ANNALS OF ECONOMICS AND FINANCE **5**, 283–298 (2004)

A Heavy-Tailed Distribution for ARCH Residuals with Application to Volatility Prediction

Dimitris N. Politis*

Department of Mathematics University of California—San Diego
La Jolla, CA 92093-0112, USA
E-mail: dpolitis@ucsd.edu

The quest for the 'best' heavy-tailed distribution for ARCH/GARCH residuals appears to still be ongoing. In this connection, we propose a new distribution that arises in a natural way as an outcome of an implicit model. The challenging application of prediction of squared returns is also discussed; an optimal predictor is formulated, and the usefulness of the new distribution for prediction is demonstrated on three real datasets. © 2004 Peking University Press

Key Words: Heteroscedasticity; Kyrtosis; Maximum likelihood; Time series.
JEL Classification Numbers: C3; C5.

1. INTRODUCTION

Consider data $X_1, \ldots, X_n$ arising as an observed stretch from a financial returns time series $\{X_t, t = 0, \pm 1, \pm 2, \ldots\}$ such as the percentage returns of a stock price, stock index or foreign exchange rate. The returns series $\{X_t\}$ will be assumed strictly stationary with mean zero which—from a practical point of view—implies that trends and other nonstationarities have been successfully removed.

The celebrated ARCH models of Engle (1982) were designed to capture the phenomenon of volatility clustering in the returns series. An ARCH(p) model can be described by the following equation:

$$X_t = Z_t \sqrt{a + \sum_{i=1}^{p} a_i X_{t-i}^2}. \tag{1}$$

* Research partially supported by NSF grant DMS-01-04059 and by the grant SES-0418136 funded jointly by the Economics and Statistics Programs of the National Science Foundation.

1529-7373/2004

The original assumption was that the series $\{Z_t\}$ is i.i.d. $N(0,1)$. Nevertheless, it was soon observed that the residuals, say $\{\hat{Z}_t\}$, from a fitted ARCH(p) model do not appear to be in accordance to the normality assumption as they are typically heavy-tailed.

Consequently, practitioners have been resorting to ARCH models with heavy-tailed errors. A popular assumption for the distribution of the $\{Z_t\}$ is the t-distribution with degrees of freedom empirically chosen to match the apparent degree of heavy tails as measured by higher-order moments such as the kyrtosis; see e.g. Bollerslev et al. (1992) or Shephard (1996) and the references therein.

Nevertheless, this situation is not very satisfactory since the choice of a t-distribution seems quite arbitrary, and the same is true for other popular heavy-tailed distributions, e.g. the double exponential. In the next section, an implicit ARCH model is developed that gives motivation towards a more 'natural'—and less *ad hoc*—distribution for ARCH/GARCH residuals. The precise definition of this new distribution is given in Section 3, together with some of its properties. The subject of maximum likelihood estimation for ARCH and GARCH models is addressed in Section 4. In Section 5, the problem of prediction of squared returns with ARCH/GARCH models is discussed, and an optimal predictor is suggested. Finally, Section 6 gives an application of volatility prediction in three datasets of interest.

2. AN IMPLICIT ARCH MODEL

Under model (1), the residuals

$$\hat{Z}_t = \frac{X_t}{\sqrt{\hat{a} + \sum_{i=1}^{p} \hat{a}_i X_{t-i}^2}} \tag{2}$$

ought to behave like i.i.d. standard normal random variables under the original ARCH assumptions; in the above, $\hat{a}, \hat{a}_1, \hat{a}_2, \ldots$ are estimates of the nonnegative parameters $a, a_1, a_2, \ldots$.

The degree of non-conformance to the normality assumption can be captured in many ways; the easiest is to compute the empirical kyrtosis and compare to the normal kyrtosis of 3. So let $K_i^j(Y)$ denote the empirical (sample) kyrtosis of the dataset $\{Y_i, Y_{i+1}, \ldots, Y_j\}$. Typically, $K_1^n(\hat{Z})$ is quite smaller than $K_1^n(X)$ but still quite bigger than 3, i.e., $3 < K_1^n(\hat{Z}) < K_1^n(X)$.

Note that, given the data, $K_1^n(\hat{Z})$ is a continuous function of $\hat{a}, \hat{a}_1, \hat{a}_2, \ldots$. The question may then be asked: is there a specification for $\hat{a}, \hat{a}_1, \hat{a}_2, \ldots$ that will make the kyrtosis $K_1^n(\hat{Z})$ of the residuals to be about 3? The answer is not in general.

Taking another look at the ratio given in eq. (2) we may interpret it as an attempt to 'studentize' the return X_t by dividing with a (time-localized) measure of standard deviation. Nevertheless, there seems to be no reason to exclude the value of X_t from an empirical (causal) estimate of the standard deviation of the same X_t. Thus, if we are to include an X_t^2 term (with its own coefficient, say $a_0 \geq 0$) in the studentization, we may define the new empirical ratio

$$\hat{W}_t = \frac{X_t}{\sqrt{\hat{a} + \hat{a}_0 X_t^2 + \sum_{i=1}^{p} \hat{a}_i X_{t-i}^2}} \tag{3}$$

that may be associated with a corresponding true equation of the type:

$$W_t = \frac{X_t}{\sqrt{a + a_0 X_t^2 + \sum_{i=1}^{p} a_i X_{t-i}^2}}. \tag{4}$$

To repeat our question in the new set-up: is there a specification for $\hat{a}, \hat{a}_0, \hat{a}_1, \ldots$ that will make the kyrtosis $K_1^n(\hat{W})$ of the new residuals $\hat{W}_t$ to be about 3? The answer in general is *yes*!

To see why, note that the simple specification: $\hat{a} = 0$, $\hat{a}_0 = 1$ and $\hat{a}_j = 0$ for $j \geq 1$ results into $\hat{W}_t = sign(X_t)$, in which case $K_1^n(\hat{W}) = 1$. But as mentioned before, a specification with $\hat{a}_0 = 0$ typically results into $K_1^n(\hat{W}) > 3$. Therefore, by the smoothness of $K_1^n(\hat{W})$ as a function of $\hat{a}, \hat{a}_1, \hat{a}_2, \ldots$, the intermediate value theorem guarantees the existence of a specification with $K_1^n(\hat{W}) = 3$.

Having residuals $\hat{W}_t$ that have kyrtosis equal to 3—as well as an approximately symmetric[1] distribution about zero—it is natural to assume that the true W_t in equation (4) follow a mean zero normal distribution—at least approximately. Furthermore, by proper re-scaling of the parameters $a, a_0, a_1, \ldots$, we may even assume that the W_t approximately follow a *standard* normal distribution

We can now re-arrange equation (4) to make it look more like model (1):

$$X_t = W_t \sqrt{a + a_0 X_t^2 + \sum_{i=1}^{p} a_i X_{t-i}^2}. \tag{5}$$

Equation (5) represents an *implicit* ARCH model; the reason for the name 'implicit' is that the term X_t appears on both sides of the equation. Nev-

[1] Some authors have raised the question of existence of skewness in financial returns; see e.g. Patton (2002) and the references therein. Nevertheless, at least as a first approximation, the assumption of symmetry is very useful for model building.

ertheless, we can solve equation (5) for X_t, to give:

$$X_t = U_t \sqrt{a + \sum_{i=1}^{p} a_i X_{t-i}^2} \tag{6}$$

where

$$U_t = \frac{W_t}{\sqrt{1 - a_0 W_t^2}}. \tag{7}$$

Interestingly, the implicit ARCH model (5) is seen to be tantamount to the regular ARCH(p) model (6) associated with the new innovation term U_t.

However, it is now apparent that exact normality may not hold for the W_t for then the denominator of (7) would become imaginary. As a matter of fact, both $\hat{W}_t$ and W_t are bounded; to see this, note that

$$\frac{1}{W_t^2} = \frac{a + a_0 X_t^2 + \sum_{i=1}^{p} a_i X_{t-i}^2}{X_t^2} \geq a_0.$$

Thus, $|W_t| \leq 1/\sqrt{a_0}$, and similarly $|\hat{W}_t| \leq 1/\sqrt{\hat{a}_0}$ almost surely.

A natural way to model a situation where the W_t are thought to be close to $N(0,1)$ but happen to be bounded is to use a *truncated* standard normal distribution, i.e., to assume that the W_t are i.i.d. with probability density given by

$$\frac{\phi(x)\mathbf{1}\{|x| \leq C_0\}}{\int_{-C_0}^{C_0} \phi(y)dy} \quad \text{for all} \quad x \in \mathbf{R} \tag{8}$$

where ϕ denotes the standard normal density, and $C_0 = 1/\sqrt{a_0}$. With a_0 small enough, the boundedness of W_t is effectively not noticeable but yields interesting implications for the distribution of the U_t defined in (7) as detailed in the following section.

3. A HEAVY-TAILED DISTRIBUTION FOR ARCH RESIDUALS

To summarize the discussion of Section 2, the newly derived implicit ARCH model consists of eq. (5) together with the assumption that the $\{W_t\}$ series is i.i.d. with density given by eq. (8).

However, if W_t is assumed to follow the truncated standard normal distribution (8), then the change of variable (7) implies that the innovation term U_t appearing in the ARCH model (6) has the density $f(u; a_0, 1)$ defined as:

$$f(u; a_0, 1) = \frac{(1 + a_0 u^2)^{-3/2} \exp(-\frac{u^2}{2(1+a_0 u^2)})}{\sqrt{2\pi}\left(\Phi(1/\sqrt{a_0}) - \Phi(-1/\sqrt{a_0})\right)} \quad \text{for all} \quad u \in \mathbf{R} \tag{9}$$

where Φ denotes the standard normal distribution function. Eq. (9) describes our proposed density for the ARCH residuals. The nonnegative parameter a_0 is a shape parameter having to do with the degree of heavy tails; note that $f(u; a_0, 1) \to \phi(u)$ as $a_0 \to 0$.

It is apparent that $f(u; a_0, 1)$ has heavy tails. Except for the extreme case where $a_0 = 0$ where all moments are finite, in general moments are finite only up to (almost) order two. In other words, if a random variable U follows the density $f(u; a_0, 1)$ with $a_0 > 0$, then it is easy to see that

$$E|U|^d < \infty \quad \text{for all} \quad d \in [0, 2) \quad \text{but} \quad E|U|^d = \infty \quad \text{for all} \quad d \in [2, \infty). \tag{10}$$

The above property is reminiscent of the t_2 distribution, i.e., Student's t distribution with 2 degrees of freedom; this is no coincidence in the sense that

$$f(u; a_0, 1) \sim c(a_0)(1 + a_0 u^2)^{-3/2} \quad \text{as} \quad u \to \infty \tag{11}$$

where $1/c(a_0) = \sqrt{2\pi}\left(\Phi(1/\sqrt{a_0}) - \Phi(-1/\sqrt{a_0})\right)\exp(1/(2a_0))$.

Eq. (11) shows that the rate by which $f(u; a_0, 1)$ tends to 0 as $u \to \infty$ is the same as in the t_2 case. Nonetheless, the tails of $f(u; a_0, 1)$ are quite lighter than those of the t_2 distribution as the constants associated with those rates are very different; in particular, the constant $c(a_0)$ is much smaller. In some sense, $f(u; a_0, 1)$ achieves its degree of heavy tails in a subtler way.

TABLE 1.

Truncated moments of the $f(u; 0.1, 1)$ density as compared to those of the f_{t_2} and f_{t_5}, i.e., the densities of the t_2 and t_5 distributions.

$a =$	1	1.9	2	2.1	3	4
$\int_{-10}^{10} \lvert u\rvert^a f(u; 0.1, 1) du$	0.905	1.444	1.561	1.695	4.401	18.74
$\int_{-100}^{100} \lvert u\rvert^a f(u; 0.1, 1) du$	0.923	1.745	1.983	2.290	20.27	875.45
$\int_{-10}^{10} \lvert u\rvert^a f_{t_2}(u) du$	1.216	2.931	3.328	3.798	14.94	89.03
$\int_{-100}^{100} \lvert u\rvert^a f_{t_2}(u) du$	1.394	6.176	7.904	10.278	194.40	9975.3
$\int_{-10}^{10} \lvert u\rvert^a f_{t_5}(u) du$	0.947	1.519	1.638	1.773	4.304	15.96
$\int_{-100}^{100} \lvert u\rvert^a f_{t_5}(u) du$	0.949	1.541	1.667	1.811	4.740	24.05

To elaborate on the latter point, Table 1 gives some moments of the $f(u; 0.1, 1)$ density truncated to either ±10 or ±100, and comparing them to the respective moments of the (truncated) f_{t_2} and f_{t_5}, i.e., the densities of the t_2 and t_5 distributions. It is apparent that up to moments of order 2 (and perhaps even order 2.1), the moments of $f(u; 0.1, 1)$ are close to those of f_{t_5}. By contrast, for moments of orders 3 and 4 the similarity with f_{t_5} breaks down; at the same time, the lighter tails of $f(u; 0.1, 1)$ as compared to those of f_{t_2} are quite clear.

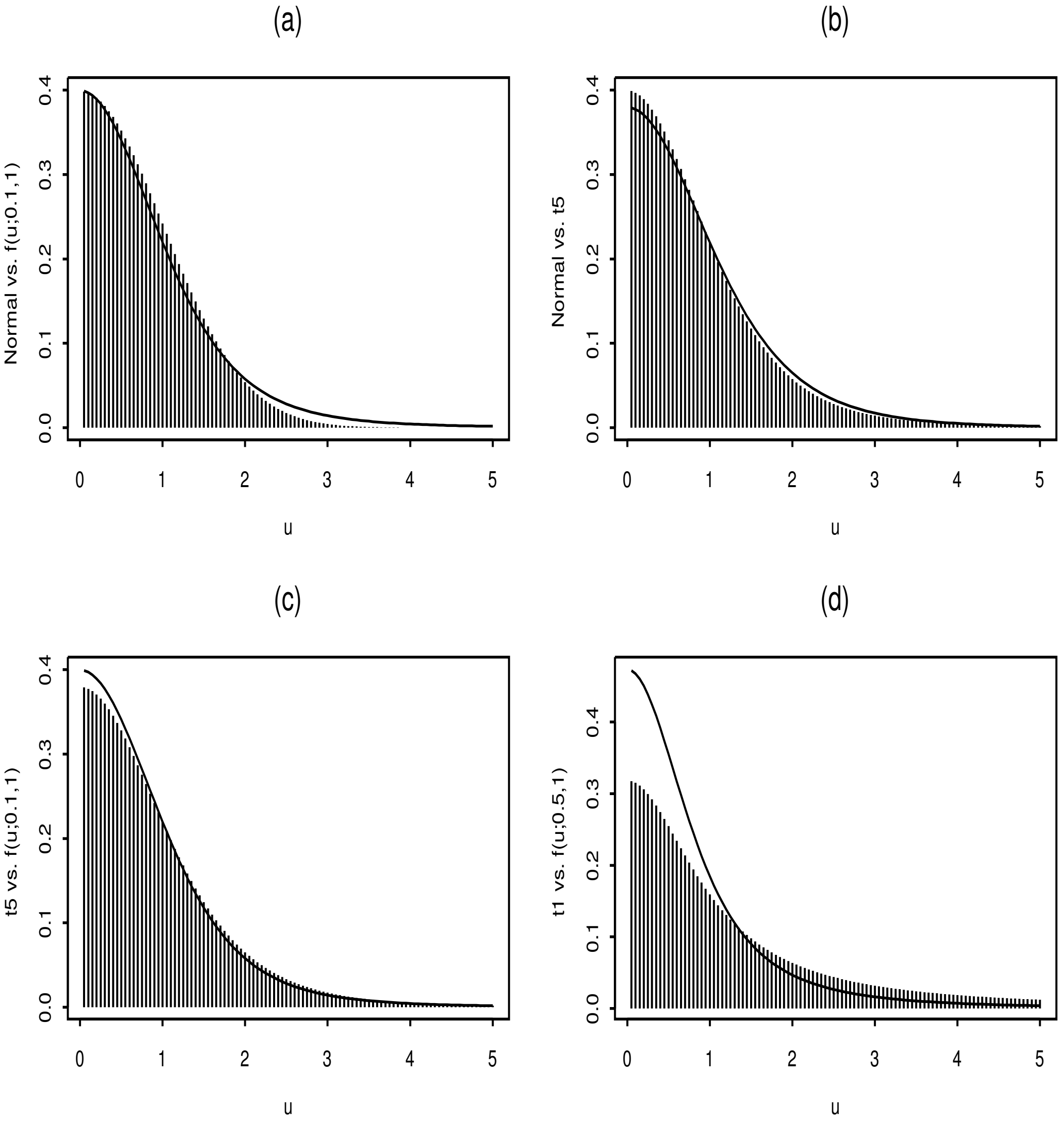

FIG. 1. (a) Standard normal density (shaded) vs. $f(u;0.1,1)$; (b) Standard normal density (shaded) vs. t with 5 degrees of freedom; (c) t with 5 degrees of freedom (shaded) vs. $f(u;0.1,1)$; (d) t with 1 degree of freedom (shaded) vs. $f(u;0.5,1)$.

Plots of the (right half of the) density $f(u; a_0, 1)$ are pictured in Figure 1 for $a_0 = 0.1$ and 0.5; they are compared to the standard normal as well as the t_5, i.e., t distribution with 5 degrees of freedom. Figures 1 (a)-(c) focus on the $f(u; 0.1, 1)$ and the t_5 since values of a_0 about 0.1 and degrees of freedom of the order of 5 seem to be typical in connection with ARCH residuals in practice; see e.g. our Section 6.

However, Figure (d) shows what happens when the degree of heavy tails is cranked up in both families, the $f(u; a_0, 1)$ and the t. The differences are quite apparent; for example, as the degree of heavy tails increases—i.e., a_0 increases in $f(u; a_0, 1)$ and degrees of freedom decreases in the t distribution—the two densities change in opposite ways around the origin: $f(0; a_0, 1)$ is increasing, while the t density is decreasing. Thus, it could be said that $f(u; a_0, 1)$ is more 'leptokurtic' than the t-family in the sense that $f(u; a_0, 1)$ becomes even more concentrated near the origin when its degree of heavy tails increases whereas, at the same time, as $u \to \infty$, $f(u; a_0, 1)$ tends to zero with slower rate than any t_d distribution with $d > 2$.

4. MAXIMUM LIKELIHOOD

In this section we consider fitting the ARCH model (6) to our data $X_1, \ldots, X_n$ under the assumption that $U_1, \ldots, U_n$ are i.i.d. according to the proposed new density $f(u; a_0, 1)$. Note that we can scale the density $f(u; a_0, 1)$ to create a two-parameter family of densities with typical member given by

$$f(x; a_0, s) = \frac{1}{s} f(\frac{x}{s}; a_0, 1) \quad \text{for all} \quad x \in \mathbf{R}. \tag{12}$$

As before, the parameter a_0 is a shape parameter, while the positive parameter s represents scale.

Consequently, for any $t > p$, the density of X_t conditionally on the observed past $\mathcal{F}_{t-1} = \{X_s, 0 < s \leq t-1\}$ is given by $f(x; a_0, s_t)$, where the volatility $s_t = \sqrt{a + \sum_{i=1}^{p} a_i X_{t-i}^2}$ is treated as constant given $\mathcal{F}_{t-1}$. Thus, the likelihood of the data $\mathbf{X} = (X_1, \ldots, X_n)$ conditionally on $\mathcal{F}_p$ (also called the 'pseudo-likelihood') is given by:

$$L(a, a_0, a_1, \ldots, a_p | \mathbf{X}) = \prod_{t=p+1}^{n} f(X_t; a_0, s_t). \tag{13}$$

As usual, define the maximum (pseudo)likelihood estimators $\hat{a}, \hat{a}_0, \hat{a}_1, \ldots, \hat{a}_p$ as the values of $a, a_0, a_1, \ldots, a_p$ that maximize $L(a, a_0, a_1, \ldots, a_p | \mathbf{X})$ subject to the nonnegativity constraints: $a \geq 0$ and $a_i \geq 0$ for all $i \geq 0$. The maximum (pseudo)likelihood estimators generally partake in the favorable properties shared by classical maximum likelihood estimators (MLE); see e.g. Gouriéroux (1997). In addition, the maximum (pseudo)likelihood estimators have recently been shown to be consistent even in the absence of finite fourth moments although in that case their rate of convergence is slower than $\sqrt{n}$; see Hall and Yao (2003). For simplicity, we will refer to the maximum (pseudo)likelihood estimators $\hat{a}, \hat{a}_0, \hat{a}_1, \ldots, \hat{a}_p$ as the MLEs in the ARCH case with $f(u; a_0, 1)$ residuals; note, however, that this maximization must be done numerically as no closed-form expressions for the MLEs seem to be available.

Now, and in the remainder of the paper, we will focus on Bollerslev's (1996) popular GARCH(1,1) model that has been shown to achieve a most parsimonious fit. Therefore, let

$$X_t = s_t U_t \quad \text{with} \quad s_t^2 = C + AX_{t-1}^2 + Bs_{t-1}^2 \tag{14}$$

and

$$U_t \sim \text{i.i.d. } f(u; a_0, 1) \tag{15}$$

where the nonnegative parameters A, B, C satisfy the weak-stationarity condition $A + B < 1$. All-in-all, the above GARCH(1,1) model given by (14)–(15) has four parameters:[2] A, B, C and a_0.

Back-solving in eq. (15) it is easy to see that the GARCH model (14) is tantamount to the ARCH model (6) with $p = \infty$ and the following identifications:

$$a = \frac{C}{1-B}, \quad \text{and} \quad a_i = AB^{i-1} \quad \text{for} \quad i = 1, 2, \ldots \tag{16}$$

Not surprisingly, the parameter a_0 does not figure in at all in eq. (16) as it is solely associated with the distribution of the errors U_t.

While it is difficult to write down exactly the (pseudo)likelihood in the GARCH case, it is easy to get an approximation. The most straightforward such approximation is to note that the exponential decay of a_i given in eq. (16) implies that $a_i \simeq 0$ for all $i \geq$ some *finite* value p_0. In this sense, the GARCH(1,1) model (14) is *approximately* equivalent to the ARCH model (6) with $p = p_0$. The MLEs of A, B, C and a_0 can then be obtained by maximizing $L(a, a_0, a_1, \ldots, a_{p_0} | \mathbf{X})$ of eq. (13) with respect to the four free parameters a_0, A, B, C, noting that $a, a_1, \ldots, a_{p_0}$ are simple functions of A, B, C by (16).

As in all numerical optimization problems, having good starting values significantly speeds up the search, and reduces the risk of finding local—but

[2] The same number of parameters (four) characterizes the GARCH (1,1) model with t–errors; the number of degrees of freedom for the best-fitting t distribution represents the fourth parameter.

not global—optimizers; to further address the latter risk, the optimization should be run a few times with different starting values each time. More practical details are given in the Section 6.

5. PREDICTION OF SQUARED RETURNS WITH ARCH/GARCH MODELS

The litmus test of any model is its predictive ability. Although ARCH models could not be expected to successfully predict the (signed) returns X_t, they are indeed expected to have some predictive ability for the squared returns X_t^2; see the discussion in the Introduction.

Nevertheless, the literature abounds with suggestions to the contrary. In particular, it is widely believed that ARCH/GARCH models are characterized by "poor out-of-sample forecasting performance vis-a-vis daily squared returns"; see Andersen and Bollerslev (1998) "numerous studies have suggested that ARCH and stochastic volatility models provide poor volatility forecasts".

It seems, however, that these negative comments have more to do with the commonly employed prediction method that seems suboptimal, namely predicting X_t^2 by the (estimated) squared volatility $\hat{s}_t^2 = \hat{C}+\hat{A}X_{t-1}^2+\hat{B}s_{t-1}^2$ where $\hat{A}, \hat{B}, \hat{C}$ are the MLEs in the GARCH model (14).

Using $\hat{s}_t^2$ as predictor for X_t^2 would be optimal if: (a) the GARCH residuals were normal $N(0,1)$; (b) Mean Squared Error (MSE) was used to measure the quality of prediction; and (c) the returns X_t had a finite fourth moment. If conditions (a),(b),(c) were to hold true, then s_t^2 would represent the conditional mean of X_t^2 given the past $\mathcal{F}_{t-1} = \{X_i, 1 \leq i \leq t-1\}$ which is the optimal (with respect to MSE) predictor of X_t^2; since $\hat{s}_t^2$ is our best proxy for s_t^2, the use of $\hat{s}_t^2$ as predictor would then be justified. However, the predictor $\hat{s}_t^2$ seems to perform similarly—and sometimes even a bit worse—as compared to the crudest possible predictor, namely the sample variance of dataset $\{X_i, 1 \leq i \leq t-1\}$, thus giving rise to the aforementioned criticisms.

The poor performance of $\hat{s}_t^2$ is not necessarily evidence against the GARCH model (14); rather, it may be seen as evidence that one or more amongst conditions (a),(b),(c) are not true. As a matter of fact, arguments against condition (a) abound as mentioned in our Introduction; see e.g. Bollerslev et al. (1992) or Shephard (1996) and the references therein. Noting that condition (b) is contingent on condition (c) we now focus on the latter.

Let $V_i^j(Y)$ and $K_i^j(Y)$ denote the empirical (sample) variance and kyrtosis (respectively) of a general dataset $\{Y_i, Y_{i+1}, \ldots, Y_j\}$. Figure 2 shows a plot of $V_1^k(X)$ and $K_1^k(X)$ as a function of k with data $X_1, X_2, \ldots$ representing daily returns of the S&P500 index spanning the period 1-1-1928 to 8-30-1991. The plot of $V_1^k(X)$ indicates convergence as k increases, giving

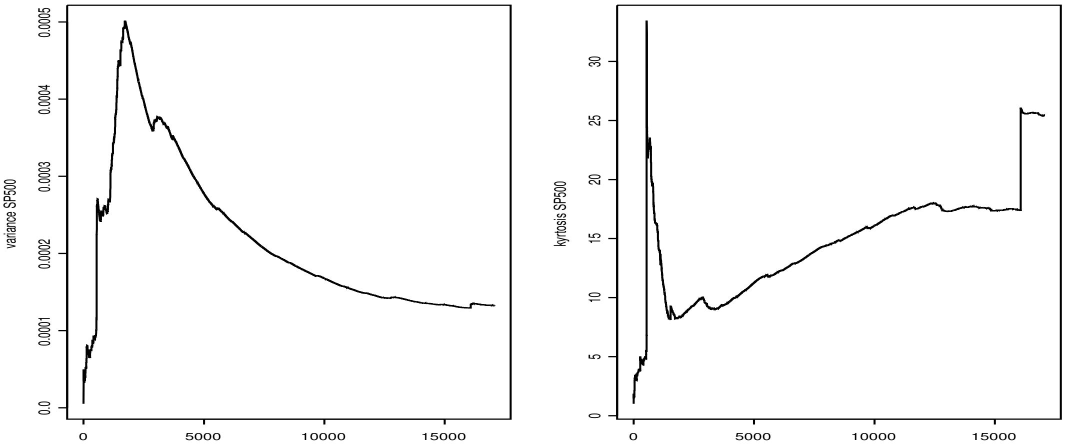

FIG. 2. Plot of $V_1^k(X)$ and $K_1^k(X)$ as a function of k; the data $X_1, X_2, \ldots$ represent daily returns of the S&P500 index spanning the period 1-1-1928 to 8-30-1991.

empirical evidence that the Strong Law of Large Numbers (SLLN) may be kicking in; the implication is that the S&P500 data may have a finite 2nd moment. On the contrary, the plot of $K_1^k(X)$ indicates divergence as k increases with the implication that the S&P500 data may *not* have a finite 4th moment.[3] Similar conclusions can be drawn using different datasets, e.g. foreign exchange rates, etc., provided the records are long enough. Hence, condition (c) seems to fail.

Thus, the failure of predictor $\hat{s}_t^2$ is justified due to the failure of conditions (a),(b),(c) that need to be modified as follows: (a′) the GARCH residuals in model (14) follow a (possibly) heavy-tailed distribution; (b′) an L_1 measure such as Mean Absolute Deviation (MAD) is used to measure the quality

[3] The point may be made that returns are ‘physically’ bounded, and hence all moments are finite. The returns are certainly bounded from below by the value -1, so the assumption of symmetry would go hand-in-hand with the boundedness assumption. Interestingly, the largest outliers ever recorded are in the negative direction, e.g. the approximately -0.2 return associated with the crash of 1987, indicating that the lower bound of -1 is really too far away to have any real import. But even adopting the viewpoint that returns are bounded, Figure 2 suggests that the 2nd moment may have a moderate value while the 4th moment is (at least) 10,000 times as large; this phenomenon may be compared with the truncation effect in Table 1: having the 4th moment equal 500 or 1,000 times the 2nd moment is tantamount (and practically indistinguishable) to having an infinite 4th moment.

of prediction; and (c′) the returns X_t have an infinite fourth moment but a finite second moment—or, at least, an 'almost' finite second moment.[4]

Under conditions (a′),(b′),(c′), the optimal predictor of X_t^2 given the past $\mathcal{F}_{t-1} = \{X_i, 1 \leq i \leq t-1\}$ is given by

$$m_2 \cdot \hat{s}_t^2 \tag{17}$$

where m_2 is the median of the (common) distribution of U_t^2. For example, $m_2 \simeq 0.455$ if $U_t \sim N(0,1)$, while $m_2 \simeq 0.528$ if $U_t \sim t_5$.

Under condition (c′), it is also possible to assume the $f(u; a_0, 1)$ distribution for the GARCH residuals, i.e., to assume model (14) together with (15). Table 2 below contains approximate values for m_2 in the case $U_t \sim f(u; a_0, 1)$ for different values of the shape parameter a_0.

TABLE 2.

Approximate values for m_1, the median of the distribution of $|U_t|$, and m_2, the median of the distribution of U_t^2, in the case $U_t \sim f(u; a_0, 1)$ for different values of the shape parameter a_0.

a_0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10
m_1	0.676	0.677	0.679	0.681	0.682	0.684	0.685	0.687	0.688	0.670
m_2	0.457	0.459	0.461	0.463	0.465	0.467	0.469	0.471	0.473	0.475

A different problem of interest is prediction of $|X_t|$ given the past $\mathcal{F}_{t-1}$. It is easy to see that the optimal predictor of $|X_t|$ with respect to MAD, i.e., L_1 loss, is given by $m_1 \hat{s}_t$, where m_1 is the median of the distribution of $|U_t|$. Note, however, that conditions (a′), (c′) afford us now the possibility of adopting an L_2 loss; in that case, the optimal predictor of $|X_t|$ with respect to MSE is given by $\mu_1 \hat{s}_t$, where μ_1 is the mean of the distribution of $|U_t|$.

Nevertheless, L_1 loss seems preferable as the MAD—which is its empirical version—is both more stable, as well as more easily interpretable. Table 2 gives approximate values for m_1 in the case $U_t \sim f(u; a_0, 1)$; the values of m_1 in case U_t follows the $N(0,1)$ or t_5 distribution are $m_1 = 0.674$ and 0.727 respectively. For concreteness, in what follows we focus exclusively on predicting the squared returns X_t^2.

[4] By 'almost' finite second moment, a condition like eq. (10) is implied. Note that, using finite-sample data such as those in Figure 2, one could never reject the hypothesis that the returns have 'almost' finite second moment only.

6. PREDICTION OF SQUARED RETURNS: THREE EXAMPLES

To evaluate and compare the predictive ability of the GARCH (1,1) model with different distributional assumptions on the errors, we focus on three well-known datasets.

- (Foreign exchange). Daily returns of the Yen vs. Dollar exchange rate from January 1, 1988 to August 1, 2002; the sample size is 3600 (weekends and holidays are excluded).
- (Stock index). Daily returns of the S&P500 stock index from October 1, 1983 to August 30, 1991; the sample size is 2000.
- (Stock price). Daily returns of the IBM stock price from February 1, 1984 to December 31, 1991; the sample size is 2000.

The Yen/Dollar data were downloaded from Datastream; the other two datasets are available as part of the `garch` module of the statistical language S+.

Table 3 shows the MLEs in the GARCH (1,1) model under three possible distributional assumptions for the GARCH errors, namely the $N(0,1)$, the t distribution (with estimated degrees of freedom), and the new $f(\cdot; a_0, 1)$ density. The computations were carried out in S+; the GARCH models associated with the first two distributions were fitted using the `garch` module, while the numerical optimization[5] for the case of the $f(\cdot; a_0, 1)$ density was performed using the function `nlminb`. Notably, in all three datasets, the degrees of freedom for the t distribution were estimated to be 5.

For the particular problem of numerical MLE under the density $f(\cdot; a_0, 1)$, good starting values for A, B, C are provided by the MLEs obtained using the aforementioned t_5 distribution for the GARCH residuals. As a matter of fact, as Table 3 shows, the actual MLEs associated with the $f(\cdot; a_0, 1)$ distribution for the residuals turn out to be remarkably close to those starting values. Perhaps of some interest is that the sum $\hat{A} + \hat{B}$ seems to consistenly take higher values under the $f(\cdot; a_0, 1)$ distribution as compared to the t_5.

Regarding $\hat{a}_0$, any number in the interval [0.07, 0.10] is a good starting value with a value around 0.08 probably being best. Recall that the truncation level for the quasi-normal residuals $\hat{W}_t$ of Section 2 is $1/\sqrt{\hat{a}_0}$. As Table 3 suggests, this number ranges from 3.35 to 3.86. Since 99.7% of the mass of the $N(0,1)$ distribution lies within ±3 anyway, this truncation does not practically spoil the normality of the $\hat{W}_t$ residuals.

In order to evaluate the out-of-sample performance of different predictors of squared returns the following procedure was implemented: the first half of each of our three datasets was used to get estimates of the GARCH coef-

[5] Some simple S+ functions associated with numerical MLE and Monte Carlo under the assumption of density $f(\cdot; a_0, 1)$ are available from: www.math.ucsd.edu/∼politis

TABLE 3.

Maximum (pseudo)likelihood estimators in the GARCH (1,1) model in the three datasets, and under three possible distributional assumptions for the GARCH errors: the $N(0,1)$, the t distribution (with estimated degrees of freedom), and the new $f(\cdot;a_0,1)$ density.

	$\hat{a}_0$	$\hat{A}$	$\hat{B}$	$\hat{C}$
Yen/Dollar–$N(0,1)$	N/A	0.062	0.898	2.29e-06
Yen/Dollar–t distr.	N/A	0.027	0.923	8.95e-07
Yen/Dollar–$f(\cdot;a_0,1)$	0.089	0.028	0.938	8.38e-07
S&P500–$N(0,1)$	N/A	0.104	0.834	6.63e-06
S&P500–t distr.	N/A	0.022	0.927	1.83e-06
S&P500–$f(\cdot;a_0,1)$	0.081	0.023	0.936	1.96e-06
IBM–$N(0,1)$	N/A	0.104	0.807	1.72e-05
IBM–t distr.	N/A	0.027	0.913	5.65e-06
IBM–$f(\cdot;a_0,1)$	0.066	0.029	0.912	6.32e-06

ficients (including a_0), while the prediction of squared returns was carried out over the second half. Table 4a tabulates the relative performance—as measured by the Mean Absolute Deviation (MAD)—of three predictors: the benchmark, the simple $\hat{s}_t^2$, and the optimal $m_2\hat{s}_t^2$. The benchmark amounts to the aforementioned crudest predictor, i.e., the sample variance of dataset $\{X_i, 1 \leq i \leq t-1\}$. The values of m_2 used in the $f(\cdot;a_0,1)$ case were obtained from Table 2 based on the estimated value for a_0; for the t distribution, the m_2 associated with t_5 was used.

It is apparent from Table 4a, that the simple predictor $\hat{s}_t^2$ seems to actually have some predictive ability, i.e., to improve upon the crude benchmark, when a heavy-tailed distribution—t or $f(\cdot;a_0,1)$—is assumed for the GARCH residuals. Detecting the presence of this predictive ability is solely due to using an L_1 measure to quantify the accuracy of prediction since, as mentioned before, the MSE of predictor $\hat{s}_t^2$ is generally comparable to that of the benchmark.

Also immediate from Table 4a is that the predictor $\hat{s}_t^2$ is always inferior to the optimal predictor $m_2\hat{s}_t^2$. Focusing on the latter, the best performance is achieved using the two heavy-tailed distributions, with the $f(\cdot;a_0,1)$ distribution having a slight edge over the t distribution in all three cases.

Note, however, that in practice the GARCH estimates would be updated daily, i.e., to predict X_t given the past $\mathcal{F}_{t-1}$, the GARCH coefficients would be estimated based on the whole of $\mathcal{F}_{t-1}$. Although it is quite feasible for a practitioner to devote 2-3 minutes daily to update those coefficients, it is unfeasible computationally to include this daily updating in our simulation.

TABLE 4a.

Entries represent the Mean Absolute Deviation (multiplied by 1,000) for the three predictors of squared returns: the benchmark, the simple $\hat{s}_t^2$, and the optimal $m_2\hat{s}_t^2$; for the last two, the GARCH(1,1) model (14) was used. The predictions were carried out over the 2nd half of each dataset, with coefficients estimated from the 1st half.

	benchmark	$\hat{s}_t^2$	$m_2\hat{s}_t^2$
Yen/Dollar–$N(0,1)$	0.0697	0.0646	0.0545
Yen/Dollar–t distr.	”	0.0550	0.0541
Yen/Dollar–$f(\cdot; a_0, 1)$	”	0.0567	0.0540
S&P500–$N(0,1)$	0.3343*	0.1042	0.0919
S&P500–t distr.	”	0.0947	0.0920
S&P500–$f(\cdot; a_0, 1)$	”	0.0975	0.0918
IBM–$N(0,1)$	0.1692	0.1918	0.1500
IBM–t distr.	”	0.1571	0.1455
IBM–$f(\cdot; a_0, 1)$	”	0.1609	0.1454

* This value is as high because the crash of 1987 is present in the 2nd half of the S&P500 dataset.

The result is that the entries of Table 4a are conservative in the sense that prediction performances would be expected to improve if the GARCH estimates were to be updated daily. To see the effect of having better GARCH estimates when the prediction is carried out we go to the other extreme: Table 4b shows the performances of our predictors carried out over the same 2nd half of each dataset but using GARCH coefficients estimated from the *whole* of the dataset, i.e., the coefficients from Table 3.

By contrast to the conservative entries of Table 4a, the entries of Table 4b are over-optimistic as the GARCH coefficients used have unrealistic accuracy; therefore, the truth should lie somewhere in-between Table 4a and Table 4b. Nevertheless, the two tables are similar enough to suggest that the effect of the accuracy of the GARCH coefficients is not so prominent, and Table 4b leads to the same conclusions as those gathered from Table 4a.

7. CONCLUSIONS

A new heavy-tailed density for ARCH/GARCH residuals was proposed in eq. (9), motivated by the development of an implicit ARCH model. The properties of the density $f(\cdot; a_0, 1)$ were studied, and the procedure for obtaining numerical MLEs was outlined.

TABLE 4b.

Entries represent the Mean Absolute Deviation (multiplied by 1,000) for the three predictors of squared returns. The predictions were carried out over the 2nd half of each dataset, with GARCH coefficients estimated from the *whole* of the dataset.

	benchmark	$\hat{s}_t^2$	$m_2\hat{s}_t^2$
Yen/Dollar–$N(0,1)$	0.0697	0.0650	0.0545
Yen/Dollar–t distr.	”	0.0554	0.0540
Yen/Dollar–$f(\cdot;a_0,1)$	”	0.0574	0.0539
S&P500–$N(0,1)$	0.3343	0.1135	0.0942
S&P500–t distr.	”	0.0948	0.0920
S&P500–$f(\cdot;a_0,1)$	”	0.0978	0.0919
IBM–$N(0,1)$	0.1692	0.1815	0.1472
IBM–t distr.	”	0.1545	0.1453
IBM–$f(\cdot;a_0,1)$	”	0.1577	0.1453

The challenging problem of prediction of squared returns was put in a rigorous framework, and the optimal predictor (17) was formulated. The usefulness of the optimal predictor was demonstrated on three real datasets.

By contrast to what is widely believed, it was found that ARCH/GARCH models *do* have predictive validity for the squared returns; this is particularly true when a heavy-tailed distribution is assumed for the GARCH residuals with the $f(\cdot;a_0,1)$ distribution appearing to have a slight edge over the popular t distribution. Notably, to appreciate and take advantage of this predictive ability one must: (a) use a more meaningful measure of prediction performance such as L_1 loss, and (b) use the optimal predictor which is given by (17) in the L_1 case.

REFERENCES

Andersen, T.G. and T. Bollerslev, 1998, Answering the sceptics: Yes, standard volatility models do provide accurate forecasts. *International Economic Review* **39, no.4**, 885-905.

Bollerslev, T., 1986, Generalized autoregressive conditional heteroscedasticity. *Journal of Econometrics* **31**, 307-327.

Bollerslev, T., Chou, R., and K. Kroner, 1992, ARCH modelling in finance: A review of theory and empirical evidence. *Journal of Econometrics* **52**, 5-60.

Engle, R., 1982, Autoregressive conditional heteroscedasticity with estimates of the variance of UK inflation. *Econometrica* **50**, 987-1008.

Gouriéroux, C., 1997, ARCH Models and Financial Applications. Springer Verlag, New York.

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

Patton, A.J., 2002, Skewness, asymmetric dependence, and portfolios. Preprint (can be downloaded from: http://fmg.lse.ac.uk/∼patton).

Shephard, N., 1996, Statistical aspects of ARCH and stochastic volatility. In: *Time Series Models in Econometrics, Finance, and Other Fields*. D.R. Cox, David, V. Hinkley, and Ole E. Barndorff-Nielsen (eds.) Chapman & Hall, London, pp. 1-67.