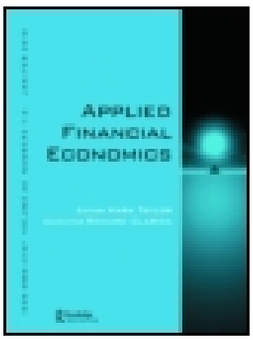

Applied Financial Economics

ISSN: 0960-3107 (Print) 1466-4305 (Online) Journal homepage: https://www.tandfonline.com/loi/rafe20

Testing for structural breaks in GARCH models

Daniel R. Smith

To cite this article: Daniel R. Smith (2008) Testing for structural breaks in GARCH models, Applied Financial Economics, 18:10, 845-862, DOI: 10.1080/09603100701262800

To link to this article: https://doi.org/10.1080/09603100701262800

Published online: 03 Jun 2008.

Applied Financial Economics, 2008, **18**, 845–862

Routledge
Taylor & Francis Group

Testing for structural breaks in GARCH models

Daniel R. Smith

Faculty of Business Administration, Simon Fraser University, 8888 University Drive, Burnaby BC V5A 1S6
E-mail: drsmith@sfu.ca

We study the ability of traditional diagnostic tests and LM and CUSUM structural break tests to detect a range of different types of breaks in GARCH models. We find that Wooldridge's (1990) robust LM tests for autocorrelation and ARCH have no power to detect structural breaks in GARCH models. However, CUSUM- and LM-based structural break tests have excellent size when the data is Gaussian, but the CUSUM tests tend to overreject even in quite large samples when returns have fat tails. However, the LM-based tests have approximately the correct size and exhibit impressive power to detect a range of breaks in the dynamics of conditional volatility. We apply these tests to a range of financial time series using returns starting only in 1990 and find that many GARCH models that pass standard specification tests fail the structural break tests.

I. Introduction

Volatility modelling and the GARCH model in particular have become important tools in many areas of both economics and finance. However despite its widespread application in the literature there is a distinct paucity of structural break tests applied. Structural break tests are an essential diagnostic tool in econometrics. To appeal to standard distributional properties of common econometric estimation methods (i.e., maximum likelihood, least squares, generalized method of moments, etc.) the model must be well specified which in turn requires that the model's parameters to be constant through time. When modelling time-varying volatility we require that the parameters which describe the data generating process of volatility be stable through time. Parameter instability is evidence of model miss-specification and standard econometric theory no longer applies. It is particularly curious that relatively few structural break tests are conducted given the popularity of regime-shifting GARCH models (Gray, 1996; Dueker, 1997; Klaassen, 2002; Haas *et al*., 2004; Pérignon and Smith, 2007).

Given the importance placed on structural break tests when specifying means, it is curious that despite literally thousands of empirical applications of the GARCH models, only a handful of structural break tests have been implemented. Early work by Lamoureux and Lastrapes (1990) demonstrated that breaks in the unconditional level of volatility drove the estimated persistence of volatility towards integrated GARCH model of Engle and Bollerslev (1986) for a sample of 30 individual stocks. They partitioned the sample into a fixed number of equally spaced volatility episodes in which unconditional volatility is fixed. The point estimates of volatility persistence is greatly reduced after allowing for these 'breaks' in volatility. Malik (2003) develops a structural break test based on the iterated cumulated sums of squares algorithm (ICSS) and analyses five exchange rates from January 1990 to September 2000 using this test.

DOI: 10.1080/09603100701262800

He identifies a number of structural breaks in the data. The ICSS algorithm looks for breaks in unconditional volatility and is able to identify these changes in unconditional volatility. Malik (2003) documents that after accounting for these breaks the estimated persistence of volatility shocks is reduced.

An early attempt to construct GARCH structural break tests is Chu (1995) who studies the properties of supremum-type F and Lagrange multiplier (LM) tests. He finds that these tests generally have good power, but are rather sensitive to the assumption of normality. He tests a GARCH model fitted to the S&P500 time series and rejects the null of no structural break. Lin and Yang (1999) consider an empirical distribution function approach to construct structural break tests. Their test is able to detect breaks in moments higher than variance. This feature is interesting but we are typically interested in modelling the dynamics of the conditional mean and volatility, and if we are concerned about fat tails these are modelled directly using maximum likelihood. They also detect structural breaks in a GARCH model of S&P500 returns.

More recently Andreou and Ghysels (2002) analyse the ability of CUSUM and least-squares type structural break tests to detect breaks. They study the performance of these tests using Monte Carlo experiments with normally distributed data in a range of GARCH models and find that the tests perform quite well having approximately the correct size and good power to detect breaks. They also test GARCH models fitted to daily data on four international stock indices and 5 minute returns on the Yen/US Dollar exchange rate. Strong evidence of structural breaks in volatility were found and they note that the break occurred in 1997, corresponding with the Asian currency crisis. Also, Lundbergh and Terasvirta (2002) develop a number of specification tests, including a test for parameter constancy against a threshold-GARCH alternative (which is related to structural break tests), unfortunately there are no empirical applications in their paper.

Testing for breaks when the break date is known is straightforward. However, testing for breaks when the break date is unknown is somewhat more challenging. Following Davies' (1977, 1987) work, Andrews (1993) and Andrews and Ploberger (1994) construct test statistics based on the supremum or suitable averages of the traditional LM test across a range of different break dates. Unfortunately these tests do not have standard distributions but critical values are tabulated in the original paper. Hansen (1997) presents a numerical procedure for computing approximate asymptotic p-values. Furthermore, Hansen (1996) develops a solution to the nuisance parameter problem using a conditional p-value type transformation that is computed by straightforward simulation and has a uniform distribution.

Current practice for specification testing of GARCH models is to test simply for serial correlation in the standardized residuals and/or the squared standardized residuals. One such diagnostic test is the robust regression-based diagnostic tests of Wooldridge (1990). An advantage of Wooldridge's (1990) test is that it is robust in the sense that if the conditional mean and variance are correctly specified, then his test statistic will converge to a chi-squared distribution regardless of the conditional distribution of the underlying data. The tests are designed to detect general miss-specification in either the conditional mean or variance. However, the tests are not directly geared towards detecting structural breaks. There has been little analysis of the ability of general miss-specification tests to uncover breaks. We have some reason to doubt the ability of these tests as Ghysels and Hall (1990) and Hall *et al*. (2003) demonstrate that GMM overidentification tests have no local power to detect structural breaks. With this result in mind we study the ability of Wooldridge's (1990) robust LM test to detect structural breaks in GARCH models using a Monte Carlo experiment. We find that these general diagnostic tests have no power to detect structural breaks in GARCH models. This is particularly troubling since these types of tests are typically the sole diagnostic tools used to assess the specification of GARCH models in practice.

We undertake a detailed Monte Carlo investigation of the size and power of both autocorrelation and ARCH specification tests (i.e. Wooldridge, 1990) and some tests specifically designed to detect structural breaks. In particular we consider both the LM-based structural break tests of Andrews (1993) and Andrews and Ploberger (1994) and the CUSUM break test of Inclán and Tiao (1992). We generate data with both moderately and highly persistent volatility process with both normal and fat-tailed innovations. We find that the tests generally have quite good size properties. Wooldridge's (1990) specification tests for both ARCH and autocorrelation have approximately the correct size. As do the LM-based tests, though we find slight variation in performance between them. In particular we find a slight tendency to over-reject in the *sup*LM and *exp*LM tests but the empirical rejection frequency of the *ave*LM test is remarkably

close to its nominal size. The CUSUM test of Inclán and Tiao (1992, IT) applied to raw returns reject far too frequently to be acceptable even in quite large samples, but when applied to standardized normally distributed residuals the test has quite good size. Interestingly, the performance of the IT tests degrade dramatically when the standardized residuals are leptokurtic. These results complement the results of Andreou and Ghysels (2002) whose simulation experiment used only normally distributed innovations. Even with as many as 2500 observations we reject the null hypothesis four times too often at the 5% level. This is particularly troubling given the compelling evidence that virtually all financial time series have fat tails.

To understand the power of the tests to detect structural breaks we generate artificial data which exhibits a range of structural breaks in both the level of unconditional volatility and the dynamics of volatility. We find that both the LM and IT tests have good power to detect breaks in the unconditional level of volatility, but that only the LM-based tests have the ability to consistently detect breaks in volatility dynamics that do not affect the unconditional level of volatility. The LM-based tests are also able to detect breaks in only the degrees of freedom of a GARCH model with conditionally Student T innovations. We also find that the robust regression-based tests for autocorrelation and ARCH have absolutely no power to detect structural breaks. The common practice of only testing if the standardized residuals are whitened by a GARCH model will fail to detect miss-specification in the form of a structural break. Our results suggest that standard diagnostic tests are no substitute for structural break tests, and that the LM-based tests of Andrews (1993) and Andrews and Ploberger (1994) are preferable to CUSUM-based tests.

We also present an empirical application using 12 different financial time series which are routinely modelled with GARCH models. Although the GARCH models perform very well against Wooldridge's (1990) tests, there is compelling evidence that many of the series contain structural breaks. Even in our relatively short post-1989 sample period we find evidence of structural breaks (the stock index, Canadian and Pound exchange rates and several stocks exhibited evidence of structural breaks).

The remainder of the article proceeds as follows. Section II briefly discusses GARCH models and current diagnostic testing procedures. Section III discusses Woldridge's (1990) robust regression-based diagnostic test for autocorrelation and ARCH effects which are fairly standard model specification tests. The ability of these general specification tests to detect structural breaks is considered in Section IV. Section V discusses LM-based and CUSUM-based structural break tests where the break point is unknown. The results of our Monte Carlo experiment on the specifically designed structural break tests are reported in Section VI. We study the ability of the tests to detect changes in the degrees of freedom in Section VII. In Section VIII we present an empirical application to daily returns on 12 financial time series. The article results are summarized and conclusions are drawn in Section IX.

II. The GARCH Model

We model the conditional mean $\mu_t = E_{t-1}y_t$ and the conditional variance $\sigma_t^2 = E_{t-1}(y_t - \mu_t)^2$ of a series y_t respectively as

$$\mu_t = \mu + \sum_{i=1}^{m} \phi_i y_{t-i} \tag{1}$$

and

$$\sigma_t^2 = \omega + \sum_{i=1}^{p} \alpha_i e_{t-i}^2 + \sum_{i=1}^{q} \beta_i \sigma_{t-i}^2 + \sum_{i=1}^{r} \delta_i e_{t-i}^2 1_{e_{t-i}<0} \tag{2}$$

where $e_t = y_t - \mu_t$ is the unexpected component of y_t and $1_{e_{t-i}<0}$ is an indicator variable taking the value unity when the ith lagged residual is negative. We label this model as AR(m)–GJR(p, q, r). The most common variant is the GARCH(1,1) model (Bollerslev *et al.*, 1992) in which volatility is modelled as:

$$\sigma_t^2 = \omega + \alpha e_{t-1}^2 + \beta \sigma_{t-1}^2 \tag{3}$$

The original ARCH model was motivated by the observation of 'volatility clustering' that is observable in many financial and economic time series. Specifically, conditional volatility is modelled as $E_{t-1}(e_t^2) = \omega + \alpha e_{t-1}^2$. This model assumes that the relationship between lagged surprises and conditional volatility is symmetric. However, it is a well known empirical regularity that in equity market conditional volatility rises more following negative surprises than following positive returns. This observation was initially documented by Black (1976) and Christie (1982) and its most common incarnation in the GARCH literature are by Glosten *et al.* (1993) (the GJR model) and

Nelson (1991) (the EGARCH model). Engle and Ng (1993) analyses a number of different variants of GARCH models that capture this so-called leverage or volatility feedback effect, and found that in their daily stock index series that both the EGARCH and GJR models fit the data the best and perform comparably. The GJR model includes the parameters δ_j to capture the asymmetric response of volatility to past shocks.

In our empirical applications we follow Bollerslev (1987) and use a standardized Student T-distribution with ν^{-1} degrees of freedom, with ν being estimated as a free parameter. We also estimate a restricted model which uses the standard normal distribution which has one less parameter and is a limiting special case of the T model as ν approaches zero. We estimate the parameters by maximum likelihood setting the initial variance σ_0^2 to the unconditional variance. The recursive GARCH filter sequentially constructs the sequence of conditional variance and the unexpected returns and also the log-likelihood. We maximize the log-likelihood numerically using the unconstrained optimization routine fminunc in Matlab. To ensure the model is well defined (i.e., stationary and with positive variance) we employ a penalty function to force the optimization routine to consider only feasible parameter values.

For the conditional volatility to be strictly positive, constraints must be placed on the parameters: e.g. $\omega > 0$. It is common to constrain α_i, $\beta_i > 0$, however this is not necessary for strictly positive volatility forecasts. Consider, for example, the GARCH(2,1) model when if α_1, $\beta_1 > 0$, α_2 can be negative if $-\alpha_2 < \alpha_1\beta_1$ (Hamilton, 1994, p. 666). For covariance stationarity, the GARCH parameters must satisfy

$$\sum_{i=1}^{p} \alpha_i + \sum_{i=1}^{q} \beta_i + 0.5 \sum_{i=1}^{r} \delta_i < 1 \tag{4}$$

If this constraint holds, then the unconditional variance is given by

$$\sigma^2 = \frac{\omega}{1 - \sum_{i=1}^{p} \alpha_i - \sum_{i=1}^{q} \beta_i - 0.5 \sum_{i=1}^{r} \delta_i}$$

We estimate and report the unconditional variance σ^2 rather than the GARCH intercept ω so our results are easier to interpret.

Andreou and Ghysels (2002) point out that a major complication that arises in applying structural break tests is that many of the tests require distributional assumptions which GARCH models may not exhibit. Most structural break tests require uniform or ϕ-mixing while Carrasco and Chen (2001) show that most univariate GARCH models are β-mixing. The structural break test of Andrews (1993) requires strong or α-mixing (Assumption 1 on p. 830) which is implied by β-mixing.

III. Diagnostic Tests

Robust estimation requires a minimum that the conditional mean and variance be correctly specified. We test the adequacy of the conditional mean and variance using Wooldridge's (1990) robust regression-based specification test. The test is robust to a range of potential miss-specifications and can be calculated using a sequence of OLS regressions. It is currently used in the literature to test the specification of both univariate and multivariate GARCH models (see, e.g. Brenner *et al.*, 1996 who test univariate GARCH models and Pérignon and Smith, 2007 for tests of multivariate GARCH models).

This test is constructed using a miss-specification indicator variable λ_t. Below we present specific cases to test the specification of the conditional mean and the conditional variance. Let θ_μ and θ_σ denote the vectors of all parameters that describe the conditional mean and variance dynamics respectively. Wooldridge (1990) develops a unified approach to calculate robust specification test statistics using linear least-squares regressions. The tests require $\sqrt{T}$-consistent estimators of θ_μ and θ_σ which we denote by $\hat{\theta}_{\mu,T}$ and $\hat{\theta}_{\sigma,T}$.

The test for the specification of the conditional mean μ_t uses the first order condition from a nonlinear least-squares estimation problem, such that

$$T^{-1} \sum_{t=1}^{T} \frac{e_t}{\sigma_t^2} = 0$$

where the inverse of the variance is used as a weighting function on the residuals (though one could use alternative weights). We test for pth order serial correlation which defines the miss-specification indicator variable as $\lambda_t = (e_{t-1}, \ldots, e_{t-p})^{\mathsf{T}}$[1] using the test statistic

$$T^{-1} \sum_{t=1}^{T} \frac{e_t}{\sigma_t^2} \lambda_t$$

[1] The miss-specification indicator variable will generally be a function of the mean parameters θ_μ and possibly a nuisance parameter ξ: $\lambda_t(\theta_\mu, \xi)$. The asymptotic distribution of the test statistics requires a $\sqrt{T}$-consistent estimator of this nuisance parameter $\hat{\xi}_T$.

which is close to zero when the conditional mean is correctly specified. The robust test statistic for the specification of the conditional mean is computed as follows. The conditional mean specification test requires a $\sqrt{T}$-consistent estimator of θ_μ which is denoted by $\hat{\theta}_{\mu,T}$. The first step is to compute the residuals $\hat{e}_t$, the gradient $\nabla_{\theta_\mu}\mu_t(\hat{\theta}_{\mu,T}) = (\partial/\partial\theta_\mu)\mu_t(\hat{\theta}_{\mu,T})$ and the indicator function $\lambda_t(\hat{\theta}_{\mu,T})$. We then form the standardized variables $\tilde{e}_t = \hat{\sigma}_t^{-1}\hat{e}_t$ using a consistent estimator of the variance $\hat{\sigma}_t^2$, $\nabla_{\theta_\mu}\tilde{\mu}_t = \hat{\sigma}_t^{-1}\nabla_{\theta_\mu}\mu_t$ and $\tilde{\lambda}_t = \hat{\sigma}_t^{-1}\lambda_t$. We next regress $\tilde{\lambda}_t$ on the scaled gradient vector $\nabla_{\theta_\mu}\tilde{\mu}_t$ and store the residuals as $\ddot{\lambda}_t$ The robust chi-squared test statistic is then computed as TR_u^2 where R_u^2 is the uncentered R^2 from the least-squares regression of a constant 1 on $\tilde{\phi}_t\ddot{\lambda}_t$.

The test for the specification of the conditional volatility σ_t^2 notes that when the conditional volatility is correctly specified, the demeaned standardized squared residual will be zero and not forecastable using any lagged information. If the variance is correctly specified, the following

$$T^{-1}\sum_{t=1}^{T}\frac{e_t^2-\sigma_t^2}{\sigma_t^2}\lambda_t$$

should be close to zero. We consider the test for pth order ARCH effects which considers lagged squared residuals as the information variable. In particular, we define the indicator variable $\lambda_t = (e_{t-1}^2,\ldots,e_{t-p}^2)^{\mathsf{T}}$ for use in the ARCH tests. We again require $\sqrt{T}$-consistent estimators of both θ_μ and θ_σ which we denote by $\hat{\theta}_{\mu,T}$ and $\hat{\theta}_{\sigma,T}$. We compute the residuals $\hat{e}_t$ the gradient $\nabla_{\theta_\sigma}\sigma_t^2(\hat{\theta}_{\sigma,T}) = (\partial/\partial\theta_\sigma)\,\sigma_t^2(\hat{\theta}_{\sigma,T})$ and the indicator function $\lambda_t(\hat{\theta}_{\sigma,T})$. We next define $\tilde{\phi}_t \equiv (\hat{e}_t^2-\hat{\sigma}_t^2)/\hat{\sigma}_t^2 = \hat{e}_t^2/\hat{\sigma}_t^2 - 1, \nabla_{\theta_\sigma}\hat{\sigma}_t^2(\hat{\theta}_{\sigma,T}) = \nabla_{\theta_\sigma}\hat{\sigma}_t^2/\hat{\sigma}_t^2(\hat{\theta}_{\sigma,T})$ and $\tilde{\lambda}_t \equiv \hat{\lambda}_t/\hat{\sigma}_t^2$, and regress $\tilde{\lambda}_t$ on $\nabla_{\theta_\sigma}\tilde{\sigma}_t^2$ and store the residuals as $\ddot{\lambda}_t$. The robust chi-squared test statistic is then computed as TR_u^2 where R_u^2 is the uncentered R^2 from the least-squares regression of a constant 1 on $\tilde{\phi}_t\ddot{\lambda}_t$.

IV. Can Standard Diagnostic Tests Detect Structural Breaks?

The standard approach to specification testing for GARCH models is to assess the extent to which the model whitens the standardized residuals using tests such as in Wooldridge (1990). However, there is good reason to be skeptical about these tests' ability to detect structural breaks. In a related context Ghysels and Hall (1990) and Hall *et al.* (2003) demonstrate that GMM overidentification tests have no local power to detect structural breaks. We therefore undertake a Monte Carlo simulation study to assess the ability of Wooldridge's (1990) tests diagnostic tests to detect structural breaks. We study both the power and size of the tests using artificial data simulated using GARCH models with and without structural breaks.

We first study the size of Wooldridge's (1990) autocorrelation and ARCH tests. We simulate 1000 different time-series of returns with a range of different sample sizes $T=\{500,\ 1000,\ 2500\}$. We simulate data with both Gaussian disturbances and from a standardized Student T distribution with 7.5 degrees of freedom to assess the effect of fat tails on the properties of the tests. We choose the parameterizations to ensure that the unconditional variances exist (see Bollerslev, 1987; He and Terasvirta, 1999). We simulate data using GARCH process with different degrees of persistence:

- Model A: A GARCH process with moderate persistence

$$\sigma_t^2 = 0.4 + 0.3e_{t-1}^2 + 0.3\sigma_{t-1}^2$$

- Model B: A GARCH process with high persistence

$$\sigma_t^2 = 0.1 + 0.1e_{t-1}^2 + 0.8\sigma_{t-1}^2$$

For each artificial data series we estimate a GARCH model under the maintained null hypothesis of no structural break and then subject it to Wooldridge's (1990) robust regression-based specification tests for both autocorrelation and ARCH effects in the standardized residuals at lags 1–5.[2]

The empirical distribution of these test statistics are reported in Table 1. For each parametrization and sample size we report the empirical rejection frequency for nominal sizes 0.10, 0.05 and 0.01. It is clear from this table that Wooldridge's (1990) robust regression-based specification tests for autocorrelation and ARCH effects have excellent size properties even in quite modest sample sizes.

To investigate the power of these tests to identify breaks we simulate data with the same three sample sizes, and introduce a structural break after $\pi=\{0.3,\ 0.5\}$ of the observations. We report

[2] The results are similar when including only one lag.

Table 1. The size of Wooldridge's (1990) tests

		Normal		T(7.5)	
Model	π	WAR	WARCH	WAR	WARCH
Panel A: $T=500$					
A: $(0.5, 1, 0.3, 0.3)^{\mathsf{T}}$	0.10	0.095	0.058	0.097	0.098
	0.05	0.048	0.027	0.040	0.046
	0.01	0.008	0.003	0.005	0.009
B: $(0.05, 1, 0.1, 0.8)^{\mathsf{T}}$	0.10	0.094	0.097	0.085	0.107
	0.05	0.041	0.048	0.041	0.054
	0.01	0.012	0.007	0.005	0.008
Panel B: $T=1000$					
A: $(0.05, 1, 0.3, 0.3)^{\mathsf{T}}$	0.10	0.102	0.095	0.085	0.086
	0.05	0.050	0.031	0.032	0.036
	0.01	0.005	0.006	0.006	0.007
B: $(0.05, 1, 0.1, 0.8)^{\mathsf{T}}$	0.10	0.097	0.089	0.083	0.109
	0.05	0.048	0.046	0.035	0.061
	0.01	0.003	0.008	0.005	0.010
Panel C: $T=2500$					
A: $(0.05, 1, 0.3, 0.3)^{\mathsf{T}}$	0.10	0.097	0.098	0.089	0.089
	0.05	0.040	0.046	0.038	0.055
	0.01	0.005	0.009	0.005	0.011
B: $(0.05, 1, 0.1, 0.8)^{\mathsf{T}}$	0.10	0.085	0.107	0.090	0.118
	0.05	0.041	0.054	0.037	0.060
	0.01	0.005	0.008	0.010	0.009

Notes: We report the empirical size of Wooldridge's robust regression-based Autocorrelation (WAR) and ARCH test (WARCH) for standardized residuals in artificial data that do not contain a structural break. The empirical size is the fraction of the 1000 generated samples (of varying lengths, $T=\{500, 1000, 2500\}$) that are greater than the reported critical value at each p-value. The parameter vector describing the data is $\theta=(\mu,\sigma^2,\alpha,\beta)^{\mathsf{T}}$. We simulate data from two different parameterizations for both normally and conditionally Student T (with 7.5 degrees of freedom) distributed data in which the GARCH process are both moderately and highly persistent.

the fraction of the 1000 artificial data series in which the AR and ARCH specification tests are significant. All simulation experiments are conducted with both Gaussian and standardized $T(7.5)$ innovations. All series initially have unit unconditional variance (this may or may not increase after the induced break). We consider a range of scenarios:

- The first two series test for power to detect shifts in the unconditional variance for both levels of persistence are considered in Model A and B above. We simulate data such that after the structural break occurs unconditional volatility jumps from its initial level of one by 40% (in scenario 1) and 80% (in scenario 2).
- A third power test examines the ability to detect increases in persistence. The unconditional variance (set at 1) does not change and the coefficient on the lagged squared residual is set to $\alpha=0.1$. β is initially set at $\beta=0.5$ but increases to $\beta=0.6$ (in scenario 1), and to $\beta=0.8$ (in scenario 2).
- The fourth power test leaves both the unconditional variance (set to $\sigma^2=1$) and persistence unchanged (fixed at $\alpha+\beta=0.9$) after the break, but allows the mix of α and β to change. We initially simulate the data from the high persistence GARCH model above, but after the break changes to (in scenario 1):

$$\sigma_t^2 = 0.1 + 0.3e_{t-1}^2 + 0.6\sigma_{t-1}^2$$

and a more dramatic change after the break to (in scenario 2):

$$\sigma_t^2 = 0.1 + 0.45e_{t-1}^2 + 0.45\sigma_{t-1}^2$$

The power experiments are reported in Table 2. To conserve space we only report the results for the 500 and 2500 observation samples. For each model we report two sets of results: the first row corresponds to a break arbitrarily introduced after 30% of the sample, while the second row is when the break is at the mid-point of the data. The most striking result to take from the tables are the negligible power of traditional diagnostic tests to

Table 2. The power of the Wooldridge (1990) tests to detect structural breaks

		Normal		T(7.5)	
Model	π	WAR	WARCH	WAR	WARCH
Panel A: $T=500$					
Pre: $(0.05, 1, 0.3, 0.3)^{\mathsf{T}}$	0.3	0.058	0.030	0.055	0.018
Post: $(0.05, 1.4, 0.3, 0.3)^{\mathsf{T}}$	0.5	0.062	0.033	0.052	0.033
Pre: $(0.05, 1, 0.3, 0.3)^{\mathsf{T}}$	0.3	0.057	0.049	0.056	0.030
Post: $(0.05, 1.8, 0.3, 0.3)^{\mathsf{T}}$	0.5	0.056	0.065	0.047	0.032
Pre: $(0.05, 1, 0.1, 0.8)^{\mathsf{T}}$	0.3	0.060	0.041	0.059	0.041
Post: $(0.05, 1.4, 0.1, 0.8)^{\mathsf{T}}$	0.5	0.059	0.046	0.061	0.045
Pre: $(0.05, 1, 0.1, 0.8)^{\mathsf{T}}$	0.3	0.055	0.042	0.049	0.038
Post: $(0.05, 1.8, 0.1, 0.8)^{\mathsf{T}}$	0.5	0.058	0.051	0.054	0.051
Pre: $(0.05, 1, 0.1, 0.5)^{\mathsf{T}}$	0.3	0.062	0.040	0.062	0.041
Post: $(0.05, 1, 0.1, 0.6)^{\mathsf{T}}$	0.5	0.061	0.043	0.046	0.046
Pre: $(0.05, 1, 0.1, 0.5)^{\mathsf{T}}$	0.3	0.055	0.035	0.057	0.033
Post: $(0.05, 1, 0.1, 0.8)^{\mathsf{T}}$	0.5	0.057	0.036	0.050	0.042
Pre: $(0.05, 1, 0.1, 0.5)^{\mathsf{T}}$	0.3	0.060	0.032	0.051	0.029
Post: $(0.05, 1, 0.2, 0.4)^{\mathsf{T}}$	0.5	0.060	0.032	0.047	0.027
Pre: $(0.05, 1, 0.1, 0.5)^{\mathsf{T}}$	0.3	0.058	0.031	0.054	0.022
Post: $(0.05, 1, 0.3, 0.3)^{\mathsf{T}}$	0.5	0.060	0.025	0.050	0.027
Panel B: $T=2500$					
Pre: $(0.05, 1, 0.3, 0.3)^{\mathsf{T}}$	0.3	0.044	0.061	0.052	0.051
Post: $(0.05, 1.4, 0.3, 0.3)^{\mathsf{T}}$	0.5	0.046	0.059	0.047	0.042
Pre: $(0.05, 1, 0.3, 0.3)^{\mathsf{T}}$	0.3	0.044	0.096	0.053	0.073
Post: $(0.05, 1.8, 0.3, 0.3)^{\mathsf{T}}$	0.5	0.048	0.140	0.041	0.085
Pre: $(0.05, 1, 0.1, 0.8)^{\mathsf{T}}$	0.3	0.046	0.059	0.050	0.046
Post: $(0.05, 1.4, 0.1, 0.8)^{\mathsf{T}}$	0.5	0.046	0.069	0.060	0.053
Pre: $(0.05, 1, 0.1, 0.8)^{\mathsf{T}}$	0.3	0.046	0.081	0.043	0.052
Post: $(0.05, 1.8, 0.1, 0.8)^{\mathsf{T}}$	0.5	0.045	0.096	0.053	0.056
Pre: $(0.05, 1, 0.1, 0.5)^{\mathsf{T}}$	0.3	0.046	0.047	0.054	0.034
Post: $(0.05, 1, 0.1, 0.6)^{\mathsf{T}}$	0.5	0.046	0.047	0.054	0.031
Pre: $(0.05, 1, 0.1, 0.5)^{\mathsf{T}}$	0.3	0.046	0.065	0.047	0.061
Post: $(0.05, 1, 0.1, 0.8)^{\mathsf{T}}$	0.5	0.046	0.062	0.046	0.054
Pre: $(0.05, 1, 0.1, 0.5)^{\mathsf{T}}$	0.3	0.042	0.051	0.057	0.048
Post: $(0.05, 1, 0.2, 0.4)^{\mathsf{T}}$	0.5	0.042	0.057	0.042	0.060
Pre: $(0.05, 1, 0.1, 0.5)^{\mathsf{T}}$	0.3	0.044	0.054	0.053	0.051
Post: $(0.05, 1, 0.3, 0.3)^{\mathsf{T}}$	0.5	0.041	0.054	0.049	0.052

Notes: We report the empirical power of Wooldridge's (1990) robust regression-based autocorrelation (WAR) and ARCH (WARCH) specification tests to detect structural breaks of various forms in a range of samples. We generate 1000 artificial samples with 500 or 2500 observations each. We introduce a structural break after the fraction of observations $\pi = \{0.3, 0.5\}$. The parameter vector describing the data generating process is $\theta = (\mu, \sigma^2, \alpha, \beta)^{\mathsf{T}}$ and the value before the break is denoted by 'Pre', and after the break by 'Post'. For each combination of Pre and Post parameters we report two rows of results, one for each π. We report the rejection frequencies using the 5% critical value with data generated using both normally distributed and conditionally student T distributed (with 7.5 degrees of freedom) innovations.

detect structural breaks. When returns are Gaussian both the robust autocorrelation and ARCH tests of Wooldridge (1990) reject at or below the nominal 5% rejection frequency even where the breaks are involved. The only scenario where standard diagnostic tests have even marginal power is in large samples (i.e., 2500 observations) with low persistence when there is an 80% increase in unconditional volatility at precisely the mid-point of the sample. In this idealized example with Gaussian returns we reject in less than 15% of samples at the nominal 5% level. These tests still lack power to detect breaks when volatility persistence is at a more realistic level (i.e., $\alpha + \beta = 0.9$). The lack of power is even worse when the data exhibits fat tail where even with 2 500 observations the tests still reject about as frequently as their nominal size.

Although these tests are not specifically designed to detect structural breaks they are generally the only specification tests used in empirical work. Our finding that these tests have no power to detect structural breaks suggests the need to consider tests specifically designed to detect structural breaks, which we do in the following section.

V. Structural Break Tests

In this section we consider two types of structural break tests that are useful for testing for breaks in GARCH models: Lagrange Multiplier-based tests and CUSUM-type tests. An important feature of both tests is that they do not require that the econometrician know the break date.

Lagrange-multiplier tests

Consider an econometric model parameterized by θ_t which is fit to time series data y_t for $t = 1, \ldots, T$. We are particularly interested in testing for a one time change in the parameters. The change point occurs at observation πT with $\pi \in (0, 1)$ being the proportion of the sample occurring before the break. The model parameters before the break are θ_1 and after the break θ_2. In particular the parameter vector describing the data generating process at time t is given by

$$\theta_t = \begin{cases} \theta_1(\pi) & \text{for } t = 1, \ldots, [\pi T] \\ \theta_2(\pi) & \text{for } t = [\pi T] + 1, \ldots, T \end{cases} \tag{5}$$

where $[\pi T]$ rounds down to the nearest integer. The null hypothesis of no structural break is that:

$$H_0 : \theta_t = \theta_0 \quad \forall t$$

When the break point π is known, testing for structural breaks is a standard and relatively trivial exercise. The Chow test is a very well known textbook structural break test and is constructed as an F-test for the null hypothesis that $\theta_1 = \theta_2$ for known break point π. However, when the break point is unknown, standard tests no longer have standard distributions. To that end, Andrews (1993) and Andrews and Ploberger (1994) develop tests for structural breaks which are valid for unknown break points, and tabulate asymptotic critical values for these test statistics.

The Lagrange-multiplier (LM hereafter) test is often used in constructing specification tests since it only requires parameter estimates obtained under the null hypothesis. This is important because it is typically much easier to estimate the null model, and specification tests are applied to an already estimated model which is correct under the null to examine the alternative of incorrect model specification. Under the null of no structural break there is one common parameter vector which holds for the whole sample and this is estimated using maximum likelihood. If one knew the break point was π the LM test for the structural break alternative would be constructed as

$$LM_T(\pi) = \frac{T}{\pi(1-\pi)} \bar{g}_{1T}(\tilde{\theta}, \pi)^{\mathrm{T}} S_T^{-1} \times D_T (D_T^{\mathrm{T}} S_T^{-1} D_T)^{-1} D_T^{\mathrm{T}} S_T^{-1} \bar{g}_{1T}(\tilde{\theta}, \pi)$$

Here $g(y_t; \tilde{\theta}) = \partial \log f(y_t; \tilde{\theta}) / \partial \tilde{\theta}$ is the score, or partial derivative of the log-density with respect to the parameter vector, $\bar{g}_{1T}(\tilde{\theta}, \pi) = (1/T) \sum_{t=1}^{\pi T} g(y_t; \tilde{\theta})$, $S_T = (1/T) \sum_{t=1}^{T} (g(y_t; \tilde{\theta}) - \bar{g}_T(\tilde{\theta}))(g(y_t; \tilde{\theta}) - \bar{g}_T(\tilde{\theta}))^{\mathrm{T}}$ with $\bar{g}_T(\tilde{\theta}, \pi) = (1/T) \sum_{t=1}^{T} g(y_t; \tilde{\theta})$, and $D_T = (1/T) \sum_{t=1}^{T} (\partial g(y_t; \tilde{\theta}) / \partial \tilde{\theta}^{\mathrm{T}})$. We compute the derivatives numerically. This LM statistic is asymptotically distributed as a chi-squared random variable with degrees of freedom equal to the dimensionality of the parameter vector.

Maximum likelihood sets the score equal to zero, so D_T is the Hessian-based estimate of the information matrix, and S_T is the outer product-based estimate. The sample mean of the moment condition is set to zero by construction, so the two partial means must add to zero. If the null hypothesis is true, then each sub-sample score will be close to zero since the model is correctly specified. However, if there is a structural break in the model, the estimate which sets the whole sample score to zero will do a poor job in each sub-sample. In other words, $\bar{g}_{1T}$ need not equal zero and in fact $\bar{g}_{1T}(\tilde{\theta}, \pi) = -\bar{g}_{2T}(\tilde{\theta}, \pi)$ which can take literally any value. This last result illustrates why the LM test only uses the first partial mean.

Unfortunately the break point is typically unknown and standard distributional theory no-longer applies since the information matrix is singular under the null hypothesis of no structural break – the log-likelihood takes the same value for all possible break points when $\theta_1 = \theta_2$. Andrews (1993) considers the *sup*LM test which is inspired by the work of Davies (1977, 1987):

$$\sup_{\pi \in \Pi} LM_T(\pi) \tag{6}$$

The null hypothesis is rejected for large values of this test statistic. Andrews (1993) notes that the *sup*LM statistic is poorly behaved when $\Pi = [0, 1]$ since in the boundary region it diverges as the sample size increases: $\lim_{T \to \infty} \sup_{\pi \in [0, 1]} LM(\pi) = \infty$. However, when π is bounded away from zero and one, the test statistic has a well defined asymptotic distribution which is dependent on the number of parameters and the proportion of the sample excluded by this bounding exercise. Critical values of the test statistic are tabulated in Andrews (1993) which depend on

a parameter π_0 which is the proportion of the observations dropped at the beginning and end of the sample, i.e. $\Pi = [\pi_0, 1-\pi_0]$.

One limitation of the *sup*LM test is that it depends on only one sample point. The power of the tests can be improved by considering test statistics over a range of different possible break points. Andrews and Ploberger (1994) develop optimal tests which are (weighted) averages of the individual $LM_T(\pi)$ tests over different possible break points π. Assuming that π is uniformly distributed over $[\pi_1, \pi_2]$ ($\pi_1 \leq \pi_2$ and bounded away from zero and one), the exponential Lagrange multiplier statistic (*exp*LM hereafter) is given by

$$\begin{aligned} exp\mathrm{LM}_T = \log\Bigg(\frac{1}{T(1-2\pi_0)}\Bigg[\sum_{t=[\pi_0 T]+1}^{T-[\pi_0 T]-1} \exp(LM_T(t/T)/2) \\ + ([\pi_0 T]+1-\pi_0 T)\{\exp(LM_T([\pi_0 T]/T)/2) \\ + \exp(LM_T((T-[\pi_0 T])/T)/2)\}\Bigg]\Bigg) \end{aligned} \tag{7}$$

and the average Lagrange multiplier statistic (*ave*LM hereafter)

$$\begin{aligned} ave\mathrm{LM}_T = \frac{1}{T(1-2\pi_0)}\Bigg[\sum_{t=[\pi_0 T]+1}^{T-[\pi_0 T]-1} LM_T(t/T) \\ + ([\pi_0 T]+1-\pi_0 T)\big\{LM_T([\pi_0 T]/T) \\ + LM_T((T-[\pi_0 T])/T)\big\}\Bigg]. \end{aligned} \tag{8}$$

The *ave*LM and *exp*LM consider the value of the structural break test at all possible break points and has more power than the *sup*LM test. As in the *sup*LM test, the distribution of the test depends on the amount of the sample dropped and on the number of parameters. Although the test statistics do not have a closed form distribution, critical values for the *ave*LM and *exp*LM tests are tabulated in Andrews and Ploberger (1994). Also, Hansen (1997) presents a numerical procedure for computing approximate asymptotic *p*-values.

These tests can be used to test for a break jointly in any of the parameters, or can be restricted to testing for a break in only a subset of the parameters. We consider six different partitions of the structural break tests when reporting our empirical results but to save space we do not employ them in all our Monte Carlo experiments: (a) all parameters jointly (ALL); (b) only the unconditional mean (UC mean); (c) the autoregressive parameters, if any (AR); (d) only the unconditional volatility (USC volatility); (e) all GARCH and GJR parameters jointly (ARCH) and (f) Only the degrees of freedom parameter (DFL). It is worth pointing out that these LM-based tests are only capable of identifying the existence of at least one break, although the tests could be used to sequentially test for multiple breaks. In particular, after identifying a break we could estimate an augmented model which allows certain parameter to be different before and after the break date identified by the *sup*LM test. This augmented model could then be tested for a structural break. This extension is beyond the scope of the current article.

CUSUM tests

Andreou and Ghysels (2002) consider both the CUSUM tests of Inclán and Tiao (1992, IT hereafter) Kokoszka and Leipus (1998, 2000, KL hereafter) and least-squares tests which test for a break in the unconditional level of volatility. We focus on the IT test because although the KL test exhibited slightly better power than the IT test, there were small size distortions in the KL statistic. In particular, Andreou and Ghysels (2002) find that even with 3000 observations when volatility is persistent ($\alpha+\beta=0.9$) the KL test rejects the null hypothesis in around 20% of samples when there is no break while the IT test on standardized residuals rejects only 5.2% of the time.[3]

We consider the IT test of Inclán and Tiao (1992)

$$IT = \sqrt{T/2}\max_k \left| \sum_{j=1}^{k} X_j / \sum_{j=1}^{T} X_j - \frac{k}{T} \right|$$

where $X_t = e_t^2$ or $X_t = e_t^2/\sigma_t^2$, and under the null hypothesis of no structural break *IT* has the Kolmogorov–Smirnov distribution.[4] An advantage of the IT test is that it is capable of detecting multiple breaks whereas the LM tests cannot. A disadvantage of IT is that it is only capable of detecting breaks in the unconditional level of volatility but the LM-based tests are capable of detecting changes in the persistence of volatility and even changes in α and β such that volatility persistence $\alpha+\beta$ remains unchanged.

[3] Note that the IT test on raw returns also rejected too frequently.
[4] The critical values 1.22, 1.36 and 1.63 are at 10, 5 and 1%, respectively.

Table 3. The size of the structural break tests

Model	p	*sup*LM	*ave*LM	*exp*LM	IT	IT(z)	*sup*LM	*ave*LM	*exp*LM	IT	IT(z)
Panel A: $T=500$, Normal											
A: $(0.5, 1, 0.3, 0.3)^{\mathsf{T}}$	0.10	0.135	0.089	0.133	0.609	0.066	0.204	0.210	0.215	0.831	0.326
	0.05	0.092	0.037	0.076	0.490	0.025	0.135	0.106	0.116	0.758	0.217
	0.01	0.041	0.005	0.033	0.317	0.000	0.056	0.018	0.040	0.592	0.085
B: $(0.05, 1, 0.1, 0.8)^{\mathsf{T}}$	0.10	0.171	0.103	0.156	0.638	0.054	0.222	0.214	0.228	0.869	0.316
	0.05	0.118	0.051	0.104	0.521	0.020	0.159	0.108	0.132	0.786	0.219
	0.01	0.057	0.006	0.041	0.329	0.000	0.070	0.022	0.060	0.622	0.096
Panel B: $T=1000$, Normal											
A: $(0.05, 1, 0.3, 0.3)^{\mathsf{T}}$	0.10	0.146	0.104	0.131	0.666	0.079	0.227	0.215	0.231	0.859	0.335
	0.05	0.092	0.044	0.072	0.539	0.038	0.142	0.109	0.128	0.764	0.218
	0.01	0.033	0.006	0.025	0.366	0.008	0.054	0.022	0.041	0.596	0.087
B: $(0.05, 1, 0.1, 0.8)^{\mathsf{T}}$	0.10	0.149	0.100	0.129	0.708	0.071	0.243	0.213	0.227	0.877	0.332
	0.05	0.102	0.046	0.087	0.572	0.028	0.157	0.102	0.146	0.807	0.215
	0.01	0.054	0.007	0.042	0.397	0.002	0.080	0.028	0.055	0.653	0.092
Panel C: $T=2500$, Normal											
A: $(0.05, 1, 0.3, 0.3)^{\mathsf{T}}$	0.10	0.124	0.109	0.121	0.732	0.077	0.204	0.237	0.218	0.912	0.330
	0.05	0.080	0.057	0.067	0.613	0.041	0.122	0.122	0.110	0.851	0.223
	0.01	0.029	0.009	0.022	0.421	0.008	0.039	0.021	0.025	0.703	0.089
B: $(0.05, 1, 0.1, 0.8)^{\mathsf{T}}$	0.10	0.136	0.089	0.130	0.776	0.075	0.212	0.196	0.191	0.934	0.329
	0.05	0.099	0.053	0.084	0.655	0.034	0.133	0.099	0.117	0.866	0.223
	0.01	0.042	0.011	0.035	0.472	0.004	0.055	0.020	0.042	0.744	0.080

Notes: We report the empirical size of the *sup*LM, *ave*LM and *exp*LM tests (using the 10, 5 and 1%, critical values found in Andrews (1993) and Andrews and Ploberger (1994)) and the IT tests using raw (IT) and standardized innovations (IT(z)) in artificial data that do not contain a structural break. The empirical size is the fraction of the 1000 generated samples (of varying lengths, $T=\{500, 1000, 2500\}$) that are greater than the reported critical value at each p-value. The parameter vector describing the data is $\theta=(\mu,\sigma^2,\alpha,\beta)^{\mathsf{T}}$. We simulate data from two different parameterizations for both normally and conditionally Student T (with 7.5 degrees of freedom) distributed data in which the GARCH process are both moderately and highly persistent.

VI. The Size and Power of the Structural Break Tests

Using a Monte Carlo experiment we previously found that standard specification tests lack power to detect structural breaks. We now use the same simulation framework to study the size and power of structural break tests to detect breaks. In particular we use the same artificial return series generated in Section IV and fit a GARCH model estimated under the null hypothesis of no structural break, and compute the following structural break tests:

- The *sup*LM, *ave*LM and *exp*LM tests of Andrews (1993) and Andrews and Ploberger (1994) for the null hypothesis that all parameters are constant through the sample.
- The CUSUM test of Inclàn and Tiao (1992) for a break in the unconditional volatility using both the raw and standardized residuals.

The empirical distribution of these test statistics are reported in Table 3. We report the empirical rejection frequencies at the nominal sizes 0.10, 0.05 and 0.01 for each parametrization and sample size. Consider first the LM-based tests. All three statistics exhibit rejection frequencies close to their nominal size in samples of 2500 observations. The *ave*LM test clearly exhibits the best size properties, and the *sup*LM and *exp*LM tests tend to slightly over-reject, particularly in smaller samples. This tendency to over-reject is slightly worse in Model B which has more persistent volatility shocks. The results do not appear to depend on whether we use the normal or Student T(7.5) distributions for the innovations except in very small samples when the *sup*LM and *exp*LM tests over-reject slightly more frequently when the data is Student T. Consider next the CUSUM test of Inclán and Tiao (1992). The test using the raw returns over-rejects at all nominal sizes. At the 5% level the rejection frequency is over 50% in modest samples, and this problem grows with the sample size. Interestingly, when returns are normally distributed the test based on standardized residuals has approximately the correct sample size. However, when returns have fat tails even the IT test using standardized residuals rejects far too frequently.

We next study the power of the structural break tests to identify breaks. The empirical power is reported in Table 4. As previously, we report the results for sample sizes of 500 and 2500 observations and report results for a break induced after 30% of

Table 4. The power of the structural break tests

Model	π	*sup*LM	*ave*LM	*exp*LM	IT	IT(z)	*sup*LM	*ave*LM	*exp*LM	IT	IT(z)
Panel A: $T=500$, Normal											
Pre: $(0.05, 1, 0.3, 0.3)^{\top}$	0.3	0.135	0.114	0.142	0.695	0.176	0.193	0.195	0.205	0.809	0.374
Post: $(0.05, 1.4, 0.3, 0.3)^{\top}$	0.5	0.170	0.154	0.169	0.763	0.257	0.235	0.262	0.265	0.830	0.442
Pre: $(0.05, 1, 0.3, 0.3)^{\top}$	0.3	0.317	0.351	0.364	0.890	0.510	0.398	0.450	0.440	0.916	0.630
Post: $(0.05, 1.8, 0.3, 0.3)^{\top}$	0.5	0.351	0.434	0.416	0.940	0.719	0.382	0.496	0.458	0.949	0.775
Pre: $(0.05, 1, 0.1, 0.8)^{\top}$	0.3	0.188	0.089	0.149	0.708	0.055	0.236	0.165	0.217	0.815	0.301
Post: $(0.05, 1.4, 0.1, 0.8)^{\top}$	0.5	0.147	0.090	0.131	0.748	0.086	0.237	0.176	0.211	0.838	0.341
Pre: $(0.05, 1.8, 0.1, 0.8)^{\top}$	0.3	0.307	0.179	0.284	0.879	0.135	0.345	0.288	0.336	0.923	0.462
Post: $(0.05, 1.8, 0.1, 0.8)^{\top}$	0.5	0.188	0.131	0.178	0.924	0.213	0.290	0.262	0.270	0.943	0.548
Pre: $(0.05, 1, 0.1, 0.5)^{\top}$	0.3	0.118	0.059	0.107	0.197	0.024	0.186	0.137	0.175	0.431	0.167
Post: $(0.05, 1, 0.1, 0.6)^{\top}$	0.5	0.127	0.063	0.111	0.191	0.024	0.222	0.138	0.209	0.448	0.185
Pre: $(0.05, 1, 0.1, 0.5)^{\top}$	0.3	0.116	0.063	0.108	0.438	0.029	0.167	0.129	0.171	0.632	0.192
Post: $(0.05, 1, 0.1, 0.8)^{\top}$	0.5	0.131	0.083	0.129	0.370	0.035	0.199	0.159	0.194	0.619	0.211
Pre: $(0.05, 1, 0.1, 0.5)^{\top}$	0.3	0.115	0.073	0.110	0.285	0.032	0.174	0.142	0.166	0.500	0.175
Post: $(0.05, 1, 0.2, 0.4)^{\top}$	0.5	0.141	0.092	0.131	0.256	0.034	0.214	0.176	0.206	0.504	0.211
Pre: $(0.05, 1, 0.1, 0.5)^{\top}$	0.3	0.150	0.136	0.164	0.413	0.068	0.212	0.219	0.240	0.618	0.227
Post: $(0.05, 1, 0.3, 0.3)^{\top}$	0.5	0.188	0.176	0.195	0.357	0.068	0.252	0.264	0.278	0.596	0.234
Panel B: $T=2500$, Normal											
Pre: $(0.05, 1, 0.3, 0.3)^{\top}$	0.3	0.637	0.675	0.701	0.971	0.858	0.609	0.666	0.649	0.980	0.870
Post: $(0.05, 1.4, 0.3, 0.3)^{\top}$	0.5	0.783	0.814	0.816	0.991	0.943	0.697	0.784	0.742	0.991	0.926
Pre: $(0.05, 1, 0.3, 0.3)^{\top}$	0.3	0.999	0.998	1.000	1.000	0.999	0.991	0.992	0.995	0.999	0.999
Post: $(0.05, 1.8, 0.3, 0.3)^{\top}$	0.5	0.999	0.999	0.999	1.000	1.000	0.998	0.998	0.998	1.000	1.000
Pre: $(0.05, 1, 0.1, 0.8)^{\top}$	0.3	0.456	0.445	0.471	0.976	0.627	0.480	0.499	0.484	0.975	0.745
Post: $(0.05, 1.4, 0.1, 0.8)^{\top}$	0.5	0.410	0.487	0.451	0.991	0.785	0.438	0.532	0.478	0.990	0.826
Pre: $(0.05, 1, 0.1, 0.8)^{\top}$	0.3	0.963	0.937	0.962	0.998	0.981	0.926	0.924	0.929	0.999	0.982
Post: $(0.05, 1.8, 0.1, 0.8)^{\top}$	0.5	0.861	0.908	0.905	1.000	0.996	0.871	0.921	0.906	0.999	0.994
Pre: $(0.05, 1, 0.1, 0.5)^{\top}$	0.3	0.068	0.046	0.061	0.215	0.035	0.122	0.119	0.122	0.524	0.194
Post: $(0.05, 1, 0.1, 0.6)^{\top}$	0.5	0.067	0.051	0.070	0.206	0.036	0.123	0.120	0.120	0.531	0.232
Pre: $(0.05, 1, 0.1, 0.5)^{\top}$	0.3	0.155	0.188	0.189	0.569	0.068	0.253	0.331	0.310	0.805	0.294
Post: $(0.05, 1, 0.1, 0.8)^{\top}$	0.5	0.308	0.336	0.359	0.485	0.081	0.421	0.501	0.490	0.754	0.314
Pre: $(0.05, 1, 0.1, 0.5)^{\top}$	0.3	0.140	0.186	0.176	0.317	0.057	0.206	0.272	0.246	0.640	0.261
Post: $(0.05, 1, 0.2, 0.4)^{\top}$	0.5	0.212	0.258	0.266	0.288	0.062	0.270	0.340	0.308	0.614	0.282
Pre: $(0.05, 1, 0.1, 0.5)^{\top}$	0.3	0.602	0.623	0.659	0.525	0.170	0.569	0.646	0.635	0.752	0.380
Post: $(0.05, 1, 0.3, 0.3)^{\top}$	0.5	0.725	0.781	0.795	0.455	0.187	0.673	0.773	0.738	0.731	0.425

Notes: We report the empirical power of the *sup*LM, *ave*LM and *exp*LM tests (using the 10, 5 and 1%, critical values found in Andrews (1993) and Andrews and Ploberger (1994)) and the IT tests using raw (IT) and standardized innovations (IT(z)) to detect structural breaks of various forms in a range of samples. We generate 1000 artificial samples with 500 or 2500 observations each. We introduce a structural break after the fraction of observations $\pi=\{0.3, 0.5\}$. The parameter vector describing the data generating process is $\theta=(\mu,\sigma^2,\alpha,\beta)^{\top}$ and the value before the break is denoted by 'Pre' and after the break by 'Post'. For each combination of Pre and Post parameters we report two rows of results, one for each π. We report the rejection frequencies using the 5% critical value with data generated using both normally distributed and conditionally student T distributed (with 7.5 degrees of freedom) data.

the sample (the first row) and after half of the sample (the second row). The *LM*-based tests exhibit reasonable power to detect breaks in the unconditional level of volatility (though the power of the *ave*LM test for small samples is poor) and a change in the GARCH parameters that leaves persistence unchanged. The power is quite impressive when sample sizes are at least 2500 which is quite common when using daily data, even when the change is only very modest. However, the tests have very low power to detect small increases in β when α is left fixed, even in very large samples. The broad conclusion that can be taken from this simple experiment is that all LM-based structural break tests perform admirably when applied to GARCH models.

Finally, consider the IT CUSUM-based tests. The IT(z) test has power comparable to the LM-based tests to detect breaks in the unconditional volatility. However, these tests have no power to detect parameter changes that affect volatility dynamics that do not affect the unconditional level of volatility. The results for the IT tests based on unstandardized returns appear to have tremendous

Table 5. Power to detect a break in the degrees of freedom

Parameters	π	*sup*LM	*ave*LM	*exp*LM	IT	IT(z)	WAR	WARCH
Panel A: $T=500$								
$(0.05, 1, 0.3, 0.3)^{\mathrm{T}}$	0.3	0.216	0.184	0.221	0.693	0.218	0.065	0.028
	0.5	0.244	0.224	0.255	0.724	0.240	0.054	0.027
$(0.05, 1, 0.1, 0.8)^{\mathrm{T}}$	0.3	0.227	0.174	0.221	0.735	0.193	0.064	0.046
	0.5	0.222	0.193	0.222	0.754	0.252	0.064	0.042
Panel B: $T=2500$								
$(0.05, 1, 0.3, 0.3)^{\mathrm{T}}$	0.3	0.496	0.554	0.548	0.865	0.254	0.049	0.039
	0.5	0.526	0.649	0.609	0.900	0.348	0.055	0.044
$(0.05, 1, 0.1, 0.8)^{\mathrm{T}}$	0.3	0.511	0.561	0.571	0.883	0.248	0.044	0.044
	0.5	0.532	0.645	0.598	0.921	0.330	0.044	0.043

Notes: We report the empirical power of the *sup*LM, *ave*LM and *exp*LM tests (using critical values found in Andrews (1993) and Andrews and Ploberger (1994)), the IT tests using raw (IT) and standardized innovations (IT(z)) and Wooldridge's autocorrelation (WAR) and ARCH test (WARCH) to detect structural breaks in the degrees of freedom. We generate 1000 artificial samples with 500 or 2500 observations each. We introduce a structural break after the fraction of observations $\pi=\{0.3, 0.5\}$. The parameter vector describing the data generating process is $\theta=(\mu,\sigma^2,\alpha,\beta)^{\mathrm{T}}$. Before the break the returns are conditionally Student T with 10 degrees of freedom and after the break the degrees of freedom decreases to 5. We report the rejection frequencies using the appropriate 5% critical values.

power but this is illusory because the rejection frequency is around 50% when the data does not contain a structural break.

To summarize the main results, we find that traditional diagnostic tests for autocorrelation and ARCH have no power to detect structural breaks. The IT test based on raw returns rejects too frequently when the null model is true. The IT test based on standardized residuals has good size when the data is normally distributed and has good power to detect breaks in unconditional volatility. However, the test cannot detect breaks in only volatility dynamics, and rejects too frequently when the data are fat tailed. Finally, the LM-based tests have about the correct size even when returns are fat-tailed, and have impressive power to detect a range of changes in the dynamics of volatility.

VII. Detecting Breaks in the Degrees of Freedom

Until now our focus has been on the ability of standard diagnostic tests (i.e., for serial correlation and volatility clustering) and structural break tests to detect breaks in the level and dynamics of conditional volatility. In this section we study the ability of these same tests to detect breaks in the degrees of freedom of the conditionally Student T GARCH model of Bollerslev (1987). Part of the motivation of this is the findings of Dueker (1997) that shifts in the degrees of freedom are important in Markov regime-switching GARCH models of stock returns.[5]

We conduct a Monte Carlo experiment in which data is simulated from the GARCH-T model with both high and low persistence with both 500 and 2500 observations. We introduce a break after both 30 and 50% of the sample. Before the break the innovations are generated as Student T with 10 degrees of freedom. After the break we increase the leptokurtosis in the data by decreasing the degrees of freedom to only five. The empirical rejection frequencies are reported at the nominal 5% level in Table 5. We find results that are quite similar to the previous tests. Standard diagnostic tests which are routinely applied have no power to detect changes in the degrees of freedom, while the LM-based tests reject the null about 20% the time when the sample size is 500, and the power increases to around 50 to 60% when the sample is increased to 2500 observations. Furthermore, the power is higher when the break occurs roughly in the middle of the sample. The CUSUM tests appear to have impressive power, but recall that the nominal size is severely distorted. The IT test rejects at the 5% level in 75–85% of samples, and the IT(z) test rejects in about 20% of the time when the returns are conditionally Student T. If anything the IT and IT(z) tests reject less frequently when there is a break in the degrees of freedom than when there is no break in the data.

[5] I thank an anonymous referee for suggesting this experiment.

Table 6. Summary statistics

Series	Sample	NOBS	Mean	Variance	Skewness	Kurtosis	BJ	$Q(5)$	$Q^{\mathrm{ARCH}}(5)$	$Q^2(5)$
VWNYSE	1/90–12/02	3280	0.0220	1.0255	−0.1252*	7.1627*	2376.79*	16.72*	7.42	669.63*
CAD	1/90-3/03	3327	0.0070	0.0984	0.0394	4.9820*	545.41*	11.63*	8.15	167.88*
JPY	1/90-3/03	3327	−0.0058	0.5067	−0.5600*	7.1798*	2595.84*	13.46*	7.46	288.91*
GDM	1/90–12/01	2263	−0.0011	0.4476	0.0162	4.7687*	295.06*	7.65	5.09	185.00*
GBP	1/90–3/03	3327	−0.0010	0.3286	−0.2447*	5.6552*	1010.51*	19.75*	12.25*	339.78*
ATT	1/90–12/02	3251	0.0356	4.3966	0.2645*	7.8121*	3174.66*	10.74	5.38	363.45*
BS	1/90–12/02	3124	−0.0101	19.3385	0.5689*	76.9588*	712167.30*	26.92*	2.13	731.18*
GE	1/90–12/02	3270	0.0729	3.0657	0.1647*	6.7296*	1909.97*	16.66*	8.03	476.28*
GM	1/90–12/02	3255	0.0455	4.2901	0.1052*	4.8325*	461.46*	10.49	7.18	262.63*
IBM	1/90–12/02	3263	0.0741	4.5080	0.3138*	8.7794*	4594.83*	10.39	6.60	82.95*
PM	1/90–12/02	3264	0.0750	4.0087	−0.4815*	14.3871*	17760.84*	12.31*	7.80	43.89*
TX	1/90–12/02	2961	0.0594	2.1823	0.3470*	5.5308*	849.65*	18.80*	12.49*	139.58*

Notes: BJ is the Bera–Jarque statistic and is distributed as chi-squared with 2 degrees of freedom, $Q(5)$ is the Box-Pierce Portmanteau statistic, $Q^{\mathrm{ARCH}}(5)$ is the Box-Pierce Portmanteau statistic adjusted for ARCH effects following Diebold (1986) and $Q^2(5)$ is the Box–Pierce test for serial correlation in the squared residuals. The three Q statistics are calculated with 5 lags and are distributed as chi-squared with 5 degrees of freedom.
* Denotes a skewness, kurtosis, BJ or Q statistic significant at the 5% level.

VIII. An Empirical Application

We illustrate the various structural break tests using 12 daily financial time-series starting in January 1990. We focus on this relatively short sample period because it seems reasonable to expect that a GARCH model's parameters may be constant over this short period of a little over a decade. This period also avoids the 1987 stock market crash which is particularly relevant for the equity series.[6]

Data and summary statistics

We consider GARCH models for 12 common time series which are observed daily:

- The excess returns on value-weighted NYSE stock index from CRSP (VWNYSE) which spans January 1990 to December 2002.
- Four US dollar denominated exchange rate series: the Canadian Dollar (CAD), the Japanese Yen (JPY), the German Deutsche Mark (GDM) and the British Pound (GBP). Our data starts in January 1990 and finishes in March 2003 except for the GDM which finishes in December 1999 when it was replaced with the Euro.
- Daily returns on seven stocks from January 1990 to December 2002 are used: ATT, Bethlehem Steel (BS), General Electric (GE), General Motors (GM), IBM (IBM), Philip Morris (PM), Texaco (TX).

Summary statistics of the data are reported in Table 6. The data present the standard set of well-known stylized facts about financial time series: non-normality, limited evidence of short-term predictability and strong evidence of predictability in volatility. All series are presented in daily percentage growth rates/returns. Individual stocks have the highest unconditional variance, while the exchange rate series have the lowest. The mean return on the individual stock is significantly higher because they are measured as raw returns, while the index is measured in excess returns.

The Bera–Jarque test conclusively rejects normality in all series. The smallest test statistic is almost 300 which is 50 times larger than the 5% critical value of 5.99. The stock index is negatively skewed and has fat tails. Roughly half the exchange rates and individual stock returns are negatively skewed and the others are positively skewed, though the negatively skewed series are significantly larger in magnitude. The asymptotic SE of the skewness statistic under the null of normality is $\sqrt{6/T}$, and the SE of the kurtosis statistic is $\sqrt{24/T}$. Approximately two-thirds of the skewness statistics are statistically different from zero (using the asymptotic Z-score), while all 12 series exhibit statistically significant leptokurtosis, suggesting that accounting for fat tails is more pressing than modelling skewness. Individual stocks in particular exhibit significant excess kurtosis. This kurtosis could alternatively be interpreted as realizations from a

[6] We have also considered results for longer sample periods dating back to July 1963 for the stock returns and 1974 for the foreign exchange rates. The full sample period results are much the same as our post-1989 data and are available on request. It should be no surprise that the evidence of structural breaks that we document for the post-1989 samples is much stronger in the longer sample.

distribution with extremely high kurtosis, or as a series containing outliers. We choose to follow Bollerslev (1987) and model the standardized residual using the fat-tailed Student T random variable, and report robust SEs to account for possible unmodelled skewness.

We use the Box–Pierce portmanteau, or Q, statistic with five lags to test for serial correlation in the data, and we adjust the Q statistic for ARCH following Diebold (1986). The standard Q-test indicates autocorrelation in all but GDM, however after adjusting for possible ARCH effects the evidence becomes much weaker. The evidence of linear dependence in the squared demeaned returns, which is an indication of ARCH effects in the data, is much stronger.

Empirical results

For each time series we must choose the correct time-series model. We fit all possible AR(m)–GJR(p, q, r) models with $m, p, q, r \leq 2$. We estimate both Gaussian and Student T-based models (the degrees of freedom is estimated as a free parameter). We choose and report the model that has the lowest Bayesian information criteria (BIC). When the chosen conditional mean or volatility is rejected as being miss-specified using Wooldridge's tests we consider the model with the highest Akaiki information criteria (AIC). The AIC invariably chooses models that are more heavily parameterized. We accept this model if it's conditional mean and variance are better specified (again using the robust specification tests) and have a BIC in roughly the vicinity of the maximal BIC.

We report the parameter estimates of these models along with robust SDs in parentheses in Table 7. All series required the fat-tailed Student T distribution. We find that symmetric GARCH models are suitable for the exchange rates, but the VWNYSE index requires an asymmetry term. Three of the seven individual stocks also required asymmetric terms. The most parsimonious models were found for most series: only four out of the 12 series required ARCH or GARCH models with more than one lag. This is interesting in light of the conclusions of Hansen and Lunde (2005) who found that the GARCH(1,1) model performs better than most rivals in terms of forecasting volatility. Two of the individual stocks are GARCH(2,1) in which α_2 is negative. As noted above, whenever α_1 and β_1 are positive, but α_2 is negative with $-\alpha_2 < \alpha_1\beta_1$, the volatility forecasts are strictly positive.

We report diagnostic tests of the GARCH models in Table 8. Virtually all series pass Wooldridge's robust regression based tests for the specification of the conditional mean and variance, as described in Section III. The models thus appear to be correctly specified. We also report the three LM-based structural break tests for each of the six parameter groupings, and the IT and IT(z) CUSUM tests. Despite the models passing traditional diagnostic tests, a very different picture emerges from the structural break tests. There is strong evidence of a structural break in the S&P500 series, CAD and all of the individual stocks with the exception of ATT. There is only weak or no evidence of a break in the other three exchange rates (JPY, GBP, GDM) and ATT.

Considering the LM-based tests, there is some evidence of structural breaks in all series: at least one of the structural break tests finds a structural break in at least one parameter grouping. The weakest evidence of structural breaks is in the GDM where only the *sup*LM test finds a structural break in only the degree of freedom parameter. It is interesting that evidence of a break in the degrees of freedom parameter is rather robust across all series: only four individual stocks have all three structural break test statistics that are insignificant (though the degree of freedom parameters for ATT, GE, GM, TX are not significant). Most of the other series have several of the structural break test statistics that are significant in most of the parameter groupings. In the individual stock return series the evidence of structural breaks is strongest in the volatility parameters. It seems that both the unconditional volatility and the GARCH parameters move around even in only very recent and short samples.

Recall that we discard a fraction of the observations at the beginning and end of the sample so the LM test statistics do not diverge. When a *sup*LM test statistic does not converge the supremum occurs at the boundary. We therefore report the point at which the supremum in the *sup*LM statistic occurs. It is comforting that most of the *sup*LM statistics occur quite far from the boundaries. An interesting exception is for GBP where among the LM-tests the strongest evidence for a break (in ALL, UC Mean and GARCH) which all occur near the initial boundary, but only very weak or no evidence from the other tests. There is also only weak evidence for a break from the IT(z) test. This evidence suggests that the basic GARCH model for the GBP series is well specified. In a similar vein we find evidence from the IT(z) test for a break in the ATT series. The only LM test suggesting a break is the *sup*LM test for a break in the volatility dynamics (GARCH) and the supremum occurs at the end of the sample. Again the evidence for a break must be interpreted with caution.

In all 12 series the IT test based on raw data appears to be significant. However, the evidence from

Table 7. Parameter estimates for the GARCH models

Data	μ		ϕ		σ^2		α		β_1		β_2		δ		ν		LL
VWNYSE	0.0345	(0.0116)	0.0941	(0.0175)	0.9324	(0.3004)	0.0114	(0.0067)	0.9203	(0.0158)	–	–	0.1202	(0.0278)	0.1326	(0.0180)	−4030.4494
CAD	0.0014	(0.0044)	0.0452	(0.0175)	0.1807	(0.0765)	0.1061	(0.0153)	0.0894	(0.0471)	0.7991	(0.0469)	–	–	0.1424	(0.0148)	−610.8807
JPY	0.0151	(0.0101)	–	–	0.4953	(0.0710)	0.0339	(0.0068)	0.9544	(0.0097)	–	–	–	–	0.1931	(0.0159)	−3285.3870
GDM	0.0025	(0.0118)	–	–	0.5598	(0.2151)	0.0533	(0.0115)	0.9387	(0.0138)	–	–	–	–	0.1429	(0.0195)	−2143.4955
GBP	0.0125	(0.0078)	–	–	0.4600	(0.1470)	0.0544	(0.0127)	0.9395	(0.0151)	–	–	–	–	0.1777	(0.0157)	−2547.0368

Data	μ		ϕ_1		ϕ_2		σ^2		α_1		α_2		β_1		β_2		δ_1		ν		LL
ATT	0.0109	(0.0258)	0.0068	(0.0193)	−0.0549	(0.0185)	3.8352	(1.3288)	0.1526	(0.0469)	−0.1312	(0.0471)	0.9771	(0.0136)	–	–	–	–	0.1617	(0.0161)	−6435.5171
BS	−0.1395	(0.0467)	−0.0339	(0.0198)	−0.0499	(0.0183)	11.2855	(2.0873)	0.2343	(0.0406)	−0.1976	(0.0413)	0.9489	(0.0191)	–	–	–	–	0.2102	(0.0198)	−7867.8929
GE	0.0715	(0.0239)	–	–	–	–	3.0525	(0.6585)	0.0137	(0.0074)	–	–	0.9436	(0.0138)	–	–	0.0711	(0.0173)	0.0851	(0.0149)	−6028.3604
GM	0.0208	(0.0329)	−0.0195	(0.0184)	–	–	4.5840	(0.5176)	0.0302	(0.0155)	–	–	0.9182	(0.0251)	–	–	0.0539	(0.0137)	0.1091	(0.0157)	−6811.6376
IBM	0.0403	(0.0296)	−0.0181	(0.0176)	–	–	4.9335	(0.8662)	0.0163	(0.0067)	–	–	0.9274	(0.0187)	–	–	0.0789	(0.0239)	0.1890	(0.0159)	−6683.7419
PM	0.1034	(0.0256)	–	–	–	–	4.7448	(1.3509)	0.1042	(0.0231)	–	–	0.8669	(0.0318)	–	–	–	–	0.2219	(0.0195)	−6406.7283
TX	0.0489	(0.0226)	−0.0387	(0.0196)	−0.0763	(0.0194)	1.8265	(0.3876)	0.0507	(0.0118)	–	–	0.0886	(0.0549)	0.8554	(0.0602)	–	–	0.0979	(0.0151)	−5063.3049

Note: We report the quasi-maximum likelihood parameter estimates of the AR(m)-GJR(p, q, r) model with conditional mean and variance given in Equations 1 and 2. We also report the value of the Log-likelihood function (LL) and robust SE in parenthesis below each parameter estimate.

Table 8. Diagnostic tests for the GARCH and GJR models

Data	WAR	WARCH	IT	All	UC mean	AR	UC volatility	GARCH	DF
VWNYSE	9.201	6.244	13.315***	50.193***	10.855***	18.763***	13.759***	16.663***	25.487***
	(0.101)	(0.283)	(1.915)***	(27.882)***	(2.094)*	(7.378)***	(3.635)***	(8.771)***	(11.930)***
				[21.315]***	[2.204]**	[6.325]***	[3.885]***	[5.811]**	[9.966]***
				0.686	0.388	0.686	0.532	0.437	0.521
CAD	16.079	4.829	6.115***	48.260***	13.610***	1.940	18.131***	29.442***	14.886***
	(0.007)	(0.437)	(0.991)	(16.450)***	(1.768)	(0.530)	(3.784)***	(10.587)***	(5.272)***
				[17.577]***	[2.280]**	[0.296]	[3.779]***	[8.974]***	[4.754]***
				0.138	0.139	0.105	0.138	0.138	0.512
JPY	8.431	8.328	4.249***	13.098	10.198**	–	7.495	6.979	9.154*
	(0.134)	(0.139)	(1.085)	(6.696)	(1.857)		(1.720)	(1.876)	(2.553)*
				[4.075]	[2.101]***		[1.311]	[1.361]	[1.961]*
				0.390	0.400		0.532	0.823	0.587
GDM	7.272	8.256	4.644***	15.860	5.248	–	3.977	8.577	7.805*
	(0.201)	(0.143)	(1.132)	(5.718)	(1.799)		(1.492)	(2.125)	(1.776)
				[4.681]	[1.159]		[0.935]	[1.717]	[1.425]
				0.620	0.620		0.637	0.620	0.525
GBP	2.479	12.007	9.279***	26.850***	10.403**	–	3.460	17.237***	6.338
	(0.780)	(0.035)	(1.255)*	(8.229)	(2.324)		(0.847)	(2.680)	(2.048)*
				[6.969]*	[1.579]*		[0.544]	[2.991]*	[1.404]
				0.069	0.050		0.282	0.068	0.690
ATT	0.010	3.201	13.128***	20.006	4.403	9.632	4.272	8.978***	6.912
	(1.000)	(0.669)	(2.230)***	(8.316)	(0.823)	(3.034)	(1.061)	(2.448)	(1.141)
				[5.591]	[0.535]	[2.445]	[0.601]	[1.598]	[0.814]
				0.947	0.271	0.223	0.359	0.950	0.947
BS	0.996	3.060	20.485***	107.931***	4.732	3.634	15.563***	62.940***	24.061***
	(0.963)	(0.691)	(4.890)***	(28.943)***	(1.503)	(1.530)	(3.387)***	(7.904)***	(2.505)***
				[47.377]***	[0.961]	[0.883]	[4.232]***	[24.933]***	[7.080]***
				0.928	0.670	0.476	0.624	0.928	0.921
GE	9.401	10.391	11.686***	26.283***	7.050	–	17.847***	21.719***	4.992
	(0.094)	(0.065)	(2.219)***	(14.649)***	(1.587)		(5.474)***	(6.712)**	(1.929)
				[9.998]***	[1.331]		[5.686]***	[5.161]**	[1.121]
				0.080	0.368		0.314	0.080	0.081
GM	4.770	7.819	6.101***	31.365***	2.027	14.275***	10.928***	17.922**	2.704
	(0.445)	(0.166)	(1.910)***	(13.058)***	(0.156)	(6.019)***	(2.737)**	(3.008)	(0.344)
				[10.609]***	[0.086]	[4.321]***	[2.700]**	[4.202]*	[0.202]
				0.950	0.947	0.698	0.636	0.936	0.868
IBM	6.819	4.495	7.960***	49.074***	8.435*	4.943	33.555***	31.898***	13.370***
	(0.234)	(0.481)	(2.993)***	(25.883)***	(2.177)*	(1.785)	(14.929)***	(10.325)***	(5.893)***
				[20.311]***	[1.835]*	[1.044]	[12.979]***	[11.325]***	[4.065]***
				0.213	0.291	0.721	0.207	0.207	0.517
PM	4.338	4.567	7.977***	23.228***	1.953	–	17.679***	13.265**	14.430***
	(0.502)	(0.471)	(2.697)***	(9.122)**	(0.455)		(5.442)***	(4.883)**	(3.794)**
				[8.098]**	[0.256]		[5.702]***	[3.978]**	[3.834]***
				0.574	0.337		0.523	0.523	0.574
TX	6.209	8.729	10.000***	22.165*	2.971	13.035***	7.452	15.027*	2.216
	(0.286)	(0.120)	(1.580)**	(12.286)*	(0.597)	(3.624)*	(2.885)**	(7.348)**	(0.240)
				[7.824]*	[0.354]	[3.033]*	[1.746]*	[4.505]**	[0.140]
				0.188	0.429	0.198	0.053	0.235	0.055

Notes: We report Wooldridge's robust LM test for fifth order autocorrelation (denoted WAR) and ARCH effects (denoted WARCH) with *p*-values in parenthesis, Inclan and Tiao's (1994) CUSUM structural break test (denoted IT) using the raw returns and in parenthesis using the standardized residuals, and the *sup*LM, *ave*LM (in parenthesis) and the *exp*LM [in brackets] structural break tests of Andrews (1993) and Andrews and Ploberger (1994). Below the *ave*LM test we report the *sup*LM-based estimate of the break date. The first two tests are distributed as chi-squared with 5 degrees of freedom, the IT test has the Kolmogorov–Smirnof distribution and the structural break tests have nonstandard distributions with tabulated critical values reported in the source articles. The level of significance for the structural break tests is indicated by: *, ** and *** at the 10, 5 and 1% levels.

the standardized IT(z) test and the LM-based tests yield much weaker evidence. This provides more reasons for caution when applying CUSUM tests based on raw returns when those returns exhibit leptokurtosis beyond that which can be explained by GARCH alone.

Interestingly, the standardized CUSUM test IT(z) provides very strong evidence of a break in all the stock return series (i.e., both index and individual returns), but no or only very weak evidence for the exchange rates where the most evidence is for the GBP and that is only at the 10% level.

IX. Conclusions

In this article, we have considered a range of diagnostic tests which can be used to determine the adequacy of GARCH models which have been applied to numerous financial time series. In particular, we employ the robust-regression based diagnostic tests of Wooldridge (1990) to test the specification of the conditional mean and variance. We also use the LM-based statistics proposed by Andrews (1993) and Andrews and Ploberger (1994) and the CUSUM test of Inclán and Tiao (1992) to test for structural stability.

We conduct a simulation experiment to assess the size and power properties of these miss-specification and structural break tests. We find that the robust diagnostic tests of Wooldridge (1990) and the LM-based tests generally have good size properties in samples of reasonable size. The *ave*LM test in particular has excellent size even in modest samples that are as small as 500 observations. The IT test applied with standardized residuals has approximately the correct size when the data are conditionally normal, but rejects too frequently when the data has fat tails. The test also rejects far too frequently when applied to the raw data. If there is reason to believe the data may be fat tailed one should steer clear of the IT tests. The finding of the weakness in the IT test is particularly unfortunate because the IT test is much easier to calculate than the LM statistics. Furthermore, the LM tests consider only a single break date, while the IT tests are capable of detecting multiple breaks.

The LM-based structural break tests generally exhibit impressive power to detect modest changes in the dynamics of volatility particularly in large samples. The tests are able to detect a range of changes in the GARCH model including changes in the level of unconditional volatility and in the dynamics of volatility. However although the IT test is quite impressive in detecting changes in the level of unconditional volatility, it has trouble in detecting changes in volatility dynamics when unconditional volatility is constant. Our results suggest that the extra computation required for the LM-based tests of Andrews (1993) and Andrews and Ploberger (1994) are worth the cost as they are robust to different conditional distributions and can detect a broader scope of structural breaks.

In our empirical application we find strong evidence of structural breaks in GARCH models of several financial time series, including US equity index returns, foreign exchange rates, and individual stock returns. These GARCH models passed standard tests for the specification of the conditional mean and variance. This and our simulation evidence that generic diagnostic tests can easily miss structural breaks suggests that applied econometricians should add structural break tests to their battery of diagnostic tests for GARCH models.

Acknowledgements

I thank Adlai Fisher, an anonymous referee and participants at the 2003 Northern Finance Association meeting in Quebec city, and the 24th International Symposium of Forecasting in Sydney Australia for their comments. I also thank Kyung Shim for his research assistance. Any remaining errors are my own responsibility.

References

Andreou, E. and Ghysels, E. (2002) Detecting multiple breaks in financial market volatility dynamics, *Journal of Applied Econometrics*, **17**, 579–600.

Andrews, D. W. K. (1993) Tests for parameter instability and structural change with unknown change point, *Econometrica*, **61**, 821–56.

Andrews, D. W. K. and Ploberger, W. (1994) Optimal tests when a nuisance parameter is present only under the alternative, *Econometrica*, **62**, 1383–414.

Black, F. (1976) Studies of stock price volatility changes, *Proceedings of the Business and Economic Statistical Section, American Statistical Association*, pp. 177–81.

Bollerslev, T. (1987) A conditional heteroskedastic time series model for speculative prices and rates of return, *Review of Economics and Statistics*, **69**, 542–7.

Bollerslev, T., Chou, R. Y. and Kroner, K. F. (1992) GARCH modelling in finance, *Journal of Econometrics*, **52**, 5–59.

Brenner, R. J., Harjes, R. and Kroner, R. (1996) Another look at models of short term interest rates, *Journal of Financial and Quantitative Analysis*, **31**, 85–107.

Carrasco, M. and Chen, X. (2001) Mixing and moment properties of various garch and stochastic volatility models, *Econometric Theory*, **18**, 17–39.
Christie, A. A. (1982) The stochastic behavior of common stock variances: value, leverage, and interest rate effects, *Journal of Financial Economics*, **19**, 407–32.
Chu, C. S. J. (1995) Detecting parameter shift in GARCH models, *Econometric Reviews*, **14**, 241–66.
Davies, R. B. (1977) Hypothesis testing when a nuisance parameter is present only under the alternative, *Biometrika*, **64**, 247–54.
Davies, R. B. (1987) Hypothesis testing when a nuisance parameter is present only under the alternative, *Biometrika*, **74**, 33–43.
Diebold, F. X. (1986) Testing for serial correlation in the presence of ARCH, *Proceedings of the Business and Economic Statistics Section of the American Statistical Association*, pp. 323–8.
Dueker, M. J. (1997) Markov switching in GARCH processes and mean-reverting stock-market volatility, *Journal of Business and Economic Statistics*, **15**, 26–35.
Engle, R. F. and Bollerslev, T. (1986) Modelling the persistence of conditional variances, *Econometric Reviews*, **5**, 1–50.
Engle, R. F. and Ng, V. K. (1993) Measuring and testing the impact of news on volatility, *Journal of Finance*, **48**, 1749–78.
Ghysels, E. and Hall, A. (1990) Are consumption-based intertemporal capital asset pricing models structural?, *Journal of Econometrics*, **45**, 121–40.
Glosten, L. R., Jagannathan, R. and Runkle, D. E. (1993) On the relation between expected value and the volatility of the nominal excess return on stocks, *Journal of Finance*, **48**, 1779–801.
Gray, S. F. (1996) Modeling the conditional distribution of interest rates as a regime-switching process, *Journal of Financial Economics*, **42**, 27–62.
Haas, M., Stefan, M. and Paolella, M. S. (2004) A new approach to makov-switching GARCH models, *Journal of Financial Econometrics*, **2**, 493–530.
Hall, A. R., Atsushi, I. and Peixe, F. P. M. (2003) Covariance matrix estimation and the limiting behavior of the overidentifying restrictions test in the presence of neglected structural instability, *Econometric Theory*, **19**, 962–83.
Hamilton, J. D. (1994) *Time Series Analysis*, Princeton University Press, Princeton, USA.
Hansen, B. E. (1996) Erratum: the likelihood ratio test under nonstandard conditions: testing the markov-switching model of GNP, *Journal of Applied Econometrics*, **11**, 195–8.
Hansen, B. E. (1997) Approximate asymptotic *P* values for structural-change tests, *Journal of Business and Economic Statistics*, **15**, 60–7.
Hansen, P. R. and Lunde, A. (2001) A forecast comparison of volatility models: does anything beat a GARCH(1,1)?, Working Paper, 01-04 Department of Economics, Brown University.
Hansen, P. R. and Lunde, A. (2005) A forecast comparison of volatility models: does anything beat a GARCH(1,1)?, *Journal of Applied Econometrics*, **20**, 873–89.
He, C. and Terasvirta, T. (1999) Fourth moment structure of the GARCH(p,q) Process, *Econometric Theory*, **15**, 824–46.
Inclán, C. and Tiao, G. C. (1992) Use of cumulative sums of squares for retrospective detection of change of variance, *Journal of the American Statistical Association*, **89**, 913–23.
Klaassen, F. (2002) Improving GARCH volatility forecasts with regime-switching GARCH, *Empirical Economics*, **27**, 363–94.
Kokoszka, P. and Leipus, R. (1998) Change-point in the mean of dependent observations, *Statistics and Probability Letters*, **40**, 385–93.
Kokoszka, P. and Leipus, R. (2000) Change-point estimation in ARCH models, *Bernoulli*, **6**, 1–28.
Lamoureux, C. G. and Lastrapes, W. D. (1990) Persistence in variance, structural change and the GARCH model, *Journal of Business and Economic Statistics*, **8**, 225–34.
Lin, S. J. and Yang, J. (1999) Testing Shifts in Financial Models with Conditional Heteroscedasticity: an Empirical Distribution Function Approach, Working Paper, 30 Quantitative Finance Research Group, University of Technology, Sydney.
Lundbergh, S. and Terasvirta, T. (2002) Evaluating GARCH models, *Journal of Econometrics*, **110**, 417–35.
Malik, F. (2003) Sudden changes in Variance and volatility persistence in foreign exchange markets, *Journal of Multinational Financial Management*, **13**, 217–30.
Nelson, D. (1991) Conditional heteroscedasticity in stock returns: a new approach, *Econometrica*, **59**, 347–70.
Pérignon, C. and Smith, D. R. (2007) Yield-factor volatility models, *Journal of Banking and Finance*, **31**, 3125–44.
Wooldridge, J. M. (1990) A unified approach to robust, regression-based specification tests, *Econometric Theory*, **6**, 17–43.