

Estimating GARCH Models: When to Use What?

Da Huang¹ Hansheng Wang¹ Qiwei Yao^{2,1}

¹Guanghua School of Management, Peking University, Beijing 100871, China

²Department of Statistics, London School of Economics, London, WC2A 2AE, UK

Abstract

The class of 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 quasilielihood estimator based on a Gaussian likelihood (GMLE) and the log-transform based least absolutely deviations estimator (LADE); see Peng and Yao (2003). 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 quasilielihood 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.

Some key words: Estimation procedure selection; GARCH; Gaussian likelihood; Heavy tail; Laplace distribution; Least absolute deviations estimator; Maximum quasilielihood estimator; Time series.

1 Introduction

There exist more than one methods for estimating parameters in GARCH models with unknown innovation distributions. The maximum quaslikelihood 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, eg. Hall & Yao (2003). To overcome the drawbacks due to the possible slow convergence rates of the GMLE, Peng & Yao (2003) proposes 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 & 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 quaslikelihood 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 below. 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, by Lee & 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 & Yao (2003), Berkes, Horváth & Kokoszka (2003), Straumann & Mikosch (2003), and Mikosch & Straumann (2003). See also Straumann (2005). Complex asymptotic properties were also observed from a Whittle estimator by Giraitis & Robinson (2001) for heavy tailed GARCH(1,1) models. Pan, Tong & Yao (2006) studied the LADE in a more general framework including asymmetric power GARCH(p, q) models. The asymptotic properties of L_p -estimators for ARCH(p) models were established by Horvath 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 illustration 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 generalised autoregressive conditional heteroscedastic, GARCH, model with orders $p \geq 1$ and $q \geq 0$ is defined as

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

where $c > 0$, $b_j \geq 0$ and $a_j \geq 0$ are unknown parameters, $\theta = (c, b_1, \dots, b_p, a_1, \dots, a_q)^T$, $\{\varepsilon_t\}$ 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 $\{X_t, t = 0, \pm 1, \pm 2, \dots\}$ with $EX_t^2 < \infty$ is that

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

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

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

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

2.2 Two estimators

The GMLE is defined as

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

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

$$\begin{aligned} \tilde{\sigma}_t(\theta)^2 &= \frac{c}{1 - \sum_{j=1}^q a_j} + \sum_{i=1}^{\min(p, t-1)} b_i X_{t-i}^2 + \sum_{i=1}^p b_i \sum_{k=1}^{\infty} \sum_{j_1=1}^q \cdots \sum_{j_k=1}^q a_{j_1} \cdots a_{j_k} \\ &\times X_{t-i-j_1-\dots-j_k}^2 I(t-i-j_1-\dots-j_k \geq 1), \end{aligned} \quad (2.5)$$

which depends on the observations X_{t-1}, \dots, X_1 only; c.f. (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 $X_{\nu+1}, \dots, X_n$ is then proportional to

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

Maximising this (conditional) likelihood with $\sigma_t(\theta)^2$ replaced by $\tilde{\sigma}_t(\theta)^2$ 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

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

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

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

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

$$\hat{\alpha} = \arg \min_{\alpha} \sum_{t=\nu+1}^n |\log(X_t^2) - \log\{\tilde{s}_t(\alpha)^2\}|, \quad (2.8)$$

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

$$\begin{aligned} \tilde{s}_t(\alpha)^2 &= \frac{\gamma}{1 - \sum_{j=1}^q a_j} + \sum_{i=1}^{\min(p, t-1)} \beta_i X_{t-i}^2 + \sum_{i=1}^p \beta_i \sum_{k=1}^{\infty} \sum_{j_1=1}^q \cdots \sum_{j_k=1}^q a_{j_1} \cdots a_{j_k} \\ &\times X_{t-i-j_1-\cdots-j_k}^2 I(t-i-j_1-\cdots-j_k \geq 1), \end{aligned} \quad (2.9)$$

which directly follows from (2.5).

In fact the LADE may also be viewed as a maximum quasilielihood 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 $X_{\nu+1}, \dots, X_n$ (given $\{X_k, k \leq \nu\}$) is then proportional to

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

Maximising this with $s_t(\alpha)^2$ replaced by $\tilde{s}_t(\alpha)^2$ leads to the LADE $\hat{\alpha}$; 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 $\hat{\theta}$ and $\hat{\alpha}$ hinges critically on the tail heaviness of the innovation distribution. When $E(|\varepsilon_t|^{4-\delta}) < \infty$ for any $\delta > 0$, $\hat{\theta}$ 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 $\hat{\theta}$ 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. See, for example, Hall and Yao (2003). On the other hand, the LADE $\hat{\alpha}$ 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 perforation 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 questions 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 $\hat{\varepsilon}_t = X_t/\tilde{\sigma}_t(\hat{\theta})$ be the residuals derived from the GMLE. In practice, we standardize $\hat{\varepsilon}_t$ such that the first two sample moments are 0 and 1. Let $\hat{e}_t = X_t/\tilde{s}_t(\hat{\alpha})$ be the residuals derived from the LADE. In practice, we ‘standardize’ \hat{e}_t such that the sample median of \hat{e}_t^2 is 1 and the sample mean of $|\log(\hat{e}_t^2)|$ is 2. This may be achieved by letting $\log(\hat{e}_t^2) = C_1 \log\{C_2 X_t^2/\tilde{s}_t(\hat{\alpha})^2\}$ for appropriate positive constants C_1 and C_2 . (Therefore the first two sample moments of $\log(\hat{e}_t^2)$ are the same as the first two moments of the Laplace distribution with density $0.25 \exp(-|x|/2)$.) 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 $\hat{F}_{n,1}(\cdot)$ be the empirical distribution of $\{\Phi(\hat{\varepsilon}_t), \nu < t \leq n\}$, and $\hat{F}_{n,2}(\cdot)$ the empirical distribution of $[G\{\log(\hat{e}_t^2)\}, \nu < t \leq n]$. We define the goodness-of-fit statistics below to measure the distances between $\hat{F}_{n,i}$ and the uniform distribution on $(0, 1)$.

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

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

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

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

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 1 below indicates that the selection rule defined above is consistent. Its proof is given in the Appendix.

Theorem 1. Let $\{X_t\}$ be defined by (2.1) for which condition (2.2) holds. Let $\nu \rightarrow \infty$ and $\nu/n \rightarrow 0$ as $n \rightarrow \infty$. Suppose that for some constant $\kappa_1, \kappa_2 > 0$,

$$\|\hat{\theta} - \theta\| = O_P(n^{-\kappa_1}), \quad \|\hat{\alpha} - \alpha\| = O_P(n^{-\kappa_2}). \quad (2.12)$$

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

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

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

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

Condition (2.12) requires that both the GMLE and the LADE are, respectively, n^{κ_1} and n^{κ_2} consistent, which have been established under certain regularity conditions. For the LADE, $\kappa_2 = 1/2$ (Peng & Yao 2003, and Pan, Tong & Yao 2006). 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 & Yao 2003, Mikosch & Straumann 2003, and Straumann 2005). 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 ARCH(2) and GARCH(1,1) models. In both cases we took the errors ε_t to have the standard normal distribution or a standardised Student's t -distribution with $d = 3$ or 4 degrees of freedom. We also took ε_t such that $\log(\varepsilon_t^2)$ have the Laplace distribution. We standardised those distributions to ensure that their first two moments are, respectively, 0 and 1. Note that, when $\varepsilon_t \sim t(d)$, $E|\varepsilon_t|^d = \infty$. We used $c = 1$, $(b_1, b_2) = (0.7, 0.2)$ for ARCH(2) model, and $(b_1, a_1) = (0.2, 0.7)$ for GARCH(1,1) model. Setting the sample size $n = 250, 500$, we drew 200 samples respectively for each setting. We used $\nu = 20$ in the estimation for GARCH models.

Figures 1 and 2 display the plots of T_{LADE} against T_{MLE} in the 200 replications for, respectively, the ARCH(2) and GARCH(1, 1) models. We also superimposed the plots of \mathcal{E}_{LADE} against \mathcal{E}_{MLE} in the same figures, where the estimation errors are defined as

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

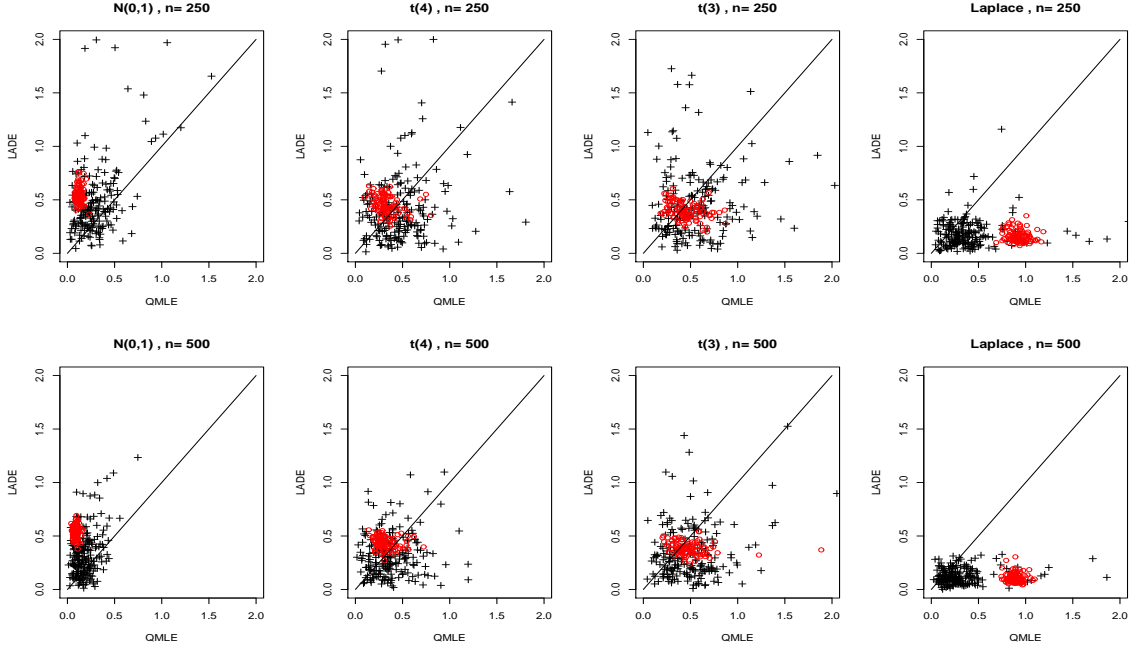


Figure 1: *Simulation results for ARCH(2) model: Plots of T_{LADe} against T_{MLE} (circles) and plots of \mathcal{E}_{LADe} against \mathcal{E}_{MLE} (crosses).*

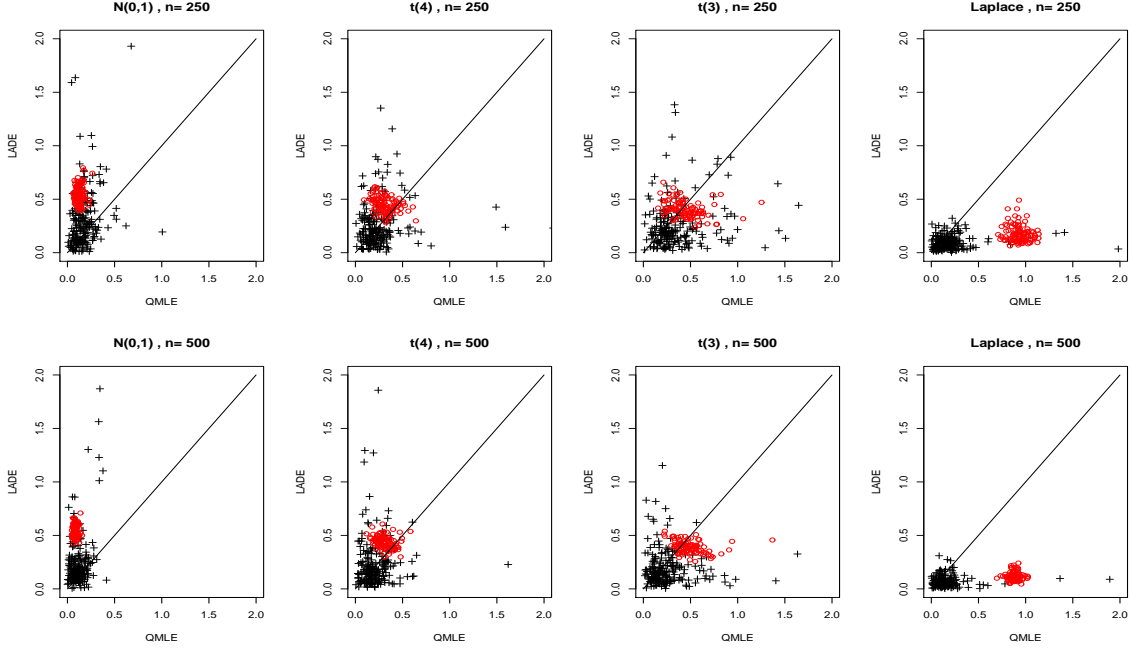


Figure 2: *Simulation results for GARCH(1,1) model: Plots of T_{LADe} against T_{MLE} (circles) and plots of \mathcal{E}_{LADe} against \mathcal{E}_{MLE} (crosses).*

see (2.6) and (2.1). We rescaled these two estimation errors by a common positive constant to optimize the visual effect of the figures. To facilitate comparison, we draw a diagonal line $y = x$ in each panel. The points under the line indicate that the LADe is preferred than the QMLE, and

the vice verse. For the models with normal innovations, the QMLE is the genuine MLE and it was always selected by the proposed procedure. On the other hand, the LADE was 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 QMLE is preferred when $d = 4$ while the LADE is preferred when $d = 3$. The preferences for the QMLE when $d = 4$ and for the LADE when $d = 3$ increases when the same size n increases. There is a clear synchrony between the T -plots and \mathcal{E} -plots; 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 centered 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 series are, respectively, 1859 and 946. For each of those two series, we fit the first half series with GARCH(1,1) models using both the GMLE and the LADE. The sample size used in estimation is $n = 930$ for SWI, and 473 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 QMLE for SWI, and the LADE for SHB.

With the sample sized 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 QMLE and the LADE. The root mean squares error of the prediction based on the QMLE 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 QMLE 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 model

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

When f is known upto some unknown parameter, 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 absolute deviations estimation (LADE) are well-developed in both parametric and nonparametric 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

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

where $L(0, \lambda)$ denotes the Laplace distribution centered 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.

ACKNOWLEDGMENT

Q. Yao’s work was partially supported by the U.K. Engineering and Physical Sciences Research Council.

Appendix: Proof of Theorem 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 $P(A_n) \rightarrow 1$ and $P(B_n) \rightarrow 1$. 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 1 always hold in this appendix.

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

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

$$\begin{aligned} E|\sigma_t(\theta)^2 - \widetilde{\sigma}_t(\theta)^2| &\leq E(X_t^2) \sum_{i=1}^p b_i \sum_{k \geq (t-p)/q} \sum_{j_1=1}^q \cdots \sum_{j_k=1}^q a_{j_1} \cdots a_{j_k} \\ &\leq E(X_t^2) \sum_{i=1}^p b_i \frac{(a_1 + \cdots + a_q)^{(t-p)/q}}{1 - (a_1 + \cdots + a_q)^{1/q}}, \end{aligned}$$

see also (2.2). Hence

$$\sum_{t>\nu} [E|\sigma_t(\theta)^2 - \widetilde{\sigma}_t(\theta)^2|]^{1/2} \leq C \sum_{t>\nu} (a_1 + \cdots + a_q)^{(t-p)/(2q)} \leq C \frac{(a_1 + \cdots + a_q)^{(\nu-p)/(2q)}}{1 - (a_1 + \cdots + a_q)^{1/(2q)}} \rightarrow 0. \quad (\text{A.1})$$

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

$$\sum_{t=\nu+1}^n E\{|X_t| |\sigma_t(\theta) - \widetilde{\sigma}_t(\theta)|\} \leq C \sum_{t>\nu} [E\{|\sigma_t(\theta) - \widetilde{\sigma}_t(\theta)|^2\}]^{1/2} \leq C \sum_{t>\nu} [E\{|\sigma_t(\theta)^2 - \widetilde{\sigma}_t(\theta)^2|\}]^{1/2} \rightarrow 0.$$

This completes the proof. ■

Lemma 2. As $n \rightarrow \infty$, $(n - \nu)^{-1} \sum_{\nu < t \leq n} E|X_t \{\widetilde{\sigma}_t(\widehat{\theta}) - \widetilde{\sigma}_t(\theta)\} I(A_n)| \rightarrow 0$.

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

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

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

On the other hand,

$$\begin{aligned} & \frac{1}{n-\nu} \sum_{t=\nu+1}^n E\{|X_t| |\check{\sigma}_t(\hat{\theta}) - \check{\sigma}_t(\theta)| I(A_n)\} \leq \frac{C}{n-\nu} \sum_{t=\nu+1}^n [E\{|\check{\sigma}_t(\hat{\theta}) - \check{\sigma}_t(\theta)|^2 I(A_n)\}]^{1/2} \\ & \leq \frac{C}{n-\nu} \sum_{t=\nu+1}^n [E\{\check{\sigma}_t(\hat{\theta})^2 - \check{\sigma}_t(\theta)^2 I(A_n)\}]^{1/2} \leq C_1 \{E(X_t^2)\}^{1/2} n^{-\kappa_1/2} n^{\kappa_1/4} \rightarrow 0. \end{aligned}$$

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

Lemma 3. For any given constant x ,

- (i) $\sup_{0 \leq x \leq 1} \frac{1}{n-\nu} \sum_{\nu < t \leq n} |I(\hat{U}_t \leq x) - I(U_t \leq x)| \xrightarrow{P} 0$, and
- (ii) $\sup_{0 \leq x \leq 1} \frac{1}{n-\nu} \sum_{\nu < t \leq n} |I(\hat{V}_t \leq x) - I(V_t \leq x)| \xrightarrow{P} 0$.

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

$$|\hat{U}_t - U_t| \leq C|\hat{\varepsilon}_t - \varepsilon_t| \leq C_1|X_t| |\check{\sigma}_t(\hat{\theta}) - \sigma_t(\theta)| / \check{\sigma}_t(\hat{\theta}).$$

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

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

Note that

$$\begin{aligned} |I(\hat{U}_t \leq x) - I(U_t \leq x)| & \leq I(\hat{U}_t \leq x, U_t > x) + I(\hat{U}_t > x, U_t \leq x) \\ & \leq I(|\hat{U}_t - U_t| > \delta) + I(U_t \in [x - \delta, x + \delta]). \end{aligned}$$

Therefore

$$\begin{aligned} & \sup_x \frac{1}{n-\nu} \sum_{t=\nu+1}^n |I(\hat{U}_t \leq x) - I(U_t \leq x)| \leq \frac{1}{n-\nu} \sum_{t=\nu+1}^n I(|\hat{U}_t - U_t| > \delta, A_n) \quad (\text{A.4}) \\ & + \sup_x \frac{1}{n-\nu} \sum_{t=\nu+1}^n I(U_t \in [x - \delta, x + \delta]) + I(A_n^c). \end{aligned}$$

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

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

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,

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

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

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

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

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

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

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

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

Proof of Theorem 1. Let $F_{n,1}$ and $F_{n,2}$ be, respectively, the empirical distribution of $\{U_t, \nu < t \leq n\}$ and $\{V_t, \nu < t \leq n\}$. By Lemma 3 and the Glivenko-Cantelli Theorem, it holds that

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

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

REFERENCES

BERKES, I., HORVÁTH, L. AND KOKOSZKA, P. (2003). GARCH processes: structure and estimation. *Bernoulli* **9**, 201–227.

- FAN, J. & YAO, Q. (2003) *Nonlinear Time Series: Nonparametric and Parametric Methods*. Springer, New York.
- GIRAITIS, L. & ROBINSON, P.M. (2001). Whittle estimation of ARCH models. *Econometric Theory* **17**, 608-23.
- HALL, P. & YAO, Q. (2003). Inference in ARCH and GARCH models with heavy-tailed errors. *Econometrica* **71**, 285-317.
- HORVATH, L. & LIESE, F. (2004). L_p -estimators in ARCH models. *Journal of Statistical Planning and Inference* **119**, 277-309.
- Konishi, S. & Kitagawa, G. (1996). Generalized information criteria in model selection. *Biometrika* **83**, 875-890.
- LEE, A.W. & HANSEN, B.E. (1994). Asymptotic theory for 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 for IGARCH(1,1) and covariance stationary GARCH(1,1) models. *Econometrica* **16**, 575-96.
- MIKOSCH, T. AND STRAUMANN, T. (2003). Stable limits of martingale transforms with application to the estimation of GARCH parameters. *A preprint*.
- PAN, J., TONG, H. & YAO, Q. (2006). Absolute Deviations Estimation for GARCH-type Models. (*Submitted.*)
- PENG, L. AND YAO, Q. (2003). Least absolute deviations estimation for ARCH and GARCH models. *Biometrika* **90**, 967-975.
- STRAUMANN, D. (2005). *Estimation in Conditionally Heteroscedastic Time Series Models*. Springer, Heidelberg.
- STRAUMANN, D. AND MIKOSCH, T. (2003). Quasi-mle in heteroscedastic times series: a stochastic recurrence equations approach. *A preprint*.
- 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.