilon)^m}{\sigma_{m,T}(\epsilon)}
\]

(5)

where

\[
\sigma_{m,T}(\epsilon) = 2 \left[ K^m + 2 \left( \sum_{j=1}^{m-1} K^{m-j} C_1(\epsilon)^{2j} \right) - (m-1)^2 C_1(\epsilon)^{2m} - m^2 KC_1(\epsilon)^{2m-2} \right]^{1/2}
\]

(6)

and \( K(\epsilon) \) is estimated by

\[
K(\epsilon) = \frac{6 \sum_{1 < j < k} h_s(x_t^m, x_j^m, x_k^m)}{[(T-m+1)(T-m)(T-m-1)]}
\]

(7)

and \( h_s(i, j, k) = |I_s(i, j)I_s(j, k) + I_s(i, k)I_s(k, j) + I_s(j, i)I_s(i, k)|/3 \).

Two parameters are to be chosen by the user: the value of \( \epsilon \) (the radius of the hypersphere which determines whether two points are ‘close’ or not), and \( m \) (the value of the embedding dimension). Brock et al. (1991) recommend that \( \epsilon \) is set to between half and three halves the standard deviation of the actual data and \( m \) is set in line with the number of observations available (e.g. use only \( m \leq 5 \) for \( T \leq 500 \) etc.). In this study, all combinations of \( \epsilon/\sigma = 1 \) and \( 1.5 \) (where \( \sigma \) is the standard deviation of the sample data), and \( m = 2 \) and \( 5 \) are used.

The common finding among almost all researchers who apply the test is that the iid null is rejected, although this rejection could be the result of either linear or non-linear structure in the data. The BDS test can be used to test against only nonlinear alternatives by filtering the data using an autoregressive or ARMA model of sufficiently high order to ensure that the residuals are serially uncorrelated. The BDS test is used to determine whether the residuals are iid. If rejection occurs, and if the focus is
limited to univariate time series, it must by definition imply that the data generating mechanism has inherent nonlinearities since linear dependence has been filtered out. Brock et al. (1991) show that the BDS test is asymptotically nuisance parameter free (NPF) when applied to the residuals of a linear model, implying that the same set of (standard normal) critical values can be used when the test is applied to residuals as to the raw data. The critical values for the BDS test applied to the standardised residuals of a GARCH(1,1) are taken from Brock et al. (1991) and Brooks and Heravi (1999). To conserve space the results for $e/\sigma = 1$ and $m = 2$ or 5 are reported. The remaining results are available on request from the authors. Engle and Ng (1993) present a test for size and sign bias in conditionally heteroscedastic models. Define $S_{t-1}^-$ as an indicator dummy that takes the value of 1 if $u_{t-1} < 0$ and the value zero otherwise. The test for sign bias is based on the significance of $\phi_1$ in

$$ u_t^2 = \phi_0 + \phi_1 S_{t-1}^- + v_t, \quad (8) $$

Where $v_t$ is a white noise error term. If positive and negative innovations to $u_t$ impact on the conditional variance of $x_t$ differently to the prediction of the model, then $\phi_1$ will be statistically significant. It may also be the case that the source of the bias is caused not only by the sign, but also the magnitude or size of the shock. The negative size bias test is based on the significance of the slope coefficient $\phi_1$ in

$$ u_t^2 = \phi_0 + \phi_1 S_{t-1}^- u_{t-1} + v_t, \quad (9) $$

Likewise, defining $S_{t-1}^+ = 1 - S_{t-1}^-$, then the Engle and Ng (1993) joint test for asymmetry in variance is based on the regression

$$ u_t^2 = \phi_0 + \phi_1 S_{t-1}^- u_{t-1} + \phi_2 S_{t-1}^- u_{t-1} + \phi_3 S_{t-1}^+ u_{t-1} + v_t, \quad (10) $$

where $v_t$ is a white noise disturbance term. Significance of the parameter $\phi_3$ indicates the presence of sign bias. That is, positive and negative realisations of $u_t$ affect future volatility differently to the prediction of the model. Similarly significance of $\phi_2$ or $\phi_3$ would suggest size bias, where not only the sign, but also the magnitude of innovation in $x_t$ is important. A joint test for sign and size bias, based upon the Lagrange Multiplier Principle, may be performed as $T.R^2$ from the estimation of (10).

3. The Monte Carlo simulation

The Monte Carlo study employed in this paper comprises two parts. The first part confirms the simulated size of the BDS and size/sign bias test statistics by generating data from a GARCH(1,1) model. Using the generated data, a GARCH(1,1) is estimated and the BDS and Engle–Ng test statistics are then calculated on the resulting standardised residuals. The second part seeks to determine the power of the tests to detect neglected asymmetries. Data is generated from either an EGARCH or a GJR model. A GARCH(1,1) model is estimated and the standardised residuals again tested using BDS and Engle–Ng. We might anticipate that both sets of tests should strongly reject their respective null hypotheses since the fitted model is mis-specified in that it does not parameterise the asymmetries manifest in the data. All experiments employ the parameter values given in the Monte Carlo study of Engle and Ng (1993), and are based on 5000
Table 1
Size of the BDS and Engle–Ng Tests on the residuals of a correctly speciﬁed GARCH model

<table>
<thead>
<tr>
<th>Nominal %</th>
<th>BDS: ( m = 2 )</th>
<th>BDS: ( m = 5 )</th>
<th>N-sign</th>
<th>N-size</th>
<th>P-size</th>
<th>Joint</th>
</tr>
</thead>
<tbody>
<tr>
<td>Panel A: Sample Size = 500 Observations</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1%</td>
<td>0.34</td>
<td>1.40</td>
<td>0.58</td>
<td>0.40</td>
<td>0.40</td>
<td>0.48</td>
</tr>
<tr>
<td>5%</td>
<td>1.22</td>
<td>2.50</td>
<td>3.98</td>
<td>2.40</td>
<td>2.40</td>
<td>2.18</td>
</tr>
<tr>
<td>10%</td>
<td>2.74</td>
<td>3.80</td>
<td>8.64</td>
<td>6.06</td>
<td>5.68</td>
<td>4.14</td>
</tr>
<tr>
<td>Panel B: Sample Size = 1000 Observations</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1%</td>
<td>0.40</td>
<td>0.34</td>
<td>0.66</td>
<td>0.50</td>
<td>0.28</td>
<td>0.60</td>
</tr>
<tr>
<td>5%</td>
<td>0.72</td>
<td>1.10</td>
<td>3.64</td>
<td>2.84</td>
<td>2.48</td>
<td>2.52</td>
</tr>
<tr>
<td>10%</td>
<td>1.40</td>
<td>2.52</td>
<td>8.64</td>
<td>6.44</td>
<td>6.36</td>
<td>5.16</td>
</tr>
<tr>
<td>Panel C: Sample Size = 3000 Observations</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1%</td>
<td>0.22</td>
<td>0.38</td>
<td>0.82</td>
<td>0.38</td>
<td>0.40</td>
<td>0.54</td>
</tr>
<tr>
<td>5%</td>
<td>0.58</td>
<td>1.02</td>
<td>4.16</td>
<td>2.28</td>
<td>2.26</td>
<td>2.08</td>
</tr>
<tr>
<td>10%</td>
<td>1.14</td>
<td>2.02</td>
<td>9.26</td>
<td>5.46</td>
<td>5.86</td>
<td>4.88</td>
</tr>
</tbody>
</table>

* All cell entries give the percentage of rejections over 5000 Monte Carlo replications. Nominal % denotes the nominal size of the test; N-sign, N-size, P-size and Joint denote the negative sign bias, the negative size bias, the positive bias, and the joint tests of Engle and Ng (1993).

replications with the first 500 observations discarded at each replication in order to mitigate the effect of start-up values.

Table 1 presents the simulated size of the test statistics at sample sizes of 500, 1000 and 3000 observations. The most important feature of the results is that both sets of tests are under-sized, in all cases rejecting their null on fewer occasions than the nominal significance level. However, the degree of under-sizing improves with sample size for the Engle and Ng tests; for the BDS test, if anything the test becomes more conservative as the sample size increases.

Tables 2 and 3 present the results for an application of the tests to the standardised residuals of a GARCH(1,1) model when the data generating process is an EGARCH and a GJR model respectively. The extremely low rejection frequencies suggest that the BDS test struggles to detect the neglected asymmetries in the standardised residuals. The rejection rate is never much greater than the nominal size of the test for samples containing 1000 observations or less. Moreover, the power of the test to detect unparameterised asymmetries rises only very slowly as the sample size is increased. For example, the percentage of correct rejections of the iid null rises from less than 2% (with \( m = 2, \varepsilon/\sigma = 1 \) and a nominal size of 5%) for 500 observations to 14% at sample size 3000. Thus the asymptotic properties of the test seem to be approached extremely slowly in the case where the test is applied to the standardised residuals of a GARCH model.

The size and sign bias tests, on the other hand, have very good power against asymmetries of the GJR type, and reasonable power against those originating from an EGARCH model, and there is a steady increase in power with sample size. For the EGARCH the power of the Engle–Ng joint test increases to over 50% (with a nominal size of 5%) for 3000 observations. For the GJR model, the power of the Engle & Ng joint test increases to virtually 100% when the sample size is increased to 3000.
Table 2
Power of the BDS and Engle–Ng tests on the residuals of a GARCH model when the DGP is EGARCH*

<table>
<thead>
<tr>
<th>Nominal %</th>
<th>BDS: $m = 2$</th>
<th>BDS: $m = 5$</th>
<th>N-sign</th>
<th>N-size</th>
<th>P-size</th>
<th>Joint</th>
</tr>
</thead>
<tbody>
<tr>
<td>Panel A: Sample Size = 500 observations</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1%</td>
<td>2.98</td>
<td>2.98</td>
<td>2.68</td>
<td>6.44</td>
<td>2.82</td>
<td>5.08</td>
</tr>
<tr>
<td>5%</td>
<td>6.52</td>
<td>6.52</td>
<td>10.76</td>
<td>15.82</td>
<td>11.84</td>
<td>13.18</td>
</tr>
<tr>
<td>10%</td>
<td>11.48</td>
<td>11.48</td>
<td>18.62</td>
<td>23.72</td>
<td>20.18</td>
<td>21.10</td>
</tr>
<tr>
<td>Panel B: Sample Size = 1000 observations</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1%</td>
<td>2.96</td>
<td>3.02</td>
<td>5.80</td>
<td>12.80</td>
<td>5.90</td>
<td>10.42</td>
</tr>
<tr>
<td>5%</td>
<td>4.72</td>
<td>3.76</td>
<td>17.48</td>
<td>27.46</td>
<td>20.42</td>
<td>22.76</td>
</tr>
<tr>
<td>10%</td>
<td>7.48</td>
<td>4.62</td>
<td>27.78</td>
<td>37.16</td>
<td>31.78</td>
<td>32.44</td>
</tr>
<tr>
<td>Panel C: Sample Size = 3000 observations</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1%</td>
<td>6.98</td>
<td>5.68</td>
<td>23.34</td>
<td>43.96</td>
<td>27.86</td>
<td>37.18</td>
</tr>
<tr>
<td>5%</td>
<td>8.82</td>
<td>7.70</td>
<td>47.06</td>
<td>64.46</td>
<td>50.62</td>
<td>57.48</td>
</tr>
<tr>
<td>10%</td>
<td>10.68</td>
<td>9.08</td>
<td>59.56</td>
<td>73.66</td>
<td>63.18</td>
<td>68.04</td>
</tr>
</tbody>
</table>

* All cell entries give the percentage of rejections over 5000 Monte Carlo replications. Nominal % denotes the nominal size of the test; N-sign, N-size, P-size, and Joint denote the negative sign bias, the negative size bias, the positive bias, and the joint tests of Engle and Ng (1993).

Table 3
Power of the BDS and Engle–Ng tests on the residuals of a GARCH model when the DGP is GJR*

<table>
<thead>
<tr>
<th>Nominal %</th>
<th>BDS: $m = 2$</th>
<th>BDS: $m = 5$</th>
<th>N-sign</th>
<th>N-size</th>
<th>P-size</th>
<th>Joint</th>
</tr>
</thead>
<tbody>
<tr>
<td>Panel A: Sample Size = 500 observations</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1%</td>
<td>0.76</td>
<td>1.86</td>
<td>7.60</td>
<td>10.64</td>
<td>8.14</td>
<td>6.92</td>
</tr>
<tr>
<td>5%</td>
<td>1.78</td>
<td>2.84</td>
<td>22.36</td>
<td>30.04</td>
<td>30.12</td>
<td>21.52</td>
</tr>
<tr>
<td>10%</td>
<td>3.62</td>
<td>4.20</td>
<td>34.42</td>
<td>43.24</td>
<td>46.22</td>
<td>33.18</td>
</tr>
<tr>
<td>Panel B: Sample Size = 1000 observations</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1%</td>
<td>0.64</td>
<td>0.78</td>
<td>20.38</td>
<td>32.28</td>
<td>29.56</td>
<td>24.44</td>
</tr>
<tr>
<td>5%</td>
<td>1.26</td>
<td>1.58</td>
<td>44.88</td>
<td>60.90</td>
<td>62.14</td>
<td>50.88</td>
</tr>
<tr>
<td>10%</td>
<td>2.44</td>
<td>3.38</td>
<td>58.90</td>
<td>73.74</td>
<td>75.62</td>
<td>64.76</td>
</tr>
<tr>
<td>Panel C: Sample Size = 3000 observations</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1%</td>
<td>10.18</td>
<td>16.16</td>
<td>75.60</td>
<td>92.94</td>
<td>94.14</td>
<td>92.86</td>
</tr>
<tr>
<td>5%</td>
<td>13.68</td>
<td>20.40</td>
<td>91.76</td>
<td>98.74</td>
<td>98.82</td>
<td>98.30</td>
</tr>
<tr>
<td>10%</td>
<td>16.22</td>
<td>24.40</td>
<td>95.40</td>
<td>99.34</td>
<td>99.46</td>
<td>99.26</td>
</tr>
</tbody>
</table>

* See notes to Table 2.

4. Conclusions

This paper sought to test the usefulness of two methods for detecting neglected asymmetries in conditional variance models. Although the sign and size bias tests of Engle and Ng performed reasonably well, the BDS test was virtually unable to detect asymmetries which were present in the data. These results add to recent research of Brooks and Heravi (1999) which showed that the BDS test can sometimes confuse quite different types of nonlinear structure (such as threshold auto-
regressive and GARCH-type models). Overall, the findings of this study should serve as a warning against the sole use of the BDS portmanteau model diagnostics to detect relatively subtle mis-specifications; a valid modelling exercise should also consider tests with power against specific alternative structures that the researcher considers might be present in the data. The results also highlight the importance of the sample size on the performance of the tests and suggest that a minimum of 3000 observations is required for the performance of the specific tests to be viewed as acceptable, while the portmanteau test does not deliver even at this sample size.

Acknowledgements

The authors would like to thank Simon Burke for helpful discussions and Saeed Heravi for assistance in programming the BDS statistic. The usual disclaimer applies to any remaining errors or omissions.

References