Econometrics Journal (2008), volume **11**, pp. 27–38.
doi: 10.1111/j.1368-423X.2008.00229.x

Estimating GARCH models: when to use what?

DA HUANG†, HANSHENG WANG† AND QIWEI YAO‡,†

†*Guanghua School of Management, Peking University, Beijing 100871, China*
E-mail: huangda@gsm.pku.edu.cn, hansheng@gsm.pku.edu.cn

‡*Department of Statistics, London School of Economics, London, WC2A 2AE, UK*
E-mail: q.yao@lse.ac.uk

First version received: January 2007; final version accepted: November 2007

Summary The class of generalized autoregressive conditional heteroscedastic (GARCH) models has proved particularly valuable in modelling time series with time varying volatility. These include financial data, which can be particularly heavy tailed. It is well understood now that the tail heaviness of the innovation distribution plays an important role in determining the relative performance of the two competing estimation methods, namely the maximum quasi-likelihood estimator based on a Gaussian likelihood (GMLE) and the log-transform-based least absolutely deviations estimator (LADE) (see Peng and Yao 2003 *Biometrika, 90*, 967–75). A practically relevant question is when to use what. We provide in this paper a solution to this question. By interpreting the LADE as a version of the maximum quasilikelihood estimator under the likelihood derived from assuming hypothetically that the log-squared innovations obey a Laplace distribution, we outline a selection procedure based on some goodness-of-fit type statistics. The methods are illustrated with both simulated and real data sets. Although we deal with the estimation for GARCH models only, the basic idea may be applied to address the estimation procedure selection problem in a general regression setting.

Keywords: *Estimation procedure selection*, *GARCH*, *Gaussian likelihood*, *Heavy tail*, *Laplace distribution*, *Least absolute deviations estimator*, *Maximum quasilikelihood estimator*, *Time series*.

1. INTRODUCTION

Several methods exist for estimating parameters in generalized autoregressive conditional heteroscedastic (GARCH) models with unknown innovation distributions. The maximum quasilikelihood estimator facilitated by hypothetically assuming the innovation distribution to be Gaussian is arguably the most frequently used estimator in practice, which we simply call the Gaussian maximum-likelihood estimator (GMLE). The asymptotic properties of the GMLE is fully understood now. In fact, it is a well-behaved estimator when the innovation distribution has finite fourth moment. However, when the innovation distribution is heavy tailed with an infinite fourth moment, the estimators may not be asymptotically normal, the range of possible limit distributions is extraordinarily large and the convergence rate is slower than the standard rate of $n^{1/2}$ (see e.g. Hall and Yao, 2003). To overcome the drawbacks due to the possible slow convergence rates of the GMLE, Peng and Yao (2003) propose a log-transform-based least absolute deviations estimator (LADE) as an alternative which is robust with respect to the heavy tails of the innovation distribution. In fact, the LADE is asymptotically normal with the standard convergence rate $n^{1/2}$ under the assumption that the second moment of the innovation distribution is finite. Monte Carlo experiments reported in Peng and Yao (2003) indicate that the relative performance of the two estimators hinges critically on the tail heaviness of the innovation distribution. Indeed, LADE is preferred for the processes with very heavy tailed innovation distributions.

In practice, the innovation distribution is unknown. A practically relevant question is how to choose an appropriate estimator for a given practical situation. In this paper, we put forward a proposal to choose between the GMLE and the LADE based on some goodness-of-fit measures. To this end, we view the LADE also as a maximum quasilikelihood estimator based on the hypothesis that the log-squared innovations follow a Laplace distribution. Our approach is based on the intuition that we should use the GMLE if the innovation distribution is close to a normal distribution, and use the LADE if the distribution of the log squared innovations is close to a Laplace distribution. Some goodness-of-fit statistics are defined to measure the closeness of those distributions (see Section 2.3). We have shown that our selection procedure is consistent in the sense that the probability of choosing the 'correct' estimator converges to 1. The numerical experiments illustrate that the proposed procedure exhibits desirable finite sample performance. Although we only deal with the estimation for GARCH models in this paper, the general idea may be applied for selecting, for example, between L_1 and L_2 estimator in a general regression setting (see the relevant discussion in Section 4).

The asymptotic properties of the GMLE have been studied initially by Weiss (1986) for pure ARCH(p) processes and by Lee and Hansen (1994) and Lumsdaine (1996) for GARCH(1,1) processes, under the assumption that the innovation distribution has finite fourth moment. Further studies for general GARCH(p, q) processes without the condition of fourth finite moment may be found in Hall and Yao (2003), Berkes et al. (2003), Straumann and Mikosch (2006) and Mikosch and Straumann (2006). See also Straumann (2005). Complex asymptotic properties were also observed from a Whittle estimator by Giraitis and Robinson (2001) for heavy tailed GARCH(1,1) models. The asymptotic properties of L_p-estimators for ARCH(p) models were established by Horváth and Liese (2004).

The rest of the paper is organized as follows. The methodology is presented in Section 2. It also contains a consistency result. Section 3 reports numerical illustrations with both simulated and real data sets. Miscellaneous remarks are given in Section 4. The technical proof is relegated to the Appendix.

2. METHODOLOGY

2.1. Model

A generalized autoregressive conditional heteroscedastic, GARCH, model with orders $p \geq 1$ and $q \geq 0$ is defined as

$$X_t = \sigma_t \varepsilon_t, \quad \text{and} \quad \sigma_t^2 \equiv \sigma_t(\theta)^2 = c + \sum_{i=1}^{p} b_i X_{t-i}^2 + \sum_{j=1}^{q} a_j \sigma_{t-j}^2, \tag{2.1}$$

where $c > 0, b_j \geq 0$ and $a_j \geq 0$ are unknown parameters, $\theta = (c, b_1, \ldots, b_p, a_1, \ldots, a_q)^{\mathrm{T}}$, $\{\varepsilon_t\}$ is a sequence of independent and identically distributed random variables with mean 0 and variance

1, and ε_t is independent of $\{X_{t-k}, k \geq 1\}$ for all t. The distribution of ε_t is unknown. When $q = 0$, (2.1) reduces to an autoregressive conditional heteroscedastic, ARCH, model. The necessary and sufficient condition for (2.1) to define a unique strictly stationary process $\{X_t, t = 0, \pm 1, \pm 2, \cdots\}$ with $EX_t^2 < \infty$ is that

$$\sum_{i=1}^{p} b_i + \sum_{j=1}^{q} a_j < 1. \tag{2.2}$$

Furthermore, for such a stationary solution, $EX_t = 0$ and $\mathrm{var}(X_t) = c/(1 - \sum_{i=1}^{p} b_i - \sum_{j=1}^{q} a_j)$ (see Giraitis et al., 2000 and also theorem 4.4 of Fan and Yao, 2003). Under condition (2.2), $\sigma_t^2 = \sigma_t(\theta)^2$ may be expressed as

$$\sigma_t(\theta)^2 = \frac{c}{1 - \sum_{j=1}^{q} a_j} + \sum_{i=1}^{p} b_i X_{t-i}^2 + \sum_{i=1}^{p} b_i \sum_{k=1}^{\infty} \sum_{j_1=1}^{q} \cdots \sum_{j_k=1}^{q} a_{j_1} \cdots a_{j_k} X_{t-i-j_1-\cdots-j_k}^2, \tag{2.3}$$

where the multiple sum vanishes if $q = 0$ (see Hall and Yao, 2003).

2.2. *Two estimators*

The GMLE is defined as

$$\widehat{\theta} = \arg\min_{\theta} \sum_{t=\nu+1}^{n} \left[\frac{X_t^2}{\widetilde{\sigma}_t(\theta)^2} + \log\left\{\widetilde{\sigma}_t(\theta)^2\right\} \right], \tag{2.4}$$

where $\widetilde{\sigma}_t(\theta)^2$ is a truncated version of $\sigma_t(\theta)^2$ defined as

$$\widetilde{\sigma}_t(\theta)^2 = \frac{c}{1 - \sum_{j=1}^{q} a_j} + \sum_{i=1}^{\min(p,t-1)} b_i X_{t-i}^2 + \sum_{i=1}^{p} b_i \sum_{k=1}^{\infty} \sum_{j_1=1}^{q} \cdots \sum_{j_k=1}^{q} a_{j_1} \cdots a_{j_k} \tag{2.5}$$

$$\times\, X_{t-i-j_1-\cdots-j_k}^2 I(t - i - j_1 - \cdots - j_k \geq 1),$$

which depends on the observations $X_{t-1}, \ldots, X_1$ only; cf. (2.3), and $\nu \geq 1$ is an integer which controls the effect of the truncation. Note for a purely ARCH model (i.e. $q = 0$), we choose $\nu = p$. The GMLE can be motivated by temporarily assuming that $\varepsilon_t \sim N(0, 1)$. Given $\{X_k, k \leq \nu\}$ with $\nu \geq \max(p, q)$, the conditional density function of $X_{\nu+1}, \ldots, X_n$ is then proportional to

$$\left\{ \prod_{t=\nu+1}^{n} \sigma_t(\theta)^2 \right\}^{-1/2} \exp\left\{ -\frac{1}{2} \sum_{t=\nu+1}^{n} \frac{X_t^2}{\sigma_t(\theta)^2} \right\}.$$

Maximizing this (conditional) likelihood with $\sigma_t(\theta)^2$ replaced by $\widetilde{\sigma}_t(\theta)^2$ leads to the GMLE estimator $\hat{\theta}$ (see 2.4).

The LADE, proposed by Peng and Yao (2003), requires a different parametrization as follows. Let $C_0 > 0$ be a constant such that the median of e_t^2 is equal to 1, where $e_t = C_0^{1/2}\,\varepsilon_t$. Then (2.1) may now be expressed as

$$X_t = s_t e_t, \quad \text{and} \quad s_t^2 \equiv s_t(\alpha)^2 = \gamma + \sum_{i=1}^{p} \beta_i X_{t-i}^2 + \sum_{j=1}^{q} a_j s_{t-j}^2, \tag{2.6}$$

where $s_t^2 = \sigma_t^2/C_0$, $\gamma = c/C_0$, $\beta_i = b_i/C_0$ and $\alpha = (\gamma, \beta_1, \ldots, \beta_p, a_1, \ldots, a_q)^{\mathrm{T}}$. Note that now

$$\log\big(X_t^2\big) = \log\big\{s_t(\alpha)^2\big\} + \log\big(e_t^2\big), \tag{2.7}$$

and the median of $\log(e_t^2)$ is 0. Thus, the true value of α minimizes $E|\log(X_t^2) - \log\{s_t(\alpha)^2\}|$. This motivates the LADE

$$\widehat{\alpha} = \arg\min_{\alpha} \sum_{t=\nu+1}^{n} \big|\log\big(X_t^2\big) - \log\big\{\widetilde{s}_t(\alpha)^2\big\}\big|, \tag{2.8}$$

where $\widetilde{s}_t(\alpha)^2$ is a truncated version of $s_t(\alpha)^2$ defined as

$$\widetilde{s}_t(\alpha)^2 = \frac{\gamma}{1 - \sum_{j=1}^{q} a_j} + \sum_{i=1}^{\min(p,t-1)} \beta_i X_{t-i}^2 + \sum_{i=1}^{p} \beta_i \sum_{k=1}^{\infty} \sum_{j_1=1}^{q} \cdots \sum_{j_k=1}^{q} a_{j_1} \cdots a_{j_k} \tag{2.9}$$

$$\times\, X_{t-i-j_1-\cdots-j_k}^2 I(t - i - j_1 - \cdots - j_k \geq 1),$$

which directly follows from (2.5).

In fact, the LADE may also be viewed as a maximum quasilikelihood estimator by temporarily assuming $\log(e_t^2)$ having a Laplace distribution with density to 0.5 $\lambda \exp(-\lambda|x|)$, where $\lambda > 0$ is a constant. By (2.7), the (conditional) likelihood function based on the observations $X_{\nu+1}, \ldots,$ X_n (given $\{X_k, k \leq \nu\}$) is then proportional to

$$\exp\left[-\lambda \sum_{t=\nu+1}^{n} \left|\log\left(X_t^2\right) - \log\left\{s_t(\alpha)^2\right\}\right|\right].$$

Maximizing this with $s_t(\alpha)^2$ replaced by $\widetilde{s}_t(\alpha)^2$ leads to the LADE $\widehat{\alpha}$ (see 2.8). Note $E(\varepsilon_t^2) < \infty$ if $\log(e_t^2)$ has the above Laplace distribution with $\lambda < 1$.

2.3. Selecting an estimation procedure

The performance of $\widehat{\theta}$ and $\widehat{\alpha}$ hinges critically on the tail heaviness of the innovation distribution. When $E(|\varepsilon_t|^{4-\delta}) < \infty$ for any $\delta > 0$, $\widehat{\theta}$ is asymptotically normal. Furthermore, the convergence rate is the standard $n^{1/2}$ provided $E(\varepsilon_t^4) < \infty$. When ε_t is heavy tailed in the sense that $E(|\varepsilon_t|^d) = \infty$ for some $2 < d < 4$, the asymptotic distribution of $\widehat{\theta}$ is no longer normal with a convergence rate slower than $n^{1/2}$, and it depends on infinite many unknown parameters of the underlying distribution. Those asymptotic results have been established under different settings by, for example, Lee and Hansen (1994), Lumsdaine (1996), Hall and Yao (2003), Berkes et al. (2003), Straumann and Mikosch (2006) and Mikosch and Straumann (2006). On the other hand, the LADE $\widehat{\alpha}$ is always asymptotically normal with the convergence rate $n^{1/2}$ provided $E(\varepsilon_t^2) < \infty$. Simulation studies also indicate that the finite sample performance of the LADE is better than that of the GMLE when, for example, $E(|\varepsilon_t|^3) = \infty$ (see Peng and Yao, 2003).

Since the distribution of ε_t is unknown in practice, it is rather difficult, if not impossible, to inference on how many moments ε_t has. A pertinent question is which estimator, between the GMLE and the LADE, we should use in practice. We provide an answer to this question below.

If we knew the distribution of innovations ε_t, the genuine maximum (conditional) likelihood estimator would be used. Intuitively, we would expect that the GMLE is a better option when the distribution of ε_t is close to $N(0, 1)$, and the LADE is better when the distribution of $\log(e_t^2)$ is approximately a Laplace distribution (see the discussion at the end of Section 2.2). This suggests that we may compare the closeness of those two pair distributions to select a good estimation procedure.

Denoted by $\Phi(\cdot)$ the $N(0, 1)$ distribution function, and by $G(\cdot)$ the distribution function with the density function $0.25 \exp(-|x|/2)$. Let $\widehat{\varepsilon}_t = X_t/\widetilde{\sigma}_t(\widehat{\theta})$ be the residuals derived from the GMLE. In practice, we standardize $\widehat{\varepsilon}_t$ such that the first two sample moments are 0 and 1. Let $\widehat{e}_t = X_t/\widetilde{s}_t(\widehat{\alpha})$ be the residuals derived from the LADE. In practice, we 'standardize' $\widehat{e}_t$ such that the sample median of $\widehat{e}_t^2$ is 1 and the sample mean of $|\log(\widehat{e}_t^2)|$ is 2. This may be achieved by letting $\log(\widehat{e}_t^2) = C_1 \log\{C_2 X_t^2/\widetilde{s}_t(\widehat{\alpha})^2\}$ for appropriate positive constants C_1 and C_2. Note that $\Phi(\varepsilon_t) \sim U(0, 1)$ when $\varepsilon_t \sim N(0, 1)$, and $G\{\log(e_t^2)\} \sim U(0, 1)$ when $G(\cdot)$ is the distribution function of $\log(e_t^2)$. Let $\widehat{F}_{n,1}(\cdot)$ be the empirical distribution of $\{\Phi(\widehat{\varepsilon}_t),\ \nu < t \le n\}$, and $\widehat{F}_{n,2}(\cdot)$ the empirical distribution of $[G\{\log(\widehat{e}_t^2)\},\ \nu < t \le n]$. We define the goodness-of-fit statistics below to measure the distances between $\widehat{F}_{n,i}$ and the uniform distribution on (0, 1).

$$T_{\text{MLE}} = \int_0^1 |\widehat{F}_{n,1}(x) - x|dx, \qquad T_{\text{LADE}} = \int_0^1 |\widehat{F}_{n,2}(x) - x|dx. \tag{2.10}$$

Obviously, these statistics are reminiscent of the Cramér–von Mises goodness-of-fit statistics. In practical implementation, we use the Riemann approximations of these integrals:

$$T_{\text{MLE}} = \sum_{t=\nu+1}^{n} \left|\frac{t-\nu}{n-\nu} - u_t\right|(u_t - u_{t-1}), \qquad T_{\text{LADE}} = \sum_{t=\nu+1}^{n} \left|\frac{t-\nu}{n-\nu} - v_t\right|(v_t - v_{t-1}), \tag{2.11}$$

where $u_{\nu+1} \le u_{\nu+2} \le \cdots \le u_n$ are the order statistics of $\{\Phi(\widehat{\varepsilon}_t),\ \nu < t \le n\}$, and $v_{\nu+1} \le v_{\nu+2} \le \cdots \le v_n$ the order statistics of $[G\{\log(\widehat{e}_t^2)\},\ \nu < t \le n]$.

Selection rule: we use the LADE if $T_{\text{MLE}} > T_{\text{LADE}}$, and the GMLE otherwise.

Let F_1 and F_2 denote, respectively, the distribution function of $\Phi(\varepsilon_t)$ and $G\{\log\ (e_t^2)\}$. Theorem 2.1. indicates that the selection role defined above is consistent. Its proof is given in the Appendix.

THEOREM 2.1 *Let $\{X_t\}$ be defined by (2.1) for which condition (2.2) holds. Let $\nu \to \infty$ and $\nu/n \to 0$ as $n \to \infty$. Suppose that for some constant $\kappa_1, \kappa_2 > 0$,*

$$||\widehat{\theta} - \theta|| = O_P(n^{-\kappa_1}), \qquad ||\widehat{\alpha} - \alpha|| = O_P(n^{-\kappa_2}). \tag{2.12}$$

Furthermore, for any constant $\delta_0 > 0$, there exists $\delta > 0$ for which

$$\sup_{0\le x\le 1} |F_1(x+\delta) - F_1(x-\delta)| < \delta_0, \qquad \sup_{0\le x\le 1} |F_2(x+\delta) - F_2(x-\delta)| < \delta_0. \tag{2.13}$$

Then as $n \to \infty$, $P(T_{\text{MLE}} > T_{\text{LADE}}) \to 1$ provided

$$\int_0^1 |F_1(x) - x|dx > \int_0^1 |F_2(x) - x|dx. \tag{2.14}$$

Condition (2.12) requires that both the GMLE and the LADE are, respectively, n^{κ_1} and n^{κ_2} consistent, which has been established under certain regularity conditions. For the LADE, $\kappa_2 = 1/2$ (Peng and Yao, 2003). For the GMLE, the value of κ_1 is related to the tail heaviness of the distribution of ε_t. In fact, such a positive κ_1 always exists for the GMLE when $E(\varepsilon_t^2) < \infty$

(Hall and Yao, 2003; Straumann, 2005; Mikosch and Straumann, 2006). Condition (2.13) is fulfilled if, for example, both F_1 and F_2 admit bounded probability density functions.

3. NUMERICAL ILLUSTRATION

In this section, we first illustrate the proposed selection procedure with the data simulated from GARCH(1,1) and ARCH(2) models. In both cases, we took the errors ε_t to be $N(0, 1)$, standardized t or skewed t with $d = 3$ or 6 degrees of freedom. A skewed t random variable is defined as

$$(0.8|V_0| + 0.6V_1)/(V_2/d)^{1/2},$$

where V_0 and V_1 are N(0, 1) random variables, $V_2 \sim \chi^2(d)$ and V_0, V_1 and V_2 are independent with each other (see Azzalini and Capitanio, 2003). We also used ε_t such that log (ε_t^2) is of a Laplace distribution. We further experimented with the semi-strong ARCH/GARCH models defined in terms of martingale difference innovations

$$\varepsilon_t = \mathrm{sgn}(\xi_t)\big[1 + \big(\eta_t^2 - 1\big)\big/\big\{1 + \exp\big(\sigma_t^2\big)\big\}\big]^{1/2},$$

where ξ_t and η_t are independent $N(0, 1)$ random variables. We used $c = 1$, $(b_1, b_2) = (0.7, 0.2)$ for the ARCH(2) model, and $(b_1, a_1) = (0.2, 0.7)$ for the GARCH(1,1) model. Setting the sample size $n = 250$, 500 and 1000, we drew 200 samples for each setting. We used $\nu = 20$ in the estimation for GARCH models.

The relative frequencies for the occurrence of the event $\{T_{\text{LADE}} < T_{\text{MLE}}\}$ in the 200 replications are listed in Table 1 for GARCH(1,1) model, and in Table 2 for ARCH(2) model. We also included in the tables the relative frequencies of the occurrence of the event $\{\mathcal{E}_{\text{LADE}} < \mathcal{E}_{\text{MLE}}\}$, where the estimation errors are defined as

$$\mathcal{E}_{\text{LADE}} = \sum_{i=1}^{p} |\widehat{\beta}_i/\widehat{\gamma} - b_i/c| + \sum_{j=1}^{q} |\widehat{a}_j - a_j|, \qquad \mathcal{E}_{\text{MLE}} = \sum_{i=1}^{p} |\widehat{b}_i/\widehat{c} - b_i/c| + \sum_{j=1}^{q} |\widehat{a}_j - a_j|$$

(see 2.6 and 2.1). For the models with normal innovations, the GMLE is the genuine MLE, and is always the preferred estimator according to our selection procedure. On the other hand, the LADE is always selected when $\log(e_t^2)$ has the Laplace distribution. For the models with t(d)-innovations, the results are less clear-cut. Overall, the GMLE is preferred when $d = 6$ while the LADE is preferred when $d = 3$. Furthermore, the preference for the GMLE when $d = 6$ and that for the LADE when $d = 3$ increases when the sample size n increases. The models with skewed t-innovations tend to be in favour of LADE more often than those with (centred) t-innovations with the same degrees of freedom. For semi-strong GARCH/ARCH models with martingale difference innovations, the LADE is preferred. This may be due to the fact that the innovation distribution is very different from normal and L_1 estimation is more robust. Overall, there is a clear synchrony between the occurrences of the two events $\{T_{\text{LADE}} < T_{\text{MLE}}\}$ and $\{\mathcal{E}_{\text{LADE}} < \mathcal{E}_{\text{MLE}}\}$, indicating that overall the preferred method by the T-measures leads to more accurate estimates for the parameters.

Now, we apply the method to two centred daily return series: the Switzerland stock index (SWI) in 2 January 1991–31 December 1998, and the B Share of the Shanghai Stock Exchange (SHB) in 2 January 2001–31 December 2004. The length of the series is, respectively, 1859 and 946. The P-value of the Jarque–Bera Test is 0.000 for both the series, and the kurtosis is 5.72665 for SWI and 5.761476 for SHB. For each of those two series, we fit the first half series

Table 1. Simulation results for GARCH(1,1) model—relative frequencies for the occurrences of the events $\{T_{\text{LADE}} < T_{\text{MLE}}\}$ and $\{\mathcal{E}_{\text{LADE}} < \mathcal{E}_{\text{MLE}}\}$ in 200 replications.

Distribution of ε_t	n	$T_{\text{LADE}} < T_{\text{MLE}}$	$\mathcal{E}_{\text{LADE}} < \mathcal{E}_{MLE}$
$N(0, 1)$	250	0.000	0.290
	500	0.000	0.235
	1000	0.000	0.295
$t(6)$	250	0.030	0.360
	500	0.000	0.425
	1000	0.000	0.450
Skewed $t(6)$	250	0.065	0.430
	500	0.065	0.470
	1000	0.005	0.490
$t(3)$	250	0.575	0.655
	500	0.680	0.650
	1000	0.790	0.695
Skewed $t(3)$	250	0.805	0.710
	500	0.925	0.725
	1000	0.965	0.775
Laplace	250	1.000	0.745
	500	1.000	0.765
	1000	1.000	0.845
Martingale difference	250	1.000	0.070
	500	1.000	0.850
	1000	1.000	0.730

with GARCH(1,1) models using both the GMLE and the LADE. The sample size used in the estimations is $n = 930$ for SWI, and 473 for SHB. The values of the goodness-of-fit statistics $(T_{\text{MLE}}, T_{\text{LADE}})$ are (0.026, 0.057) for SWI and (0.044, 0.041) for SHB. Thus, our selection rule prefers the GMLE for SWI, and the LADE for SHB.

With the sample size fixed at $n = 930$ for SWI and $n = 473$ for SHB, we also perform one-step ahead prediction of the squared returns for each of the second half series. The prediction is based on the fitted GARCH(1,1) models using both the GMLE and the LADE. With LADE, the predicted squared returns are of the form $\widehat{s}_t^2 S_e$, where S_e is the sample variance of the residuals $\widehat{e}_j \equiv X_j/\widehat{s}_j (j < t)$ (see 2.6). The root mean squares error of the prediction based on the GMLE is 1.750 for SWI, and 4.757 for SHB. The root mean squares error based on the LADE is 2.715 for SWI and 2.894 for SHB. Thus, the GMLE provided the more accurate prediction for SWI, while the LADE predicted SHB better. This shows that the estimation method preferred by our selection rule also provided better prediction.

4. MISCELLANEOUS REMARKS

Although we deal with the estimation for GARCH models only in this paper, the idea may apply to select an appropriate estimation method in, for example, a general regression

Table 2. Simulation results for ARCH(2) model—relative frequencies for the occurrences of the events $\{T_{\text{LADE}} < T_{\text{MLE}}\}$ and $\{\mathcal{E}_{\text{LADE}} < \mathcal{E}_{\text{MLE}}\}$ in 200 replications.

Distribution of ε_t	n	$T_{\text{LADE}} < T_{\text{MLE}}$	$\mathcal{E}_{\text{LADE}} < \mathcal{E}_{\text{MLE}}$
$N(0, 1)$	250	0.000	0.290
	500	0.000	0.260
	1000	0.000	0.260
$t(6)$	250	0.015	0.415
	500	0.000	0.470
	1000	0.000	0.450
Skewed $t(6)$	250	0.075	0.510
	500	0.055	0.460
	1000	0.010	0.505
$t(3)$	250	0.570	0.605
	500	0.675	0.625
	1000	0.770	0.725
Skewed $t(3)$	250	0.840	0.700
	500	0.890	0.730
	1000	0.975	0.785
Laplace	250	1.000	0.795
	500	1.000	0.825
	1000	1.000	0.800
Martingale	250	1.000	0.170
difference	500	1.000	0.930
	1000	1.000	0.995

model

$$y = f(X) + \varepsilon. \tag{4.1}$$

When f is known up to some unknown parameters, it is a parametric regression model. When f is completely unknown, it is a nonparametric regression problem. Nevertheless, both the least squares estimation (LSE) and least absolutely deviations estimation (LADE) are well developed in both parametric and non-parametric setting. Intuitively, LSE should be used when the distribution of ε is close to a normal distribution while LADE should be used when the distribution of ε is close to a Laplace distribution. The procedure presented in Section 3.2 is readily applicable for the selection between those two estimation methods.

The above problem may be seen as to choose the most relevant distribution from the union of the normal distribution family and the Laplace distribution family. In this sense, it is a kind of model selection problem. However, we argue that such an estimation-selection problem is different from the conventional model-selection problems often featured in statistical literature. The standard information criteria such as the AIC are designed to select the most relevant model from a given *smooth* parametric family under the assumption that the family contains the true model as one of its members. When the truth is not in the family, the criteria such the TIC (Takeuchi, 1976; Konishi and Kitagawa, 1996) may be used to select the 'best' approximation for the truth within the given family. However, to our best knowledge, no criteria may be applied to select an ‘optimum’ approximate model for the truth across two or more parametric families. The lack of such a criterion is due to the fact that the maximum likelihood principle may not apply across different distribution families.

We may embed the two distribution families into one via, for example, a convex combination. This is to consider, for the regression model (4.1), the error distribution family

$$\pi N(0, \sigma^2) + (1 - \pi)L(0, \lambda), \quad \pi \in [0, 1],\ \sigma^2 > 0,\ \lambda > 0,$$

where $L(0, \lambda)$ denotes the Laplace distribution centred at 0 and with scale parameter λ. Now the MLE for π is typically neither 0 nor 1. Consequently, the MLE for $f(\cdot)$ is neither LSE nor LADE. Therefore, this approach, though legitimate on its own, would not provide an answer to the problem concerned.

ACKNOWLEDGMENTS

H. Wang acknowledges the support of Chinese NSF grant #10771006. Q. Yao’s work was partially supported by EPSRC research grants GR/R97436 and EP/C549058. We thanks two referees for helpful comments and suggestions.

REFERENCES

Azzalini, A. and A. Capitanio (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew-t distribution. *Journal of the Royal Statistical Society*, Series B *65,* 367–89.

Berkes, I., L. Horváth and P. Kokoszka (2003). GARCH processes: structure and estimation. *Bernoulli 9,* 201–27.

Chow, Y. S. and H. Teicher (1997). *Probability Theory* (3rd ed.). New York: Springer.

Fan, J. and Q. Yao (2003). *Nonlinear Time Series: Nonparametric and Parametric Methods*. New York: Springer.

Giraitis, L., P. Kokoszka and R. Leipus (2000). Stationary ARCH models: dependence structure and central limit theorem. *Econometric Theory 16,* 3–22.

Giraitis, L. and P. M. Robinson (2001). Whittle estimation of ARCH models. *Econometric Theory 17,* 608–31.

Hall, P. and Q. Yao (2003). Inference in ARCH and GARCH models with heavy-tailed errors. *Econometrica 71,* 285–317.

Horváth, L. and F. Liese (2004). L_p-estimators in ARCH models. *Journal of Statistical Planning and Inference 119,* 277–309.

Konishi, S. and G. Kitagawa (1996). Generalized information criteria in model selection. *Biometrika 83,* 875–90.

Lee, S. W. and B. E. Hansen (1994). Asymptotic theory for the GARCH(1,1) quasi-maximum likelihood estimator. *Econometric Theory 10,* 29–52.

Lumsdaine, R. (1996). Consistency and asymptotic normality of the quasi-maximum likelihood estimator in IGARCH(1,1) and covariance stationary GARCH(1,1) models. *Econometrica 64,* 575–96.

Mikosch, T. and T. Straumann (2006). Stable limits of martingale transforms with application to the estimation of GARCH parameters. *Annals of Statistics 34,* 493–522.

Peng, L. and Q. Yao (2003). Least absolute deviations estimation for ARCH and GARCH models. *Biometrika 90,* 967–75.

Straumann, D. (2005). *Estimation in Conditionally Heteroscedastic Time Series Models*. Heidelberg: Springer.

Straumann, D. and T. Mikosch (2006). Quasi-MLE in heteroscedastic times series: a stochastic recurrence equations approach. *Annals of Statistics 34*, 2449–95.

Takeuchi, K. (1976). Distribution of information statistics and criteria for adequacy of models. *Mathematical Sciences 153*, 12–18 (In Japanese).

Weiss, A. (1986). Asymptotic theory for ARCH models: estimation and testing. *Econometric Theory 2,* 107–31.

APPENDIX: PROOF OF THEOREM 2.1

We use the same notation as in Section 2. Put $U_t = \Phi(\varepsilon_t)$, $\widehat{U}_t = \Phi(\widehat{\varepsilon}_t)$, $V_t = G\{\log(e_t^2)\}$ and $\widehat{V}_t = G\{\log(\widehat{e}_t^2)\}$. Let $A_n = \{||\widehat{\theta} - \theta|| < n^{-\kappa_1/2}\}$ and $B_n = \{||\widehat{\alpha} - \alpha|| < n^{-\kappa_2/2}\}$. It follows from (2.12) that $P(A_n) \to 1$ and $P(B_n) \to 1$. Denote by '$\xrightarrow{P}$' the convergence in probability, and C, C_1 and C_2 some generic positive constants which may be different at different places. We split the proof into several lemmas. We assume that the conditions of Theorem 2.1 always hold in this Appendix.

LEMMA A.1. *As $n \to \infty$, it holds that $\sum_{t>\nu} E\{|X_t||\sigma_t(\theta) - \widetilde{\sigma}_t(\theta)|\} \to 0$.*

Proof. It follows from (2.3) and (2.5) that for any $t > p$,

$$E\left|\sigma_t(\theta)^2 - \widetilde{\sigma}_t(\theta)^2\right| \le E\left(X_t^2\right)\sum_{i=1}^{p} b_i \sum_{k\ge(t-p)/q}\ \sum_{j_1=1}^{q}\cdots\sum_{j_k=1}^{q} a_{j_1}\cdots a_{j_k}$$

$$\le E\left(X_t^2\right)\sum_{i=1}^{p} b_i \frac{(a_1+\cdots+a_q)^{(t-p)/q}}{1-(a_1+\cdots a_q)^{1/q}}$$

(see also 2.2). Hence,

$$\sum_{t>\nu}\left[E\left|\sigma_t(\theta)^2 - \widetilde{\sigma}_t(\theta)^2\right|\right]^{1/2} \le C\sum_{t>\nu}(a_1+\cdots+a_q)^{(t-p)/(2q)} \tag{A.1}$$

$$\le C\frac{(a_1+\cdots+a_q)^{(\nu-p)/(2q)}}{1-(a_1+\cdots a_q)^{1/(2q)}} \to 0.$$

Note that $E(X_t^2) < \infty$, which is ensured by (2.2). By (A.1), it holds that

$$\sum_{t=\nu+1}^{n} E\{|X_t||\sigma_t(\theta) - \widetilde{\sigma}_t(\theta)|\} \le C\sum_{t>\nu}\left[E\left\{|\sigma_t(\theta) - \widetilde{\sigma}_t(\theta)|^2\right\}\right]^{1/2}$$

$$\le C\textstyle\sum_{t>\nu}\left[E\left\{\left|\sigma_t(\theta)^2 - \widetilde{\sigma}_t(\theta)^2\right|\right\}\right]^{1/2} \to 0.$$

This completes the proof. □

LEMMA A.2. *As* $n \to \infty$, $(n-\nu)^{-1}\sum_{\nu<t\le n} E|X_t\{\widetilde{\sigma}_t(\widehat{\theta}) - \widetilde{\sigma}_t(\theta)\}I(A_n)| \to 0.$

Proof. It holds on the set A_n that $\sum_{1\le j\le q}\widehat{a}_j$ is bounded from the above by a constant smaller than 1 for all sufficiently large n.

Replace the sum over $1 \le k < \infty$ in the third term on the RHS of (2.5) by the sum over $1 \le k \le n^{\kappa_1/4}$, and denote by $\breve{\sigma}_t(\theta)^2$ the resulting function on the RHS. Then, similar to Lemma A.1, we may show that

$$\sum_{t=\nu+1}^{n} E[|X_t|\{|\widetilde{\sigma}_t(\widehat{\theta}) - \breve{\sigma}_t(\widehat{\theta})| + |\widetilde{\sigma}_t(\theta) - \breve{\sigma}_t(\theta)|\}I(A_n)] \to 0. \tag{A.2}$$

On the other hand,

$$\frac{1}{n-\nu}\sum_{t=\nu+1}^{n} E\{|X_t||\breve{\sigma}_t(\widehat{\theta}) - \breve{\sigma}_t(\theta)|I(A_n)\} \le \frac{C}{n-\nu}\sum_{t=\nu+1}^{n}\left[E\left\{|\breve{\sigma}_t(\widehat{\theta}) - \breve{\sigma}_t(\theta)|^2 I(A_n)\right\}\right]^{1/2}$$
$$\le \frac{C}{n-\nu}\sum_{t=\nu+1}^{n}\left[E\left\{\left|\breve{\sigma}_t(\widehat{\theta})^2 - \breve{\sigma}_t(\theta)^2\right|I(A_n)\right\}\right]^{1/2} \le C_1\left\{E\left(X_t^2\right)\right\}^{1/2} n^{-\kappa_1/2} n^{\kappa_1/4} \to 0.$$

The result required follows from this and (A.2). □

LEMMA A.3. *For any given constant x,*

(i) $\sup_{0\le x\le 1}\frac{1}{n-\nu}\sum_{\nu<t\le n}|I(\widehat{U}_t \le x) - I(U_t \le x)| \xrightarrow{P} 0$ *and*

(ii) $\sup_{0\le x\le 1}\frac{1}{n-\nu}\sum_{\nu<t\le n}|I(\widehat{V}_t \le x) - I(V_t \le x)| \xrightarrow{P} 0.$

Proof. We prove (i) first. Since the standard normal density function is bounded, it holds that

$$|\widehat{U}_t - U_t| \le C|\widehat{\varepsilon}_t - \varepsilon_t| \le C_1|X_t||\widetilde{\sigma}_t(\widehat{\theta}) - \sigma_t(\theta)|/\widetilde{\sigma}_t(\widehat{\theta}).$$

Note that $1/\widetilde{\sigma}_t(\widehat{\theta})$ is bounded from above by a finite constant on the set A_n for all sufficiently large n. It follows from Lemmas (A.1) and (A.2) that $(n-\nu)^{-1}\sum_{\nu<t\le n} E\{|\widehat{U}_t - U_t|I(A_n)\} \to 0$. This implies that for any $\delta > 0$,

$$\frac{1}{n-\nu}\sum_{t=\nu+1}^{n} I(|\widehat{U}_t - U_t| > \delta,\ A_n) \xrightarrow{P} 0. \tag{A.3}$$

Note that

$$|I(\widehat{U}_t \le x) - I(U_t \le x)| \le I(\widehat{U}_t \le x, U_t > x) + I(\widehat{U}_t > x, U_t \le x)$$
$$\le I(|\widehat{U}_t - U_t| > \delta) + I(U_t \in [x-\delta, x+\delta]).$$

Therefore,

$$\sup_x \frac{1}{n-\nu}\sum_{t=\nu+1}^{n}|I(\widehat{U}_t \le x) - I(U_t \le x)| \le \frac{1}{n-\nu}\sum_{t=\nu+1}^{n} I(|\widehat{U}_t - U_t| > \delta, A_n) \tag{A.4}$$
$$+ \sup_x \frac{1}{n-\nu}\sum_{t=\nu+1}^{n} I(U_t \in [x-\delta, x+\delta]) + I\left(A_n^c\right).$$

For any given $\delta_0 > 0$, $P(A_n^c) < \delta_0$ for all sufficiently large n. Note that

$$\begin{aligned}
&\sup_x \frac{1}{n-\nu} \sum_{t=\nu+1}^{n} I(U_t \in [x-\delta, x+\delta]) \\
&\quad \le \sup_x \Big| \frac{1}{n-\nu} \sum_{t=\nu+1}^{n} I(U_t \in [x-\delta, x+\delta]) - F_1(x+\delta) + F_1(x-\delta) \Big| \\
&\qquad + \sup_x |F_1(x+\delta) - F_1(x-\delta)|.
\end{aligned}$$

By the Glivenko–Cantelli Theorem (p. 284 of Chow and Teicher, 1997), the first term on the RHS of the above expression converges to 0 almost surely. Condition (2.13) implies that the second term may be smaller than δ_0 by choosing δ sufficiently small. Now the result follows from (A.4) and (A.3).

Now we prove (ii). Since a Laplace density function is bounded,

$$\begin{aligned}
|\widehat{V}_t - V_t| &\le C \big| \log(\widehat{e}_t^2) - \log(e_t^2) \big| = C \big| \log\{s_t(\alpha)^2/\widetilde{s}_t(\widehat{\alpha})^2\} \big| \\
&\le C \big[\big| \log\{s_t(\alpha)^2/\widetilde{s}_t(\alpha)^2\} \big| + \big| \log\{\widetilde{s}_t(\alpha)^2/\widetilde{s}_t(\widehat{\alpha})^2\} \big| \big].
\end{aligned}$$

Note that for $x \ge 0$, $\log\,(1+x) \le x$ and $s_t(\alpha)^2 \ge \widetilde{s}_t(\alpha)^2 \ge \gamma > 0$. Hence, the above expression implies that

$$\begin{aligned}
|\widehat{V}_t - V_t| \le C \Bigg[& \frac{\big|s_t(\alpha)^2 - \widetilde{s}_t(\alpha)^2\big|}{\widetilde{s}_t(\alpha)^2} \\
& + \big|\widetilde{s}_t(\alpha)^2 - \widetilde{s}_t(\widehat{\alpha})^2\big| \left\{ \frac{I\{\widetilde{s}_t(\alpha)^2 > \widetilde{s}_t(\widehat{\alpha})^2\}}{\widetilde{s}_t(\widehat{\alpha})^2} + \frac{I\{\widetilde{s}_t(\alpha)^2 \le \widetilde{s}_t(\widehat{\alpha})^2\}}{\widetilde{s}_t(\alpha)^2} \right\} \Bigg] \\
\le \frac{C}{\gamma} & \big[\big|s_t(\alpha)^2 - \widetilde{s}_t(\alpha)^2\big| + \big|\widetilde{s}_t(\alpha)^2 - \widetilde{s}_t(\widehat{\alpha})^2\big| \big\{1 + \gamma/\widetilde{s}_t(\widehat{\alpha})^2\big\} \big].
\end{aligned}$$

When n is sufficiently large, $\widetilde{s}_t(\widehat{\alpha})^2$ is bounded from below by a positive constant on the set B_n. Thus, it holds on B_n that

$$|\widehat{V}_t - V_t| \le C_1 \big|s_t(\alpha)^2 - \widetilde{s}_t(\alpha)^2\big| + C_2 \big|\widetilde{s}_t(\alpha)^2 - \widetilde{s}_t(\widehat{\alpha})^2\big|.$$

Now using the similar arguments as in the proofs of Lemmas A.1 and A.2, we may show that

$$\sum_{t>\nu} E\big|s_t(\alpha)^2 - \widetilde{s}_t(\alpha)^2\big| \to 0, \quad \text{and} \quad \frac{1}{n-\nu} \sum_{t=\nu+1}^{n} E\big\{\big|\widetilde{s}_t(\alpha)^2 - \widetilde{s}_t(\widehat{\alpha})^2\big| I(B_n)\big\} \to 0.$$

Now proceeding as the proof for (i) above, we may obtain the required result. □

Proof of Theorem 2.1: Let $F_{n,1}$ and $F_{n,2}$ be, respectively, the empirical distribution of $\{U_t, \nu < t \le n\}$ and $\{V_t, \nu < t \le n\}$. By Lemma (A.3) and the Glivenko–Cantelli Theorem, it holds that

$$\sup_x |\widehat{F}_{n,i}(x) - F_i(x)| \le \sup_x |\widehat{F}_{n,i}(x) - F_{n,i}(x)| + \sup_x |F_{n,i}(x) - F_i(x)| \xrightarrow{P} 0,$$

for $i = 1, 2$. This and condition (2.14) entail the required result. □