CORE DISCUSSION PAPER

2003/9

BOOTSTRAP MISSPECIFICATION TESTS FOR ARCH BASED ON THE EMPIRICAL PROCESS OF SQUARED RESIDUALS

Lajos HORVÁTH[1], Piotr KOKOSZKA[2] and Gilles TEYSSIÈRE[3]

January 2003

Abstract

We propose and study by means of simulations and graphical tools a class of goodness–of–fit tests for ARCH models. The tests are based on the empirical distribution function of squared residuals and smooth (parametric) bootstrap. We examine empirical size and power by means of a simulation study. While the tests have overall correct size, their power strongly depends on the type of alternative and is particularly high when the assumption of Gaussian innovations is violated. As an example, the tests are applied to returns on Foreign Exchange rates.

Key words: ARCH model, empirical process, goodness–of–fit tests, size–power curves, smooth bootstrap, squared residuals.

JEL Classification: C12, C22.

[1]University of Utah, Department of Mathematics. 155 South 1440 East, Salt Lake City, UT 84112–0090, USA. E–mail: horvath@math.utah.edu

[2]Utah State University, Mathematics and Statistics. 3900 Old Main Hill, Logan, UT 84322–3900, USA. E–mail: piotr@math.usu.edu

[3]NBGF, GREQAM & CORE, Université Catholique de Louvain. 65, Av. Roosevelt, F-75008 Paris, France. E–mail: gilles@ehess.cnrs-mrs.fr, WWW: http://www.gillesteyssiere.net

We thank the referee for constructive advice which led to a significant improvement of the paper, and Jeroen Rombouts for a careful reading of the paper.

This text presents research results of the Belgian Program on Interuniversity Poles of Attraction initiated by the Belgian State, Prime Minister's Office, Science Policy Programming. The scientific responsibility is assumed by the authors.

1 Introduction

Suppose a sample $y_1, \ldots, y_n$ was observed and an ARCH type model has been postulated as an adequate description of the data. In this paper we are concerned with the problem of verifying if the postulated model fits the data. Tests of this kind are known as goodness–of–fit or misspecifications tests and play a central role in time series analysis.

In the classical "linear" time series analysis, such tests, known also as diagnostic checks, fall roughly into four categories: 1) examination of the residual plot, 2) portmanteau type tests based on weighted sums of covariances of residuals, 3) Lagrange–multiplier tests, 4) tests based on the empirical distribution function of the estimated residuals or on the spectral empirical distribution function (integrated periodogram).

Goodness–of–fit tests for non–linear time series models have only recently become an object of a more systematic research even though basic tools like residual plots have, of course, been used in a more or less intuitive way for a long time. Li and Mak (1994), Horváth and Kokoszka (2001) and Berkes, Horváth and Kokoszka (2002a, 2002b) studied Portmanteau type statistics based on autocorrelations of residuals and squared residuals. Davis and Mikosch (1998) and Mikosch and Stărică (2000) studied the autocorrelation function of the observations following, respectively, ARCH(1) and GARCH(1,1) models. Lundbergh and Teräsvirta (2002) proposed a broad class of Lagrange-multiplier type specification tests whereas Mikosch and Stărică (1999) proposed a goodness–of–fit test for GARCH based on the integrated periodogram of the observations.

Recently, Horváth, Kokoszka and Teyssière (2001) found the asymptotic distribution of the empirical process of ARCH(p) squared residuals and Berkes and Horváth (2002) extended these results to general GARCH(p, q) models. Relevant theoretical results are also presented in Chapter 8 of Koul (2002).

In this paper we follow up on this theoretical work by proposing and examining the finite sample performance of several goodness–of–fit tests based on the empirical distribution of squared residuals. These tests are particularly suited to detect departures from the postulated distribution of the unobservable innovations. To explain the idea, we focus on a simple ARCH(p) model. Suppose then that the postulated model is defined by

$$y_t = \sigma_t \varepsilon_t, \qquad \sigma_t^2 = b_0 + \sum_{j=1}^{p} b_j y_{t-j}^2, \qquad \varepsilon_t \sim N(0,1), \tag{1}$$

and we want to check if this model is an acceptable approximation to the data. Having obtained estimates $\hat{b}_0, \hat{b}_1, \ldots, \hat{b}_p$, we compute the squared residuals

$$\hat{\varepsilon}_k^2 = y_k^2 \left[\hat{b}_0 + \sum_{j=1}^{p} \hat{b}_j y_{k-j}^2 \right]^{-1}, \quad k = p+1, \ldots, n. \tag{2}$$

The empirical process of the squared residuals is then

$$\hat{\alpha}_n(x) = (n-p)^{-1/2} \sum_{k=p+1}^{n} \left(\mathbf{I}[\hat{\varepsilon}_k^2 \le x] - (2\Phi(\sqrt{x}) - 1) \right), \tag{3}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function and $\mathbf{I}[\cdot]$ is the indicator function. If a distribution different from normal is postulated for the ε_t in (1), then $2\Phi(\sqrt{x}) - 1$ in (3) should be replaced by the cumulative distribution function of the ε_t^2.

It is intuitively clear that if (1) is an adequate description of the data, then the $\hat{\varepsilon}_k^2$ should have distribution which is close to that of ε_k^2, so the empirical distribution function

$$\hat{F}(x) = \frac{1}{n-p} \sum_{k=p+1}^{\infty} \mathbf{I}[\hat{\varepsilon}_k^2 \le x], \tag{4}$$

should be close to $2\Phi(\sqrt{x}) - 1$ and consequently the empirical process (3) should be "small". On the other hand, if (1) is a poor approximation to the data, then $\hat{F}(x)$ will be very different from $2\Phi(\sqrt{x}) - 1$ and so $\hat{\alpha}_n$ will be "large". This may be due not only to fact that non–Gaussian innovations yield a better approximation, but also to the misspecified functional form of the model. For example, if ARCH(p') with $p' > p$ better fits to the data, then the estimates $\hat{b}_j$ will have large bias and the $\hat{\varepsilon}_k^2$ will have empirical distribution very different from $2\Phi(\sqrt{x}) - 1$, even when the ε_k can be assumed normal.

In Section 6 we present some mathematical and historical background on goodness–of–fit tests based on the empirical process of the residuals. Here we only mention that, unlike for many linear models, in the case of ARCH models the empirical process has an asymptotic distribution which depends in an intricate way on the distribution of the innovations and model parameters. For this reason asymptotically pivotal statistics based on the empirical process cannot be readily constructed. Nevertheless, by employing some form of bootstrap, standard goodness–of–fit tests might be appropriately modified. In this paper we propose a simple procedure of this kind and investigate its validity by means of a simulation study. We focus on three functionals of the empirical process $\hat{\alpha}_n$ of squared residuals:

- the Cramér–von Mises (CVM) statistic

$$\widehat{T}_n = \int |\hat{\alpha}_n(x)|^2 \, dx. \tag{5}$$

- The normalized Cramér–von Mises (NCVM) statistic

$$\widehat{NT}_n = \int |\hat{\alpha}_n(x)|^2 \, (2\pi)^{-1/2} \exp\left(-x^2/2\right) dx. \tag{6}$$

- The Kolmogorov–Smirnov (KS) statistic

$$\widehat{KS}_n = \max_{1 \le i \le n} \left|\hat{\alpha}_n(\hat{\varepsilon}_i^2)\right|. \tag{7}$$

In the case of model (1), our goodness–of–fit procedure works as follows. Denote by $\hat{\tau}_n$ the statistic of interest, where $\hat{\tau}_n = \widehat{T}_n, \widehat{NT}_n, \widehat{KS}_n$.

1. Having observed the sample $y_1, \dots, y_n$, obtain an estimate $\hat{\mathbf{b}}_n = (\hat{b}_0, \hat{b}_1, \dots, \hat{b}_p)$ using a pseudo maximum likelihood method (PML) described in Section 2.2.

2. Generate B independent sequences of iid standard normal random variables, each of length $n + w$. (The extra w values are used to avoid an initialization effect, we set here $w = 10$.) Denote each of the B sequences by $\varepsilon^*_{-w+1}, \dots, \varepsilon^*_0, \varepsilon^*_1, \dots, \varepsilon^*_n$.

3. Generate B bootstrap ARCH(p) realizations $y_1^*, \ldots, y_n^*$, where the y_t^* satisfy

$$y_t^* = \sigma_t^* \varepsilon_t^*, \quad (\sigma_t^*)^2 = \hat{b}_0 + \sum_{j=1}^{p} \hat{b}_j (y_{t-j}^*)^2. \tag{8}$$

   Thus $\{y_t^*\}$ is a smooth bootstrap version of the sequence $\{y_t\}$. The choice of B is discussed in Section 2.1.

4. For each of the B samples $y_1^*, \ldots, y_n^*$ obtain estimates $\hat{b}_0^*, \hat{b}_1^*, \ldots, \hat{b}_p^*$ and construct bootstrap residuals

$$(\hat{\varepsilon}_k^2)^* = (y_k^*)^2 \left[\hat{b}_0^* + \sum_{j=1}^{p} \hat{b}_j^* (y_{k-j}^*)^2 \right]^{-1}, \quad k = p+1, \ldots, n. \tag{9}$$

   From the $(\hat{\varepsilon}_k^2)^*$ construct the bootstrap empirical process $\hat{\alpha}_n^*$ and compute the statistic $\hat{\boldsymbol{\tau}}_{\boldsymbol{n}}^*$. Then find, say, the 5 per cent critical region which corresponds to the upper tail of the empirical distribution of the B numbers $\hat{\boldsymbol{\tau}}_{\boldsymbol{n}}^*$, or the P-value.

The paper is organized as follows. In Section 2 we describe our simulation study. Section 3 focuses on the presentation of the simulation results with graphical methods, whereas Section 4 contains an application of the tests proposed above to returns on Foreign exchange rates. We conclude in Section 5 and provide some mathematical and historical background in Section 6.

2 Simulation Study

In this section, we describe the details our simulation study.

2.1 Bootstrap tests and the choice of B

We consider $N = 10000$ replications of the DGP (data generating process), and for each replication j, we follow the test procedure described in Section 1. For a test of size α, we calculate the bootstrap critical values $\hat{C}_\alpha^*(j)$ as the $(1-\alpha)$th percentiles of the $\hat{\tau}_n^*$. We consider $\alpha = 1\%$, 5%, and 10%. For each replication j, we also calculate the bootstrap P-value $\hat{P}_j^*$, $j = 1, \ldots, N$, which is the empirical probability of observing a statistic $\hat{\tau}^*$ greater than $\hat{\tau}_n$, i.e.

$$\hat{P}_j^* = \hat{P}_j^*(\hat{\tau}_n) = \frac{1}{B} \sum_{i=1}^{B} \mathbf{I}[\hat{\tau}_i^* > \hat{\tau}_n]. \tag{10}$$

The Gaussian errors have been generated using the Box-Muller method, which uses uniform deviates randomly drawn from two different random number generators, initialized with two different seeds. The resulting sequence of uniform deviates succeeds the whole set of Marsaglia's (1996) DIEHARD tests.

An important point is the choice of the number of draws B from the bootstrap DGP, as it involves a trade-off between power loss and computing time. Given that the computation of the CVM and NCVM bootstrap tests involve numerical integration, which is the most computing time consuming operation for these tests, we have to choose B in an optimal way to avoid excessive computing time. To explain how we intend to achieve this goal, we follow below the exposition of Davidson and MacKinnon (1999, 2000).

Let P^* be the ideal bootstrap P-value, i.e. the probability that $\hat{\tau}_n^* > \hat{\tau}_n$ under the bootstrap DGP. The ideal bootstrap test rejects the null hypothesis at level α if $P^* < \alpha$, while the feasible bootstrap test rejects the null hypothesis at level α if $\hat{P}^* < \alpha$. As $B \to \infty$, $\hat{P}^* \to P^*$, but since the number of draws B is finite, bootstrap tests involve some loss of power. The loss of power caused by a finite number of draws has been studied for Monte Carlo tests. These results are of interest in our case since bootstrap tests are Monte Carlo tests for pivotal statistics. For a Monte Carlo test to be exact, $\alpha(B+1)$ must be an integer, see Dufour and Kiviet (1998) and Dufour and Khalaf (2001). Even though the statistics considered in this paper are not pivotal, we follow the recommendation of Davidson and MacKinnon (2000) and always choose B so that $\alpha(B+1)$ is an integer.

In our simulations we used a pretesting procedure suggested by Davidson and MacKinnon (2000) which, they argue, is more efficient than the three-step method of Andrews and Buchinsky (2000). This procedure consists in starting by choosing B relatively small, e.g. $B = 99$, and increasing it until we are confident at some level β, say $\beta = 0.001$, that the estimated bootstrap P-value $\hat{P}^*(\hat{\tau}_n)$ does not differ from the ideal bootstrap P-value $P^*(\hat{\tau}_n)$. More precisely, the procedure works as follows:

1. Compute $\hat{\tau}_n$, set $B = B' = 99$, and the statistics τ_j^*, $j = 1, \ldots, B$.

2. Compute the estimated bootstrap P-value $\hat{P}^*(\hat{\tau}_n)$ based on the B bootstrap samples. Depending whether $\hat{P}^*(\hat{\tau}_n) < \alpha$ or $\hat{P}^*(\hat{\tau}_n) > \alpha$, test either the hypothesis that $P^*(\hat{\tau}_n) \geq \alpha$ or $P^*(\hat{\tau}_n) \leq \alpha$.[1] If the hypothesis is rejected, then stop, else go to step 3.

3. Set $B = 2B' + 1$. If B is too large, e.g. $B > 12,799$ then stop, else calculate τ_j^* for a further $B' + 1$ bootstrap samples, set $B' = B$ and return to step 2.

2.2 Data generating processes and estimation procedure

In the simulations the Data Generating Process (DGP) follows

1. either an ARCH(2) model

$$y_t = \sigma_t \varepsilon_t, \qquad \sigma_t^2 = b_0 + b_1 y_{t-1}^2 + b_2 y_{t-2}^2, \tag{11}$$

   with

$$b_0 = 0.1, \quad b_1 = 0.2, \quad b_2 = 0.1, \tag{12}$$

   or with (c.f. **DGP 5**)

$$b_0 = 0.1, \quad b_1 = 0.1, \quad b_2 = 0.4. \tag{13}$$

2. or a GARCH(1,1) model

$$y_t = \sigma_t \varepsilon_t, \qquad \sigma_t^2 = b_0 + b_1 y_{t-1}^2 + a_1 \sigma_{t-1}^2, \tag{14}$$

   with

$$b_0 = 0.1, \quad b_1 = 0.3, \quad a_1 = 0.3. \tag{15}$$

[1] This is done by using the Normal approximation of the Binomial distribution.

Remark 2.1 A sufficient condition for the stationarity of the process (11) and $Ey_t^4 < \infty$ is $(E\varepsilon_t^4)^{1/2}(b_1 + b_2) < 1$, see e.g. Giraitis, Kokoszka and Leipus (2000). For the parameters (12) we thus require that $E\varepsilon_t^4 < (10/3)^2 = 11.11$ (for standard normal ε_t, $E\varepsilon_t^4 = 3$). A analogous sufficient condition for the process (14) with standard normal errors ε_t is $a_1^2 + 2a_1b_1 + 3b_1^2 < 1$, see e.g. He and Teräsvirta (1999).

We consider the following Data Generating Processes. For each process we also report in Table 5 the approximate skewness and kurtosis which has been computed by generating a realization of length 10^6 and finding the ratio of the fourth sample moment over the square of the second sample moment.

DGP 0 The ε_t in (11) and (14) are iid standard normal. The models (11) and (14) are respectively denoted **DGP 0A** and **DGP 0B**.

DGP 1 Consider iid $\eta_t \sim \chi^2(k)$. Thus $\mathrm{E}(\eta_t) = k$, and $\mathrm{Var}(\eta_t) = 2k$. We set $k = 5$, and define the ε_t in (11) by $\varepsilon_t = (\eta_t - k)/\sqrt{2k}$. The distribution of the ε_t has mean zero, but is asymmetric. Since the empirical distribution of asset returns is skewed, see e.g. Singleton and Wingender (1986), Badrinath and Chatterjee (1988), Engle and González-Rivera (1991), this case may be of practical relevance; it can be expected that the skewness of the ε_t will be reflected in the distribution of the y_t. Perhaps a more appropriate approach to modeling the skewness of the returns is to use a model which transforms symmetric innovations into asymmetric observations, see e.g. the asymmetric power ARCH of Ding, Granger and Engle (1993) and the threshold GARCH of Zakoian (1994), but since in this study we use only standard GARCH models as archetypes, we experiment with skewed innovations.

DGP 2 The ε_t in (11) are independent but $\varepsilon_t \sim N(0,1)$ for the first half of the sample, and $\varepsilon_t \sim N(0,2)$ for the second half. This alternative corresponds to a change–point (a break) in the distribution of the ε_t.

DGP 3 The ε_t in (11) follow a linear ARCH(1), LARCH(1), process introduced by Giraitis, Robinson and Surgailis (2000), i.e.

$$\varepsilon_t = \zeta_t \xi_t, \qquad \zeta_t = a_0 + a_1 \varepsilon_{t-1}. \tag{16}$$

Thus y_t is the composition of an ARCH process and a LARCH process. For the LARCH(1) which forms the innovations sequence of the ARCH(2) defined by (11), we set $a_0 = 1.0$, $a_1 = 0.2$ and used standard normal innovations for the sequence $\{\xi_t\}$. This alternative corresponds to a remaining heteroskedastic structure in the innovations process. Under the null hypothesis $a_1 = 0$, we obtain the **DGP 0A**. We choose a LARCH(1) process for the nested process as taking a standard ARCH(1) process yields a process $\{y_t\}$ with a very large kurtosis for $a_1 \geq 0.2$. Further details are given in appendix on page 14. For this DGP, we use as benchmark the test for correct specification of the heteroskedastic functional form $Q(M)$ by Li and Mak (1994) which is asymptotically equivalent to the test by Lundbergh and Teräsvirta (2002) defined for the surimposition of standard ARCH/GARCH processes.

DGP 4 Consider iid $u_t \sim t(k)$, i.e. t distributed with k degrees of freedom. We set in (11) $\varepsilon_t = \sqrt{k/(k-2)}u_t$ with $k = 6$. Then the ε_t have a symmetric unit variance distribution which is leptokurtic, i.e. its tails are thicker than the ones of a normal distribution.

DGP 5 We also consider the case of misspecification of the functional form of the model: the DGP is still ARCH(2) (11) but with $b_0 = 0.1$, $b_1 = 0.1$, and $b_2 = 0.4$. The estimates $\hat{b}_j$ are however computed assuming that the observed series follows an ARCH(1) model. For this DGP, we also use as benchmark the test for correct specification of the conditional variance by Li and Mak (1994).

DGP 6 We consider a GARCH(1,1) DGP, with the assumption that the error terms are t distributed with 7 degrees of freedom. The parameters b_1 and a_1 are the same as for **DGP 0B**. This choice has been motivated by the fact that Bollerslev (1987) and Teräsvirta (1996) observed that the observed kurtosis of financial data is better fitted with $t(7)$ distributed error terms. These simulations results are of interest for our application on real data, as we will consider series of returns on Foreign Exchange (FX) rates at daily frequency.

DGP 7 We consider that the error terms in (11) are iid and follow a Laplace distribution Lap(σ), the density of which is $f(t) = (2\sigma)^{-1/2}\exp\{-|t|/\sigma\}$. For this DGP, we set $\sigma = 2^{-1/2}$, thus $\varepsilon_t \sim \text{Lap}(2^{-1/2})$, which is the two–sided exponential distribution with mean zero and variance equal to one.

DGP 8 We consider that the ε_t in (11) are independent but $\varepsilon_t \sim N(0,1)$ for the first half of the sample, and $\varepsilon_t \sim \text{Lap}(2^{-1/2})$ for the second half.

DGP 9 We consider that the ε_t in (11) are independent but $\varepsilon_t \sim N(0,1)$ for the first half of the sample, and $\varepsilon_t \sim t(7)$ for the second half.

DGP 10 We consider that the ε_t in (11) are independent but $\varepsilon_t \sim N(0,1)$ for the first half of the sample, and ε_t follow a centralized and standardized $\chi^2(5)$ for the second half.

The parameters $\hat{b}_j$ are estimated by the pseudo maximum likelihood (PML) method, i.e. the estimates are the parameter values which maximize the likelihood computed under the assumption that the innovations ε_t are independent and normally distributed. In that case, the log-likelihood function is:

$$\mathcal{L}(b;y) = -\frac{1}{2}\sum_{1\le t\le n}\left(\ln(2\pi) + \ln(\sigma_t^2) + \frac{y_t^2}{\sigma_t^2}\right), \tag{17}$$

where n denotes the number of observations. In this paper we consider $n = 200$ and $n = 400$.

As explained in Subsection 2.1, for a number of bootstraps $B = 1499$, an experiment for each DGP above involves 15 million PML estimation procedures. For each of them we check whether the Hessian matrix $\partial^2\mathcal{L}(b;y)/(\partial b\partial b^T)$ evaluated at the PML estimates $\hat{b}_j$ is positive definite, and we reject the estimation results if this condition is not satisfied. The estimates were obtained using the sequential QP algorithm from the NAG library, which allows to maximize a function subject to linear and nonlinear constraints, e.g. the constraint of fourth moment existence. Although the optimization procedure makes use of the analytic gradient of the log–likelihood function, see e.g. Bollerslev (1986), the whole set of simulations for the first version of this work required over 15000 hours of computing time on several computers, including two–processor machines. A recent experiment for a sample of 200 observations, carried out on a Pentium IV with a 2.2 GHz clock, required around 55 seconds for B equal to 1499: the testing procedure is thus very affordable on modern computers.

3 Presentation of the results with graphical methods

The availability of the bootstrap P-values allows us to present the simulation results using the graphical methods advocated by Davidson and MacKinnon (1998). For a given test, define the empirical distribution function (EDF) of the P-values $P_j, j = 1\ldots, N$ as

$$\hat{F}(x_i) \equiv \frac{1}{N}\sum_{j=1}^{N} \mathbf{I}[P_j \leq x_i], \tag{18}$$

where $\{x_i\}$ is a fine grid of points in $[0, 1]$. To compare the size, we display the P-value discrepancy plots, i.e. the plots of $\hat{F}(x_i) - x_i$ against x_i. Under the null hypothesis, the P-value is uniformly distributed on the interval $[0, 1]$, therefore the P-value discrepancy plots should be close to the 0^o degree line. We estimate $\hat{F}(\cdot)$ on a grid of 215 points, $x_i = .001, .002, \ldots, 0.010, .015, \ldots,$ $.990, .991, \ldots, .999$. Even with a large number of replications, e.g. 100,000, P-value discrepancy plots are jagged due to the experimental randomness. Therefore, when the number of replications is smaller because, as in our case, the statistics are computing time consuming, a smooth version of the P-value discrepancy plots is desirable. Denote by v_i the discrepancy, i.e. $v_i = F(x_i) - x_i$. A smoothed curve is obtained by regressing v_i on smooth function of x_i. Davidson and MacKinnon (1998) suggested the following regression:

$$v_i = \sum_{l=1}^{L} \gamma_l \sin(l\pi x_i) + u_i, \tag{19}$$

which can be written in matrix notation as

$$\boldsymbol{v = Z\gamma + u} \quad \text{with} \quad \boldsymbol{E(uu^T) =: \Omega},$$

where the vector $\boldsymbol{\gamma}$ is obtained by using the feasible Generalized Least Squares estimator. The truncation order L is selected using the Akaike Information Criterion, the log-likelihood function associated with the GLS estimation being:

$$-\frac{m}{2}\log(2\pi) + \frac{1}{2}\left|\boldsymbol{\Omega}^{-1}\right| - \frac{1}{2}(\mathbf{v} - \mathbf{Z}\boldsymbol{\gamma})^T\boldsymbol{\Omega}^{-1}(\mathbf{v} - \mathbf{Z}\boldsymbol{\gamma}).$$

Confidence bands are obtained by adding to the smoothed curve twice the square root of the diagonal elements of the covariance matrix $\mathbf{Z}(\mathbf{Z}^T\boldsymbol{\Omega}^{-1}\mathbf{Z})^{-1}\mathbf{Z}^T$. The exact form of $\boldsymbol{\Omega}^{-1}$ and other details are presented in Davidson and MacKinnon (1998).

Figure 1 p. 16 shows the P-value discrepancy plots and the smoothed P-value discrepancy plots for the three tests for the **DGP 0B** (The graphs for the **DGP 0A** are very similar). Since these plots are close to the 0^o line, which is inside the confidence bands, the bootstrap tests are seen to have correct size even for small sample sizes.

For representing the power of the test for each DGP, we use a graphical representation of the size–power function, denoted by $\eta(\alpha)$, which is the probability under that DGP that the test will reject the null hypothesis when the nominal rejection probability under the null hypothesis is α. For a powerful test, $\eta(\alpha)$ is much larger than α for small α's and so the curve $\eta(\alpha)$ is Γ shaped. For a test with small power the curve $\eta(\alpha)$ is close to the 45^o line.

The size–power curve graphs the power of a test against its nominal level. The EDF of the P-values, $\hat{F}(\alpha)$, which is an estimator of the cumulative distribution function of the P-values, is an approximation to $\eta(\alpha)$. However, it is more informative to plot $\hat{F}(\alpha)$ against the real size of the test. Thus the size–power trade-off curves in Figures 2-9 are the graphs of the EDF $\hat{F}(x)$ of the P-values under the alternative hypothesis against the EDF $\hat{F}(x)$ of the P-values under the null hypothesis. The null hypothesis is always **DGP 0A**, except for **DGP 6**, where the null hypothesis is **DGP 0B**.

Figures 2-9 plot the size–power functions for selected DGP's: **DGP 4, 5, 6, 8** with the 45^o line.[2] For all DGP's considered in our simulation study, the CVM test has more power than the NCVM test, which in turn is more powerful than the KS test. The only exception are **DGP 7** and **DGP 8**, (Laplace innovations) where the NCVM test is slightly more powerful than the CVM test.

For the sample sizes considered, the tests have a high power for the **DGP 1, 4, 6, 7, 8**, i.e. for the asymmetric and leptokurtic alternatives. The curves $\eta(\alpha)$ aproach a Γ shape when the sample size increases from 200 to 400 observations.

We also calculated the empirical size and power of the test for three levels: 10%, 5%, and 1%. Table 6 p 15 shows that even for small samples of 200 observations, the empirical size is very close to the nominal size.

The comparison with the test by Li and Mak (1994) is of interest: while the latter test has no power at all for the sample sizes and the parameter chosen for **DGP 3**, i.e. the power of the test is equal to its size, even when considering bootstrap tests fort this statistic, the portmanteau test is more powerful than our tests based on the EDF for **DGP 5**. For a sample of 200 observations, the empirical frequencies of rejection of the bootstrapped $Q(M)$ statistic for a test of size 5% with $M = 3, 4, 6$, are respectively equal to 86.90%, 84.10%, and 78.65%. For sample sizes of 400 observations, the frequencies of rejection increases to 99.14%, 98.82%, and 97.84%

A comparison with the tests for change–point developed in Kokoszka and Teyssière (2002), shows that the tests developed here have more power for detecting changes in the distribution of the innovations, i.e. **DGP 8, 9, 10**.

4 Application to Foreign Exchange rates returns series

We consider here an application of the three bootstrap tests to the series of squared residuals obtained from the estimation of a GARCH(1,1) on series of Foreign Exchange (FX) rates returns $R_t = 100 \times \log(P_t/P_{t-1})$, where P_t denotes the spot FX rate at time t. We consider the series of daily US Dollar/Deutschmark (USD–DEM) and US Dollar/British Pound (USD–GBP) FX rates returns from April 1979, i.e. after the inception of the European Monetary System. At first sight, this application is not relevant since empirical evidence has shown that the sums of the coefficients of a GARCH(1,1) are close to one, and thus violate the condition for stationarity of a GARCH(1,1) given above and in He and Teräsvirta (1999). However, Mikosch and Stărică (1999, 2002a, 2002b) have argued that this property can be due to the concatenation of several GARCH(1,1) processes, which can also be viewed as a non–homogeneous GARCH(1,1) process, i.e. with changing coefficients. Indeed, estimating a GARCH(1,1) model on shorter samples gives results which satisfy the stationarity condition.

We consider 2 samples of 1000 observations. The model specification is:

$$R_t = \mu + \sigma_t\varepsilon_t, \qquad \sigma_t^2 = b_0 + a_1\varepsilon_{t-1}^2 + b_1\sigma_{t-1}^2, \qquad \varepsilon_t \sim iidN(0,1). \tag{20}$$

[2] The full set of size–power curves is available upon request.

Tables 1 and 3 below display the estimated parameters of the GARCH(1,1) while tables 2 and 4 display the bootstrap CVM, NCVM and KS statistics $\hat{\tau}$, and the associated bootstrap critical values for a test of size 1%, 5%, and 10%. For all the considered series, the hypothesis of normality of the innovations is rejected, even for the size 1%.

Table 1: Estimated parameters USD–DEM. Heteroskedastic robust t-statistics are in parentheses

Series	b_0	b_1	a_1	μ
1	0.0148	0.7599	0.1844	0.0204
	(0.9596)	(7.2343)	(2.3757)	(0.9842)
2	0.0139	0.8537	0.1231	-0.0128
	(1.2151)	(15.2439)	(2.0036)	(-0.3785)

Table 2: Bootstrap test statistics and their critical values. USD–DEM

	Series 1			Series 2		
Stat	CVM	NCVM	KS	CVM	NCVM	KS
$\hat{\tau}$	17.2911	4.4413	3.7807	7.66543	2.1729	2.88938
1 %	0.9835	0.2392	1.3809	1.0497	0.2363	1.38856
5 %	0.6999	0.1561	1.1436	0.6720	0.1545	1.14692
10 %	0.5810	0.1219	1.0255	0.5666	0.1200	1.02789

Table 3: Estimated parameters USD–GBP. Heteroskedastic robust t-statistics are in parentheses

Series	b_0	b_1	a_1	μ
1	0.0172	0.8725	0.0965	-0.0391
	(1.4735)	(15.9346)	(1.9942)	(-1.0204)
2	0.0253	0.8942	0.04715	0.0430
	(1.4374)	(17.8774)	(1.7436)	(1.0703)

Table 4: Bootstrap test statistics and their critical values. USD–GBP

	Series 1			Series 2		
Stat	CVM	NCVM	KS	CVM	NCVM	KS
$\hat{\tau}$	4.1374	1.1634	2.2537	8.5529	2.1097	2.8452
1 %	0.9961	0.2449	1.3503	1.0470	0.2480	1.4074
5 %	0.6918	0.1563	1.1468	0.6888	0.1559	1.1505
10 %	0.5837	0.1249	1.0524	0.5496	0.1147	1.0213

5 Conclusions

We have examined several goodness–of–fit tests based on the residual empirical process of an ARCH sequence. Unlike for linear time series models, asymptotic tests cannot be readily constructed, but bootstrap tests are relatively easy to perform. The bootstrap tests we studied have correct size and can detect a misspecified probability distribution of the innovations (errors) in an ARCH model with high probability. They can detect an incorrect postulated order of the model with moderate success, and are practically unable to detect a structural break in the model parameters. For the detection of a misspecified distribution of the innovations, the tests have good power even for samples of size 200 observations. In situations where the tests are applicable, the CVM test performs better than the NCVM and KS tests, except when the error terms or part of the error terms follow a Laplace distribution. In that case, the NCVM test is slightly more powerful than the CVM test. An application to returns on FX rates shows that if a GARCH(1,1) model is postulated, normal errors should not be assumed.

6 Historical and mathematical appendix

In this section we provide some mathematical and historical background for the goodness–of–fit procedures considered in this paper.

Suppose we observe a sample $\varepsilon_1, \ldots, \varepsilon_n$ of iid observations with cumulative distribution function F which we assume for simplicity to be strictly increasing and absolutely continuous. Then the empirical process of the ε_i is

$$\alpha_n(x) = n^{-1/2} \sum_{i=1}^{n} \{\mathbf{I}[\varepsilon_i \leq x] - F(x)\}. \tag{21}$$

It is well known, see e.g. Shorack and Wellner (1986), that $\alpha_n(x)$ converges in $D(-\infty, \infty)$ to $B^0(F(x))$, where B^0 is a Brownian bridge. The latter convergence is to be interpreted as the convergence of $\alpha_n(F^{-1}(\cdot))$ to $B^0(\cdot)$ in $D[0,1]$. Relation (21) is used to construct classical goodness–of–fit tests. For example, a variant of the Cramér–von Mises test (NCVM in our notation) is based on the convergence

$$\int_{-\infty}^{\infty} |\alpha_n(x)|^2 F(dx) \xrightarrow{d} \int_{-\infty}^{\infty} \left|B^0(F(x))\right|^2 F(dx) \stackrel{d}{=} \int_0^1 \left|B^0(u)\right|^2 du, \tag{22}$$

the distribution of the last integral being tabulated.

In time series models, the innovations (noise sequence) are not observable and we have to work with residuals. To illustrate the idea, suppose we observe a sample $x_1, \ldots, x_n$ from an autoregressive model of order p which, in a sense, is a linear analogue of the ARCH(p) model. Thus, we assume that the observations satisfy

$$x_t = \sum_{j=1}^{p} \phi_j x_{t-j} + \varepsilon_t, \tag{23}$$

where the ε_t are as above and, in addition, have mean zero and unit variance. Suppose we estimate the coefficients ϕ_j using estimators such that for each j, $n^{1/2}(\hat{\phi}_j - \phi_j)$ is bounded in probability (all reasonable estimators satisfy this condition) and construct the residuals

$$\hat{\varepsilon}_t = x_t - \sum_{j=1}^{p} \hat{\phi}_j x_{t-j}, \quad t = p+1, \ldots, n. \tag{24}$$

Then the empirical process of these residuals, defined analogously to (21) with the ε_i replaced by the $\hat{\varepsilon}_i$, still converges to $B^0(F(\cdot))$. This property holds for more general linear models and under weaker assumptions on the ε_t, see Chapter 7 of Koul (1992) and references therein.

The situation is different for ARCH models. It follows from Boldin (1998) that the difference between the empirical process of the residuals from an ARCH(1) and a transformed Brownian bridge tends to a normal random variable. Horváth, Kokoszka and Teyssière (2001) considered a general ARCH(p) models and squared residuals and showed the process (3) converges in $D[0, \infty)$ to $B^0(G(t)) + tg(t)\xi$, where G and g are, respectively, the cumulative distribution function and the density of ε_0^2 and ξ is a normal random variable which is correlated in an intricate way with the process $B^0(G(\cdot))$. Even though the joint covariance structure is known, it does not lead to a simple convergence like (22) and so, unlike for linear models, asymptotic tests cannot be readily constructed, but the existence of a limiting distribution suggests the applicability of bootstrap tests. For extensions to GARCH(p, q) processes and more general ARCH processes we refer, respectively, to Berkes and Horváth (2002) and Section 8.3 of Koul (2002).

References

Andrews, D. W. K. and M. Buchinsky (2000). A Three-Step Method for Choosing the Number of Bootstrap Repetitions. *Econometrica*, **68,** 23–51.

Badrinath, S. G. and S. Chatterjee (1988). On Measuring Skewness and Elongation in Common Stock Returns Distributions: The Case of Market Index. *Journal of Business*, **61,** 451–472.

Berkes, I. and L. Horváth (2002). Limit Results for the Empirical Process of Squared Residuals in GARCH Models. *Preprint.*

Berkes, I., Horváth, L. and P. S. Kokoszka (2002a). Asymptotics for GARCH Squared Residuals Correlations. *Econometric Theory*, forthcoming. `http://math.usu.edu/~piotr/research.html`

Berkes, I., Horváth, L. and P. S. Kokoszka (2002b). A Weighted χ^2–Test for GARCH(1,1) Squared Residual Correlations. *Preprint.* `http://math.usu.edu/~piotr/research.html`

Boldin, M. V. (1998). On Residual Empirical Distribution Functions in ARCH Models with Applications to Testing and Estimation. *Mitt. Math. Sem. Giessen*, **235,** 49–66.

Bollerslev, T. (1986). Generalized Autoregressive Conditional Heteroskedasticity. *Journal of Econometrics*, **31,** 307–327.

Bollerslev, T. (1987). A Conditional Heteroskedastic Time Series Model for Speculative Prices and Rates of Return. *Review of Economics and Statistics*, **69,** 542–547.

Davidson, R. and J. G. MacKinnon (1998). Graphical Methods for Investigating the Size and Power of Hypothesis Tests. *The Manchester School*, **66,** 1–26.

Davidson, R. and J. G. MacKinnon (1999). The Size Distortion of Bootstrap Tests. *Econometric Theory*, **15,** 361–376.

Davidson, R. and J. G. MacKinnon (2000). Bootstrap Tests: How Many Bootstraps? *Econometric Reviews*, **19,** 55–68.

Davis, R. A. and T. Mikosch (1998). Limit Theory for the Sample Autocorrelation of Stationary Processes with Heavy Tails with Applications to ARCH. *The Annals of Statistics*, **26,** 2049–2080.

Ding, Z., Granger, C. W. J. and R. F. Engle (1993). A Long Memory Property of Stock Market Returns and a New Model. *Journal of Empirical Finance*, **1,** 83–106.

Dufour, J. M. and L. Khalaf (2001). Monte Carlo Tests in Econometrics. In B. Baltagi editor, *Companion to Theoretical Econometrics*, 494–519. Blackwell, Oxford.

Dufour, J. M. and J. F. Kiviet (1998). Exact Inference Methods for First-Order Autoregressive Distributed Lag Models. *Econometrica*, **66,** 79–104.

Engle, R. F. and G. González-Rivera (1991). Semiparametric ARCH models. *Journal of Business and Economic Statistics*, **9,** 345–359.

Giraitis, L., Kokoszka, P. S. and R. Leipus (2000). Stationary ARCH Models: Dependence Structure and Central Limit Theorem. *Econometric Theory*, **16,** 3–22.

Giraitis, L., Robinson, P. M. and D. Surgailis (2000). A Model for Long-Memory Conditional Heteroskedasticity. *Annals of Applied Probability*, **10,** 1002–1024.

He, C. and T. Teräsvirta (1999). Fourth Moment Structure of the GARCH(p, q) Process. *Econometric Theory*, **15,** 824–846.

Horváth, L. and P. S. Kokoszka (2001). Large Sample Distribution of ARCH(p) Squared Residual Correlations. *Econometric Theory*, **17,** 283–295.

Horváth, L., Kokoszka, P. S. and G. Teyssière (2001). Empirical Process of the Squared Residuals of an ARCH Sequence. *The Annals of Statistics*, **29,** 445–469.

Kokoszka, P. S. and G. Teyssière (2002). Change-Point Detection in GARCH Models: Asymptotic and Bootstrap Tests. *Preprint.* `http://math.usu.edu/~piotr/research.html`

Koul, H. L. (1992). *Weighted Empiricals and Linear Models.* IMS, Hayward, California.

Koul, H. L. (2002). *Weighted Empiricals Processes in Dynamic Nonlinear Models.* Lecture Notes in Statistics, Springer Verlag.

Li, W. K. and T. K. Mak (1994). On the Squared Residual Autocorrelations in Non–linear Time Series with Conditional Heteroskedasticity. *Journal of Time Series Analysis*, **15,** 627–636.

Lundbergh, S. and T. Teräsvirta (2002). Evaluating GARCH Models. *Journal of Econometrics*, **110,** 417–435.

Marsaglia, G. (1996). DIEHARD: A Battery of Tests of Randomness.
`http://stat.fsu.edu/pub/diehard`.

Mikosch, T. and C. Stărică (1999). Change of Structure in Financial Time Series, Long Range Dependence and the GARCH Model. *Preprint.* `http://www.cs.nl/~eke/iwi/preprints`.

Mikosch, T. and C. Stărică (2000). Limit Theory for the Sample Autocorrelation and Extremes in GARCH(1,1) Process. *The Annals of Statistics*, **28**, 1427–1451.

Mikosch, T. and C. Stărică (2002a). Long–Range Dependence Effects and ARCH Modelling. In M. S. Taqqu, G. Oppenheim and P. Doukhan editors, *Long-Range Dependence: Theory and Applications*, 439–459. Birkhäuser, Boston.

Mikosch, T. and C. Stărică (2002b). Non–Stationarities in Financial Time Series: The Long–Range Dependence and the IGARCH Effects. *Preprint.*

Shorack, G. R. and J. A. Wellner (1986). *Empirical Processes with Applications to Statistics.* Wiley, New York.

Singleton, J. C. and J. Wingender (1986). Skewness Persistence in Common Stock Returns. *Journal of Financial and Quantitative Analysis*, **21,** 335–341.

Teräsvirta, T. (1996). Two Stylized Facts and the GARCH(1,1) Model. *Stockholm School of Economics, Working Paper Series in Economics and Finance* **96**.

Zakoian, J. M. (1994). Threshold Heteroskedastic Models. *Journal of Economics Dynamic and Control*, **18,** 931–955.

A Tables

- **DGP 0A**: ARCH(2) with $N(0,1)$ errors, i.e. the null hypothesis,
- **DGP 0B**: GARCH(1,1) with $N(0,1)$ errors, i.e. the null hypothesis,
- **DGP 1**: ARCH(2) with centralized and standardized $\chi^2(5)$ errors,
- **DGP 2**: ARCH(2) with $N(0,1)$ errors for the first half of observations, and $N(0,2)$ errors for the second half,
- **DGP 3**: surimposition of two ARCH processes, i.e. a LARCH(1) nested in an ARCH(2),
- **DGP 4**: ARCH(2) with standardized $t(6)$ errors,
- **DGP 5**: ARCH(1) model estimated on an ARCH(2) DGP,
- **DGP 6**: GARCH(1,1) with standardized $t(7)$ errors,
- **DGP 7**: ARCH(2) with Lap($2^{-1/2}$) errors,
- **DGP 8**: ARCH(2) with $N(0,1)$ errors for the first half of observations, and Lap($2^{-1/2}$) errors for the second half,
- **DGP 9**: ARCH(2) with $N(0,1)$ errors for the first half of observations, and standardized $t(7)$ errors for the second half,
- **DGP 10**: ARCH(2) with $N(0,1)$ errors for the first half of observations, and centralized and standardized $\chi^2(5)$ errors for the second half.

Table 5: Empirical skewness and kurtosis for the different DGP.

DGP	Skewness	Kurtosis
DGP 0A	0.0004	3.4111
DGP 0B	0.0005	4.0630
DGP 1	1.3694	6.8832
DGP 2	0.0015	8.1750
DGP 3	0.0108	7.2059
DGP 4	-0.0227	7.2830
DGP 5	-0.0027	5.0889
DGP 6	0.0369	6.9694
DGP 7	-0.0073	8.8038
DGP 8	0.0002	6.0825
DGP 9	-0.0185	5.9228
DGP 10	0.6812	5.2007

We have also considered for **DGP 3** the case of an ARCH(1) nested in an ARCH(2), i.e.

$$\varepsilon_t = \zeta_t \xi_t, \qquad \zeta_t^2 = a_0 + a_1 \varepsilon_{t-1}^2 \tag{25}$$

When $a_1 = 0.1$, the kurtosis of the process computed on a sample of 10^6 observation is equal to 6.6745, but increases to 4012.12 for $a_1 = 0.2$. For $a_1 \geq 0.3$, the kurtosis is huge. Thus, we stick to the case of a LARCH(1) nested in an ARCH(2) as for $a_1 = 0.2$ the kurtosis is equal to 7.20, and becomes huge for $a_1 \geq 0.4$.

Table 6: Empirical frequencies of rejection (200 observations).

	CVM			NCVM			KS		
DGP	10 %	5 %	1 %	10 %	5 %	1 %	10 %	5 %	1 %
DGP 0A	0.1020	0.0508	0.0108	0.0995	0.0505	0.0110	0.0984	0.0501	0.0101
DGP 0B	0.1032	0.0518	0.0113	0.1021	0.0518	0.0107	0.1015	0.0505	0.0105
DGP 1	0.8590	0.7698	0.5417	0.5577	0.4216	0.1959	0.4530	0.3119	0.1277
DGP 2	0.1320	0.0764	0.0180	0.1265	0.0647	0.0133	0.1028	0.0502	0.0123
DGP 3	0.1868	0.1109	0.0383	0.1790	0.1086	0.0349	0.1431	0.0788	0.0221
DGP 4	0.7227	0.6323	0.4417	0.6907	0.5889	0.3602	0.5590	0.4330	0.2224
DGP 5	0.3897	0.3139	0.1909	0.3442	0.2587	0.1459	0.2626	0.1824	0.0849
DGP 6	0.5826	0.4762	0.2893	0.5422	0.4416	0.2548	0.4173	0.3108	0.1503
DGP 7	0.9897	0.9814	0.9453	0.9918	0.9841	0.9526	0.9840	0.9682	0.9063
DGP 8	0.6580	0.5616	0.3655	0.6821	0.5847	0.3849	0.6096	0.4916	0.2886
DGP 9	0.2564	0.1765	0.0812	0.2383	0.1578	0.0669	0.1731	0.1030	0.0376
DGP 10	0.3987	0.2850	0.1365	0.2403	0.1496	0.0530	0.1733	0.1000	0.0311

Table 7: Empirical frequencies of rejection (400 observations).

	CVM			NCVM			KS		
DGP	10 %	5 %	1 %	10 %	5 %	1 %	10 %	5 %	1 %
DGP 0A	0.1019	0.0526	0.0104	0.1017	0.0540	0.0104	0.1015	0.0519	0.0133
DGP 0B	0.1015	0.0487	0.0103	0.0984	0.0494	0.0105	0.0987	0.0484	0.0114
DGP 1	0.9915	0.9801	0.9280	0.8285	0.7157	0.4641	0.8104	0.6820	0.4110
DGP 2	0.2006	0.1175	0.0374	0.1540	0.0870	0.0258	0.1262	0.0658	0.0191
DGP 3	0.3171	0.2241	0.0979	0.3183	0.2153	0.0948	0.2421	0.1572	0.0591
DGP 4	0.9323	0.8970	0.7979	0.9172	0.8720	0.7521	0.8371	0.7598	0.5665
DGP 5	0.6009	0.5181	0.3637	0.5414	0.4535	0.3095	0.4303	0.3348	0.1942
DGP 6	0.8497	0.7869	0.6287	0.8215	0.7460	0.5695	0.7030	0.5992	0.3819
DGP 7	0.9999	0.9998	0.9996	1.0000	1.0000	0.9998	0.9999	0.9999	0.9993
DGP 8	0.9255	0.8757	0.7398	0.9381	0.8962	0.7661	0.8976	0.8371	0.6592
DGP 9	0.5363	0.4366	0.2700	0.5031	0.4002	0.2355	0.3932	0.2834	0.1348
DGP 10	0.5604	0.4379	0.2364	0.3018	0.1916	0.0781	0.2212	0.1257	0.0382

B P-value discrepancy plots

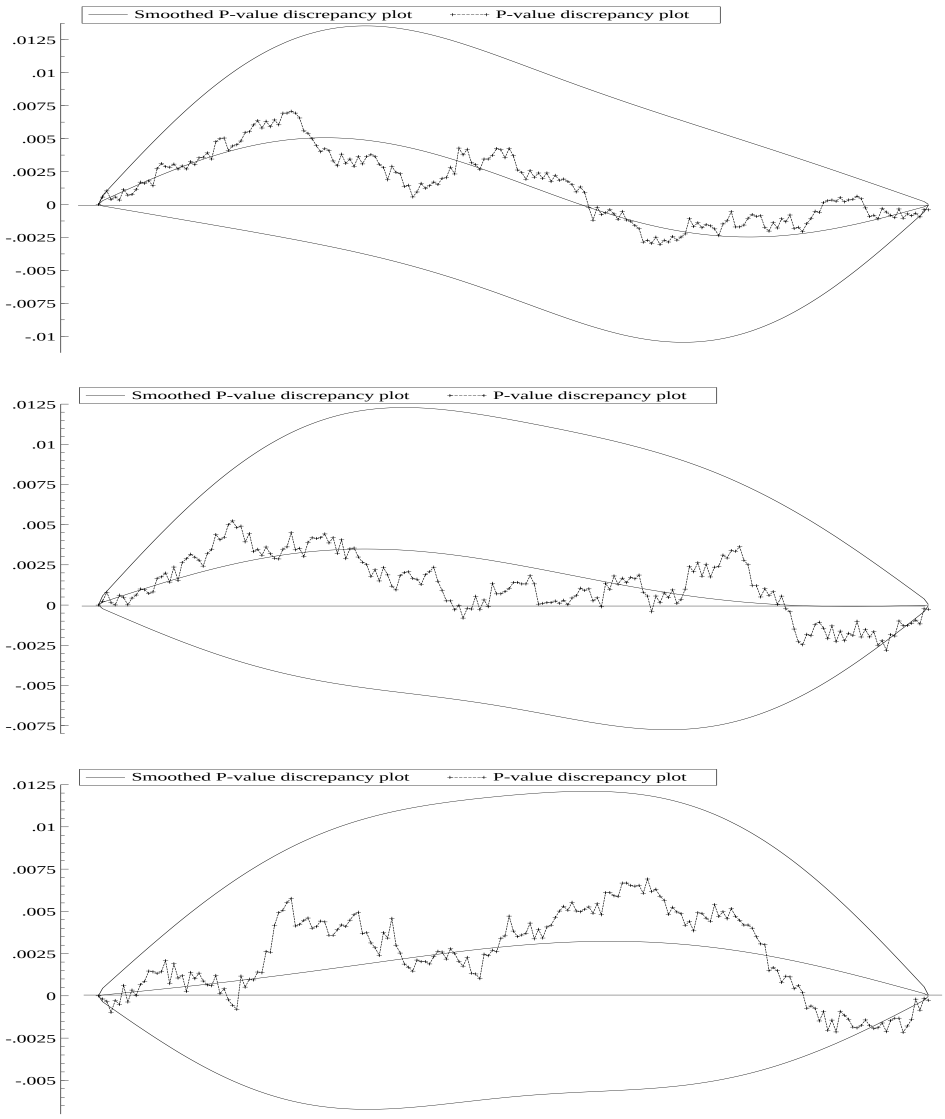

Figure 1: P-value discrepancy plots and their smoothed versions with confidence bands. Sample size is 200. Top - CVM, middle - NCVM, bottom - KS test

C Size–power curves

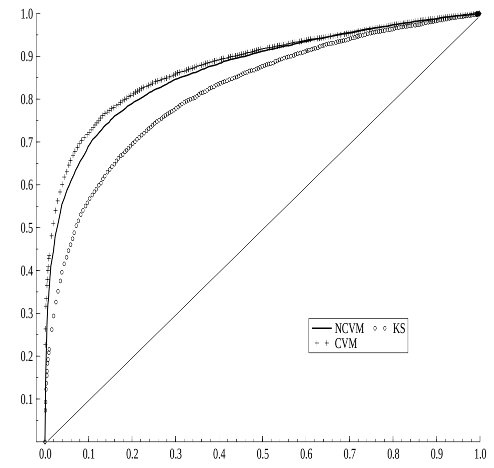

Figure 2: DGP 4 (leptokurtic errors), 200 observations.

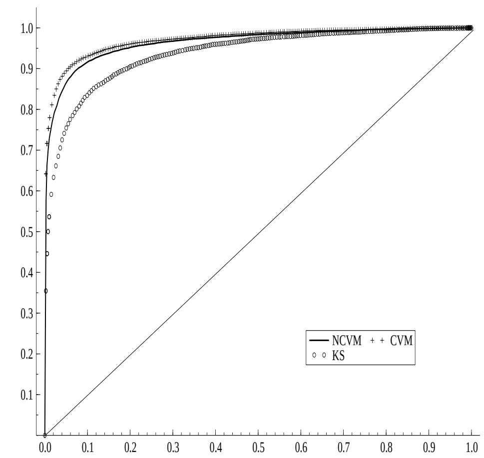

Figure 3: DGP 4 (leptokurtic errors), 400 observations.

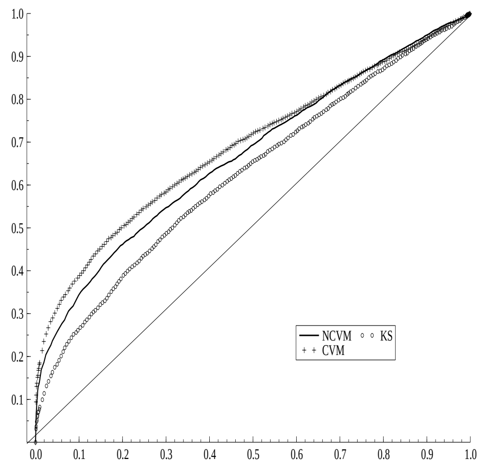

Figure 4: DGP 5 (incorrect order), 200 observations.

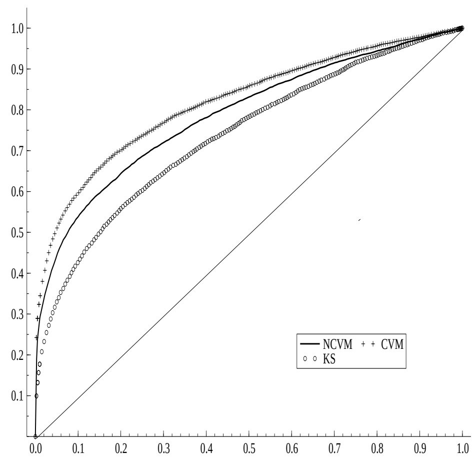

Figure 5: DGP 5 (incorrect order), 400 observations.

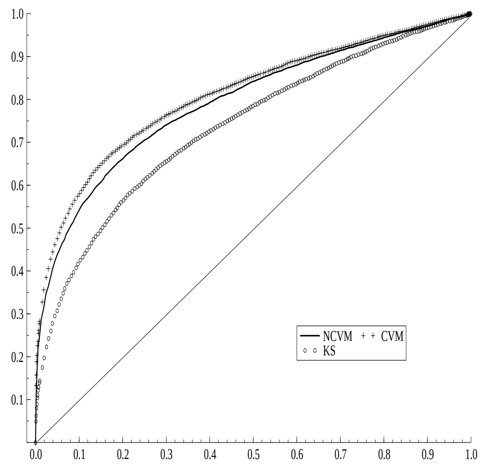

Figure 6: DGP 6 (GARCH(1,1) with $t(7)$ innovations), 200 observations.

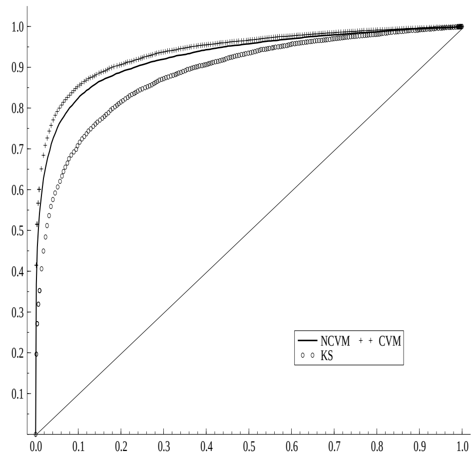

Figure 7: DGP 6 (GARCH(1,1) with $t(7)$ innovations), 400 observations.

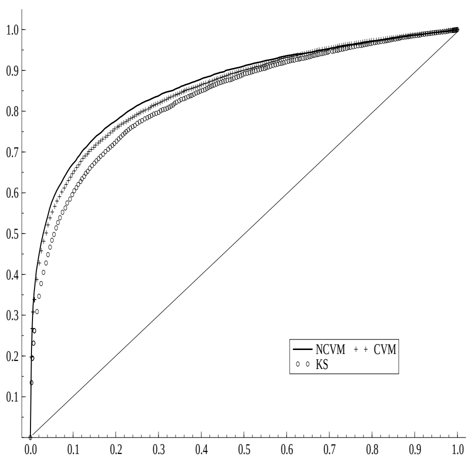

Figure 8: DGP 8 (change point in the innovations), 200 observations.

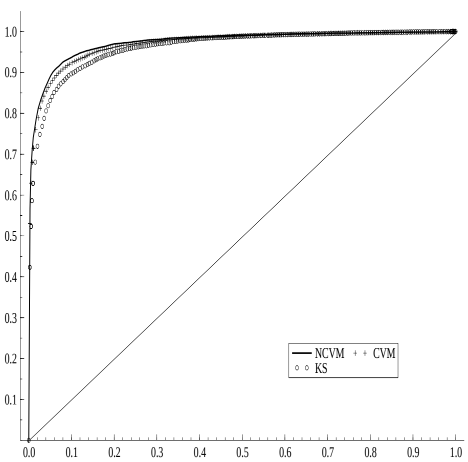

Figure 9: DGP 8 (change point in the innovations), 400 observations.