Autoregressive conditional heteroskedasticity and changes in regime

James D. Hamilton*\textsuperscript{,}a, Raul Susmelb

\textsuperscript{a}Department of Economics, University of California at San Diego, La Jolla, CA 92093-0508, USA
\textsuperscript{b}Department of Finance, University of Houston, Houston, TX 77004, USA

(Received August 1992; final version received September 1993)

Abstract

ARCH models often impute a lot of persistence to stock volatility and yet give relatively poor forecasts. One explanation is that extremely large shocks, such as the October 1987 crash, arise from quite different causes and have different consequences for subsequent volatility than do small shocks. We explore this possibility with U.S. weekly stock returns, allowing the parameters of an ARCH process to come from one of several different regimes, with transitions between regimes governed by an unobserved Markov chain. We estimate models with two to four regimes in which the latent innovations come from Gaussian and Student $t$ distributions.

Key words: ARCH models; Stock prices; Regime-switching models; Volatility

JEL classification: C22; G12

1. Introduction

Financial markets sometimes appear quite calm and at other times highly volatile. Describing how this volatility changes over time is important for two reasons. First, the riskiness of an asset is an important determinant of its price. Indeed, empirical estimation of the conditional variance of asset returns forms

One popular approach to modeling volatility is the autoregressive conditional heteroskedasticity (ARCH) specification introduced by Engle (1982). Bollerslev, Chou, and Kroner (1992) characterized this general class of models as follows. Suppose that a variable $u_t$ is governed by

$$u_t = \sigma_t \cdot v_t,$$

(1.1)

where $(v_t)$ is an i.i.d. sequence with zero mean and unit variance. The conditional variance of $u_t$ is specified to be a function of its past realizations:

$$\sigma_t^2 = g(u_{t-1}, u_{t-2}, \ldots).$$

(1.2)

Often it is assumed that $v_t \sim N(0, 1)$ and that $g(\cdot)$ depends linearly on the past squared realizations of $u$:

$$\sigma_t^2 = a_0 + \sum_{i=1}^{q} a_i u_{t-i}^2 + \sum_{i=1}^{p} b_i \sigma_{t-i}^2.$$

(1.3)

This is a Gaussian GARCH($p, q$) specification introduced by Bollerslev (1986); when $p = 0$ it becomes the ARCH($q$) specification of Engle (1982).

Section 2 discusses some of the shortcomings of such models as a description of the volatility in stock return data. We argue that a promising alternative is to allow for the possibility of sudden, discrete changes in the values of the parameters of an ARCH($q$) process as in the Markov-switching model in Hamilton (1989). Markov-switching ARCH models have previously been employed in Brunner’s (1991) study of inflation and Cai’s (forthcoming) analysis of Treasury bill yields. Section 3 of this paper motivates a parsimonious parameterization of a Markov-switching ARCH model which differs from that used by these researchers, and develops optimal forecasts for this specification. Section 4 presents an empirical application, which also differs from previous studies in allowing up to four different regimes. Another innovation is that stock returns within any given regime are modeled with a Student $t$ distribution rather than a Normal distribution. We argue that this class of models better describes a number of key features of the data.

2. Stock market volatility and traditional ARCH specifications

The stock price series used in this analysis is the value-weighted portfolio of stocks traded on the New York Stock Exchange contained in the CRISP data
tapes. The raw data are returns from Wednesday of one week to Tuesday of the following week. The original data start with the week ended Tuesday, July 3, 1962 and end with the week ended Tuesday, December 29, 1987. All results reported in this paper condition on the first four observations and begin estimation with the week ended Tuesday, July 31, 1962. Thus \( T = 1327 \) observations were used to estimate each model.

Let \( y_t \) denote the weekly stock return measured in percent; for example, \( y_t = -2.0 \) means that stock prices fell 2% in week \( t \). In all of the models we investigated, the process \( u_t \) that is described by the ARCH process is the residual from a first-order autoregression for stock returns:

\[
y_t = x + \phi y_{t-1} + u_t.
\]

By far the most popular ARCH model that has been used to describe financial market volatility is the GARCH(1, 1) specification.\(^1\) For our stock return data, we make two modifications to the usual GARCH (1, 1) specification which have received support in other studies and which clearly improve the model's ability to describe the weekly stock return series used in our study.

First, we follow a suggestion by Bollerslev (1987) and Baillie and DeGennaro (1990) and treat \( u_t \) in (1.1) as drawn from a Student \( t \) distribution with \( v \) degrees of freedom (normalized to have unit variance). The implied conditional density for \( u_t \) is then:

\[
f(u_t | u_{t-1}, u_{t-2}, \ldots) = \frac{\Gamma\left(\frac{v + 1}{2}\right)}{\sqrt{\pi} \Gamma(v/2)} (v - 2)^{-1/2} \sigma_t^{-1} \times \left[1 + \frac{u_t^2}{\sigma_t^2 (v - 2)}\right]^{-(v + 1)/2}.
\]

These authors' proposal was to treat the degrees of freedom \( v \) as a parameter to be estimated by maximum likelihood along with the others; \( v \to \infty \) corresponds to the \( N(0, \sigma_t^2) \) distribution.

Second, we follow Black (1976) and Nelson (1991), among many others, who have argued that a stock price decrease tends to increase subsequent volatility by more than would a stock price increase of the same magnitude. This is sometimes called the 'leverage' effect. Engle and Ng (1991) compared several alternative specifications of this leverage effect, and concluded that the parameterization of Glosten, Jagannathan, and Runkle (1989) was the most promising. Their GARCH-L (1, 1) specification is

\[
\sigma_t^2 = a_0 + a_1 \cdot u_{t-1}^2 + \xi \cdot d_{t-1} \cdot u_{t-1}^2 + b_1 \cdot \sigma_{t-1}^2,
\]

\(^1\) For example, Engle's (1991) survey observed that 'most investigators have found that the amazing GARCH(1, 1) is a generally excellent model for a wide range of financial data.' Bollerslev, Chou, and Kroner (1992) detailed several dozen successful applications of the GARCH(1, 1) specification.
where $d_{t-1}$ is a dummy variable that is equal to zero if $u_{t-1} > 0$ and equal to unity if $u_{t-1} \leq 0$. The leverage effect predicts that $\hat{c} > 0$.

Our maximum likelihood estimates for this specification based on weekly stock returns are as follows:\(^2\)

\[
y_t = 0.31 + 0.27 y_{t-1} + u_t,
\]
\[
\text{(0.05)} \quad \text{(0.03)}
\]

\[
\sigma_t^2 = 0.28 + 0.07 u_{t-1}^2 + 0.23 d_{t-1} \cdot u_{t-1}^2 + 0.78 \sigma_{t-1}^2,
\]
\[
\text{(0.12)} \quad \text{(0.03)} \quad \text{(0.06)} \quad \text{(0.06)}
\]

\[
\hat{\gamma} = 5.6.
\]
\[
\text{(0.8)}
\]

The numbers in parentheses are the usual 'asymptotic standard errors' based on the matrix of second derivatives of the log-likelihood function. Table 1 presents some summary statistics for this model and compares it with others we have estimated.

All of the parameters are highly statistically significant on the basis of either Wald tests or likelihood ratio tests, suggesting that this model offers a clear improvement over the null hypothesis of homoskedastic errors. However, earlier researchers have raised several concerns about the adequacy of such a specification, as we now discuss.

### 2.1. Forecasting performance

If the specification were correct and the parameters were known with certainty, then $\sigma_t^2$ would be the conditional expectation of $u_t^2$. Hence a mean squared error loss function,

\[
E \left\{ (u_t^2 - \theta_t)^2 | u_{t-1}, u_{t-2}, \ldots \right\},
\]

would be minimized with respect to $\theta_t$ by choosing $\theta_t = \sigma_t^2$ for $\sigma_t^2$ given by (2.5). The average value of $(\hat{u}_t^2 - \hat{\sigma}_t^2)^2$ turns out to be given by

\[
\text{MSE} = \frac{1}{T-1} \sum_{t=1}^{T} (\hat{u}_t^2 - \hat{\sigma}_t^2)^2 = 815.6.
\]

\(^2\)Generating the sequence of variances $\{\sigma_t^2\}$ in (2.5) requires knowledge of the starting value $\sigma_0^2$. For the reported results, $\sigma_0^2$ was treated as a separate parameter which was estimated by maximum likelihood along with the others. We also fit the models following Bollerslev's (1986) suggestion of setting $\sigma_0^2$ equal to the average value of $\hat{u}_t^2$, with very similar results.
<table>
<thead>
<tr>
<th>Model</th>
<th>No. of parameters</th>
<th>Log-likelihood</th>
<th>AIC</th>
<th>Schwarz</th>
<th>Degrees of freedom</th>
<th>Persistence (2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant variance</td>
<td>2</td>
<td>-3095.2</td>
<td>-3097.2</td>
<td>-3102.4</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Gaussian GARCH(1, 1)</td>
<td>5*</td>
<td>-2944.7</td>
<td>-2949.7</td>
<td>-2962.7</td>
<td>—</td>
<td>0.99</td>
</tr>
<tr>
<td>Student t GARCH-L(1, 1)</td>
<td>7*</td>
<td>-2822.0</td>
<td>-2829.0</td>
<td>-2847.2</td>
<td>5.6</td>
<td>0.96</td>
</tr>
<tr>
<td>GED GARCH-L(1, 1)</td>
<td>7*</td>
<td>-2839.8</td>
<td>-2846.8</td>
<td>-2865.0</td>
<td>1.2</td>
<td>0.96</td>
</tr>
<tr>
<td>Gaussian ARCH-L(2)</td>
<td>6</td>
<td>-2934.4</td>
<td>-2940.4</td>
<td>-2956.0</td>
<td>—</td>
<td>0.74</td>
</tr>
<tr>
<td>Gaussian SWARCH-L(2, 2)</td>
<td>9</td>
<td>-2836.3</td>
<td>-2845.3</td>
<td>-2868.7</td>
<td>—</td>
<td>0.42</td>
</tr>
<tr>
<td>Student t ARCH-L(2)</td>
<td>7</td>
<td>-2846.0</td>
<td>-2853.0</td>
<td>-2871.2</td>
<td>4.7</td>
<td>0.71</td>
</tr>
<tr>
<td>Student t SWARCH-L(2, 2)</td>
<td>10</td>
<td>-2818.0</td>
<td>-2828.0</td>
<td>-2854.0</td>
<td>5.7</td>
<td>0.59</td>
</tr>
<tr>
<td>Student t SWARCH-L(3, 2)</td>
<td>13</td>
<td>-2802.7</td>
<td>-2815.7</td>
<td>-2849.4</td>
<td>7.2</td>
<td>0.48</td>
</tr>
<tr>
<td>Student t SWARCH-L(4, 2)</td>
<td>15</td>
<td>-2798.1</td>
<td>-2813.1</td>
<td>-2852.0</td>
<td>8.7</td>
<td>0.50</td>
</tr>
</tbody>
</table>

The count of the number of parameters attributed to the GARCH specifications does not include the estimate of the initial variance \( \sigma_z^2 \). The count of the number of parameters for the SWARCH-L(3, 2) and SWARCH-L(4, 2) specifications does not include the transition probabilities \( p_{ij} \) imputed to be zero.

The second column reports \( L^* \), the maximum value achieved for the log of the likelihood function. The number in parentheses below each entry reports what the p-value for a likelihood ratio test of that model against the preceding specification would be under the assumption that twice the difference in log-likelihoods is distributed \( \chi^2 \) with degrees of freedom equal to the difference in number of parameters between the null and alternative.

AIC was calculated as \( L^* - k \) for \( k \) the number of parameters in column 1.

Schwarz was calculated as \( L^* - (k/2) \cdot \ln(T) \) for \( T = 1327 \).

The degree of freedom parameter is the magnitude \( v \) in expression (2.2) for the Student t distribution or the parameter \( \zeta \) for the GED distribution [see Eq. (2.4) in Nelson, 1991]. The standard error for this parameter is in parentheses.

The persistence parameter \( \lambda \) is characterized by expression (2.11) for a GARCH(1, 1) and by the largest eigenvalue of the matrix in (3.16) for an ARCH-L(2, 2) or SWARCH-L(2, 2) model.
For a standard of comparison, let $\bar{y}$ denote the sample average return and $s^2$ the unconditional sample variance:

$$
\bar{y} = T^{-1} \sum_{t=1}^{T} y_t, \quad s^2 = T^{-1} \sum_{t=1}^{T} (y_t - \bar{y})^2.
$$

If we simply forecast the variance of $y_t$ to be the constant $s^2$ throughout the sample, the corresponding value for the loss function is

$$
MSE = T^{-1} \sum_{t=1}^{T} \{(y_t - \bar{y})^2 - s^2\}^2 = 757.9. \quad (2.8)
$$

Thus the Student $t$ GARCH-L(1, 1) model actually yields poorer forecasts than just using a constant variance. The $R^2$, or percent improvement of (2.7) over (2.8), is $-0.08$.

The average squared forecast error is probably an unfair standard for judging this specification, since it is based on fourth moments of the actual data $y_t$. The unconditional fourth moment would fail to exist if (2.5) were the data-generating process. The third row of Table 2 compares the forecast $\sigma_t^2$ with the unconditional sample variance $s^2$ using three alternative loss functions. The GARCH-L(1, 1) does almost as poorly in terms of mean absolute error, but somewhat better for other loss functions.

Tables 1 and 2 also report results for a number of other simple ARCH models. The Student $t$ does better than the generalized error distribution suggested by Nelson (1991) and is vastly superior to a conditionally Gaussian model, both in terms of forecasting performance and in terms of a classical test of the null hypothesis of Gaussian residuals against the alternative of Student $t$. The leverage effect [the parameter $\xi$ in Eq. (2.3)] is likewise highly statistically significant and helps improve the forecasts, and the GARCH(1, 1) yields much better forecasts than low-order ARCH models. Thus, despite its weak forecasting performance, the model in (2.4)–(2.6) is the best representative that we have found for describing these stock return data out of the class of models given in (1.1)–(1.3).

2.2. Persistence

Our second concern with models such as that in (2.4)–(2.6) is the high degree of persistence they imply for stock volatility. To calculate a measure of this persistence, replace 'r' with 't + m' in Eq. (2.3) and write the result as

$$
E(u_{t+m}^2 | u_{t+m-1}, u_{t+m-2}, \ldots) = a_0 + a_1 \cdot u_{t+m-1}^2 + \xi \cdot d_{t+m-1} \cdot u_{t+m-1}^2 + h_1 \cdot E(u_{t+m-1}^2 | u_{t+m-2}, u_{t+m-3}, \ldots). \quad (2.9)
$$

West, Edison, and Cho (1993) suggested that an interesting alternative basis for comparing forecasts is to calculate the utility of an investor with a particular utility function investing on the basis of different variance forecasts.
Table 2
Comparison of one-period-ahead forecasts of different models

<table>
<thead>
<tr>
<th>Model</th>
<th>Loss function (percent improvement)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MSE</td>
</tr>
<tr>
<td>Constant variance</td>
<td>757.9</td>
</tr>
<tr>
<td>Gaussian GARCH(1, 1)</td>
<td>866.8</td>
</tr>
<tr>
<td></td>
<td>(- 0.14)</td>
</tr>
<tr>
<td>Student t GARCH-L(1, 1)</td>
<td>815.6</td>
</tr>
<tr>
<td></td>
<td>(- 0.08)</td>
</tr>
<tr>
<td>GED GARCH-L(1, 1)</td>
<td>840.71</td>
</tr>
<tr>
<td></td>
<td>(- 0.11)</td>
</tr>
<tr>
<td>Gaussian ARCH-L(2)</td>
<td>1094.7</td>
</tr>
<tr>
<td></td>
<td>(- 0.44)</td>
</tr>
<tr>
<td>Gaussian SWARCH-L(2, 2)</td>
<td>822.1</td>
</tr>
<tr>
<td></td>
<td>(- 0.08)</td>
</tr>
<tr>
<td>Student t ARCH-L(2)</td>
<td>904.1</td>
</tr>
<tr>
<td></td>
<td>(- 0.19)</td>
</tr>
<tr>
<td>Student t SWARCH-L(2, 2)</td>
<td>850.4</td>
</tr>
<tr>
<td></td>
<td>(- 0.12)</td>
</tr>
<tr>
<td>Student t SWARCH-L(3, 2)</td>
<td>814.3</td>
</tr>
<tr>
<td></td>
<td>(- 0.07)</td>
</tr>
<tr>
<td>Student t SWARCH-L(4, 2)</td>
<td>709.3</td>
</tr>
<tr>
<td></td>
<td>(0.06)</td>
</tr>
</tbody>
</table>

The following loss functions are reported:

\[
\text{MSE} = T^{-1} \sum_{t=1}^{T} |\hat{u}_t^2 - \sigma_t^2|^2, \quad \text{MAE} = T^{-1} \sum_{t=1}^{T} |\hat{u}_t^2 - \sigma_t^2|, \\
\text{[LE]}^2 = T^{-1} \sum_{t=1}^{T} (\ln(\hat{u}_t^2) - \ln(\sigma_t^2))^2, \quad |\text{LE}| = T^{-1} \sum_{t=1}^{T} |\ln(\hat{u}_t^2) - \ln(\sigma_t^2)|. 
\]

For the constant variance model, \(\hat{u}_t\) represents \(y_t - \bar{y}\) and \(\sigma_t^2\) represents \(\sigma^2\). The percent improvement compares each model to the constant variance specification.

Take conditional expectations of both sides of (2.9), noting that the symmetry of the distribution of \(v_{t+m-1}\) implies that \(d_{t+m-1}\) is uncorrelated with \(u_{t+m-1}^2\):

\[
E(u_{t+m-1}^2 | u_t, u_{t-1}, \ldots) = a_0 + a_1 \cdot E(u_{t+m-1}^2 | u_t, u_{t-1}, \ldots) \\
+ \bar{z} \cdot E(d_{t+m-1} | u_t, u_{t-1}, \ldots) \cdot E(u_{t+m-1}^2 | u_t, u_{t-1}, \ldots) \\
+ \bar{b}_1 \cdot E(u_{t+m-1}^2 | u_t, u_{t-1}, \ldots). \quad (2.10)
\]
Table 3
Comparison of four-period-ahead and eight-period-ahead forecasts

| Model                  | Percent improvement in loss function | MSE    | MAE    | $[LE]^2$ | $|LE|$ |
|------------------------|-------------------------------------|--------|--------|----------|--------|
| **Four-week-ahead forecasts** |                                     |        |        |          |        |
| Gaussian GARCH(1, 1)   |                                     | 0.22   | 0.24   | -0.11    | -0.06  |
| Student t GARCH-L(1, 1)|                                     | -0.12  | 0.03   | 0.08     | 0.05   |
| Student t SWARCH-L(2, 2)|                                   | -0.00  | 0.05   | 0.02     | 0.02   |
| Student t SWARCH-L(3, 2)|                                   | 0.00   | 0.07   | 0.02     | 0.04   |
| Student t SWARCH-L(4, 2)|                                   | 0.01   | 0.07   | -0.01    | 0.03   |
| **Eight-week-ahead forecasts** |                                   |        |        |          |        |
| Gaussian GARCH(1, 1)   |                                     | -0.24  | -0.41  | -0.19    | -0.15  |
| Student t GARCH-L(1, 1)|                                     | -0.09  | -0.02  | 0.09     | 0.02   |
| Student t SWARCH-L(2, 2)|                                   | 0.00   | 0.05   | 0.05     | 0.01   |
| Student t SWARCH-L(3, 2)|                                   | 0.00   | 0.07   | 0.05     | 0.03   |
| Student t SWARCH-L(4, 2)|                                   | 0.00   | 0.06   | 0.03     | 0.03   |

Table entries report the percent improvement of forecasts of each model compared to the constant-variance forecast based on each of the four loss functions described in Table 2.

Let $\sigma_{i+m|t}^2$ denote the $m$-period-ahead forecast of the variance:

$$\sigma_{i+m|t}^2 \equiv \mathbb{E}(u_{i+m}^2 | u_t, u_{t-1}, \ldots).$$

Thus for example $\sigma_{i+1|t}^2 \equiv \sigma_{i+1}^2$. Using this notation, (2.10) can be written

$$\sigma_{i+m|t}^2 = a_0 + (a_1 + b_1 + \xi/2) \cdot \sigma_{i-m-1|t}^2.$$

Thus the $m$-period-ahead forecast follows a simple first-order difference equation in the forecast horizon $m$, with decay parameter $\lambda$ given by

$$\lambda = (a_1 + b_1 + \xi/2).$$

(2.11)

For the values in (2.5), $\lambda$ is estimated to be 0.96. Thus the model implies extremely persistent movements in stock price volatility. For example, any change in the stock market this week will continue to have nonnegligible consequences a full year later: $(0.96)^{52} = 0.12.$

Such persistence is difficult to reconcile with the poor forecasting performance - if the variance were this persistent, the model should do a much better job at forecasting it. However, the model’s forecasts continue to worsen relative to a constant-variance specification as the forecast horizon increases (see Table 3).

The consequence of the big value for $\lambda$ can be seen dramatically in the effects of the stock market crash in October 1987. The top panel of Fig. 1 plots weekly stock returns for the last six months of 1987, in which the October crash is an
Fig. 1. Top panel: Weekly returns on the New York Stock Exchange during the second half of 1987. Middle panel: Widths of 95% confidence intervals for one-week-ahead forecasts of stock returns implied by the Student t GARCH-L(1, 1) model. Bottom panel: $\mu_t + 2 \cdot \sigma_t$, where $\sigma_t$ is calculated from expression (3.15) for $m = 1$ using the parameters for the Student t SWARCH-L(3, 2) model.

extreme outlier. The middle panel plots the corresponding 95% confidence interval for future forecasts of $u_t$ based on the specification of the variance in (2.5). The model anticipates possible weekly stock price movements in excess of 10% through the end of 1987 as a consequence of the October 1987 crash. This perception was evidently not shared by stock market participants, whose beliefs

---

4 Recall that a standard $t$ variable with $v$ degrees of freedom has variance $v(v - 2)$. Thus if $u_t$ has a Student $t$ distribution with variance $\sigma_t^2$ and $v$ degrees of freedom, then $u_t \cdot \sqrt{v} = [\sigma_t^2 \cdot (v - 2)]^{1/2}$ has the standard $t$ distribution with $v$ degrees of freedom. The 95% critical values for a $t$ variable with 5.6 degrees of freedom are $\pm 2.5$. Hence the 95% confidence intervals for $u_t$ would be calculated as $\pm 2.5 \div \sqrt{5.6} \cdot [\hat{\sigma}_t^2 \cdot (3.6)] = \pm 2 \cdot \hat{\sigma}_t$.
about stock market volatility should be reflected in the prices of stock options. Engle and Mustafa (1992) concluded on the basis of stock option prices that the volatility consequences of the 1987 crash disappeared much more rapidly than is suggested by the middle panel of Fig. 1. Lamoureux and Lastrapes (1993) presented related evidence based on earlier data that standard GARCH models overforecast the persistence of volatility.

3. ARCH specifications with changes in regime

A number of researchers have suggested that the poor forecasting performance and spuriously high persistence of ARCH models might both be related to structural change in the ARCH process. The high estimate for the persistence parameter $\alpha$ is known to be nonrobust across subsamples. Diebold (1986) and Lamoureux and Lastrapes (1990) argued that the high estimated value for $\alpha$ may reflect structural changes that occurred during the sample in the variance process. This is related to Perron's (1989) observation that changes in regime may give the spurious impression of unit roots in characterizations of the level of a series. Cai (forthcoming) in particular noted that the volatility in Treasury bill yields appears to be much less persistent when one models changes in the parameter through a Markov-switching process, and it seems promising to investigate whether a similar result might characterize stock returns.

For these reasons we explore a specification in which the parameters of an ARCH process can occasionally change. Let $y_t$ be a vector of observed variables and let $x_t$ denote an unobserved random variable that can take on the values 1, 2, ..., or $K$. Suppose that $x_t$ can be described by a Markov chain,

$\text{Prob}(x_t = j | x_{t-1} = i, x_{t-2} = k, \ldots, y_{t-1}, y_{t-2}, \ldots) = \text{Prob}(x_t = j | x_{t-1} = i) = p_{ij},$

for $i, j = 1, 2, \ldots, K$. It is sometimes convenient to collect the transition probabilities in a $(K \times K)$ matrix:

$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1K} \\ p_{21} & p_{22} & \cdots & p_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ p_{K1} & p_{K2} & \cdots & p_{KK} \end{bmatrix}.$

Note that each column of $P$ sums to unity.

The variable $x_t$ is regarded as the 'state' or 'regime' that the process is in at date $t$. By this we mean that $x_t$ governs that parameters of the conditional distribution of $y_t$. If the density of $y_t$ conditional on its own lagged values as well as on the
current and previous $q$ values for the state is of a known form,

$$f(y_t | s_t, s_{t-1}, \ldots, s_{t-q}, y_{t-1}, y_{t-2}, \ldots, y_0), \quad (3.2)$$

then the methods developed in Hamilton (1989) can be used to evaluate the likelihood function for the observed data and make inferences about the unobserved regimes. For example, $y_t$ could follow an ARCH($q$) process whose parameters depend on the unobserved realization of $s_t, s_{t-1}, \ldots, s_{t-q}$. Such models have been fit to inflation data by Brunner (1991) and to Treasury bill yields by Cai (forthcoming).

Although (3.2) provides a fairly general framework for describing structural change, it has the limitation that the density of $y_t$ can only depend on a finite number of lags of $s$ represented by the parameter $q$. Thus, for example, one can allow the parameters of an ARCH($q$) process to change, but changes in a GARCH($p, q$) process with $p > 0$ are not allowed as a special case of (3.2).

The objective is to select a parsimonious representation for the different possible regimes. A specification in which all of the parameters change with each regime would likely be numerically unwieldy and overparameterized. Hamilton (1989) suggested the following regime-switching model for the conditional mean:

$$y_t = \mu_{s_t} + \tilde{y}_t. \quad (3.3)$$

Here $\mu_{s_t}$ denotes the parameter $\mu_1$ when the process is in the regime represented by $s_t = 1$, while $\mu_{s_t}$ indicates $\mu_2$ when $s_t = 2$, and so on. The variable $\tilde{y}_t$ was assumed to follow a zero-mean $q$th-order autoregression:

$$\tilde{y}_t = \phi_1 \tilde{y}_{t-1} + \phi_2 \tilde{y}_{t-2} + \cdots + \phi_q \tilde{y}_{t-q} + \epsilon_t. \quad (3.4)$$

The idea behind this specification was that occasional, abrupt shifts in the average level of $y_t$ would be captured by the values of $\mu_{s_t}$.

A natural extension of this approach to the conditional variance would be to model the residual $u_t$ in (2.1) as

$$u_t = \sqrt{g_{s_t}} \times \tilde{u}_t. \quad (3.3)$$

Here $\tilde{u}_t$ is assumed to follow a standard ARCH-L($q$) process,

$$\tilde{u}_t = h_t \cdot v_t,$$

with $v_t$ a zero mean, unit variance i.i.d. sequence, while $h_t$ obeys

$$h_t^2 = a_0 + a_1 \tilde{u}_{t-1}^2 + a_2 \tilde{u}_{t-2}^2 + \cdots + a_q \tilde{u}_{t-q}^2 + \xi \cdot d_{t-1} \cdot \tilde{u}_{t-1}^2, \quad (3.4)$$

where $d_{t-1} = 1$ if $\tilde{u}_{t-1} \leq 0$ and $d_{t-1} = 0$ for $\tilde{u}_{t-1} > 0$. The underlying ARCH-L($q$) variable $\tilde{u}_t$ is then multiplied by the constant $\sqrt{g_{s_t}}$ when the process is in the regime represented by $s_t = 1$, multiplied by $\sqrt{g_2}$ when $s_t = 2$, and so on. The
factor for the first state $y_1$ is normalized at unity with $g_j \geq 1$ for $j = 2, 3, \ldots, K$. The idea is thus to model changes in regime as changes in the scale of the process. Conditional on knowing the current and past regimes, the variance implied for the residual $u_t$ is

$$
E(u_t^2 | s_t, s_{t-1}, \ldots, s_{t-q}, u_{t-1}, u_{t-2}, \ldots, u_{t-q})
$$

$$
= g_{s_t} \{ a_0 + a_1 \cdot (u_{t-1}^2 / g_{s_{t-1}}) + a_2 \cdot (u_{t-2}^2 / g_{s_{t-2}}) + \ldots + a_q \cdot (u_{t-q}^2 / g_{s_{t-q}}) 
+ \xi \cdot d_{t-1} \cdot (u_{t-1}^2 / g_{s_{t-1}}) \}
$$

$$
\equiv \sigma_t^2(s_t, s_{t-1}, \ldots, s_{t-q}).
$$

(3.5)

where $d_{t-1} = 1$ for $u_{t-1} \leq 0$ and $d_{t-1} = 0$ for $u_{t-1} > 0$.

In the absence of a leverage effect ($\xi = 0$), we will say that $u_t$ in (3.3) follows a $K$-state, $q$th-order Markov-switching ARCH process, denoted $u_t \sim$ SWARCH($K, q$). In the presence of leverage effects ($\xi \neq 0$), we will call it a SWARCH-L($K, q$) specification. We investigated both Gaussian ($v_t \sim N(0, 1)$) and Student $t$ ($v_t$ distributed $t$ with $\nu$ degrees of freedom and unit variance) versions of the model.

The appendix describes the algorithm used to evaluate the sample log-likelihood function,

$$
\ell_p = \sum_{t=1}^{T} \ln f(y_t | y_{t-1}, y_{t-2}, \ldots, y_{t-3}),
$$

(3.6)

which can be maximized numerically with respect to the population parameters $\phi$, $\theta_0$, $\theta_1, \ldots, g_k$, $g_1, g_2, \ldots, g_k$, $\xi$, and $\nu$ subject to the constraints that $g_1 = 1$, $\sum_{j=1}^{K} \rho_{ij} = 1$ for $i = 1, 2, \ldots, K$, and $0 \leq \rho_{ij} \leq 1$ for $i,j = 1, 2, \ldots, K$. The appendix also describes the inference about the particular state the process was in at date $t$. When this inference is based on information observed through date $t$, it is called the 'filter probability':

$$
p(s_t, s_{t-1}, \ldots, s_{t-q} | y_t, y_{t-1}, \ldots, y_{t-3}).
$$

(3.7)

Expression (3.7) denotes the conditional probability that the date $t$ state was the value $s_t$, the date $t - 1$ state was the value $s_{t-1}$, and the date $t - q$ state was the value $s_{t-q}$. These probabilities condition on the values of $y$ observed through date $t$. Since there are $K^{q+1}$ possible configurations for $(s_t, s_{t-1}, \ldots, s_{t-q})$, there

\[\frac{\partial}{\partial \theta_{ij}} \ell_p = \sum_{t=1}^{T} y_t^{-1} (y_t - \hat{y}_t)^2 \] for $j = 1, 2, \ldots, K - 1$,

\[\frac{\partial}{\partial \theta_{ij}} \ell_p = \sum_{t=1}^{T} y_t^{-1} (y_t - \hat{y}_t)^2 \] for $j = K$.

and estimating $\theta_{ij}$ for $i = 1, 2, \ldots, K$ and $j = 1, 2, \ldots, K - 1$ without restrictions.
are \( K^{q+1} \) separate numbers of the form of (3.7); these \( K^{q+1} \) values sum to unity by construction.

Alternatively, the full sample of observations can be used to construct the ‘smoothed probability’:

\[
p(s_t | Y_T, Y_{T-1}, \ldots, Y_{-1}) . \tag{3.8}
\]

Expression (3.8) denotes \( K \) separate numbers for each date \( t \) in the sample; again these \( K \) numbers sum to unity.

3.1. Forecasts

To calculate \( m \)-period-ahead forecasts of \( u_{t+m}^2 \), consider first a hypothetical situation in which we knew the values of \( s_t, s_{t-1}, \ldots, s_{t-q+1} \) with certainty, meaning we would also know with certainty the values of \( \tilde{u}_t = u_t / \sqrt{g_{s_t}} \) for \( \tau = t, t - 1, \ldots, t - q + 1 \). For this information set the forecast of \( u_{t+m}^2 \) would be

\[
E(u_{t+m}^2 | s_t, s_{t-1}, \ldots, s_{t-q+1}, \tilde{u}_t, \tilde{u}_{t-1}, \ldots, \tilde{u}_{t-q+1})
\]

\[
= E(g_{s_{t+m}} \cdot \tilde{u}_{t+m}^2 | s_t, s_{t-1}, \ldots, s_{t-q+1}, \tilde{u}_t, \tilde{u}_{t-1}, \ldots, \tilde{u}_{t-q+1})
\]

\[
= E(g_{s_{t+m}} | s_t, s_{t-1}, \ldots, s_{t-q+1}) \cdot E(\tilde{u}_{t+m}^2 | \tilde{u}_t, \tilde{u}_{t-1}, \ldots, \tilde{u}_{t-q+1}) . \tag{3.9}
\]

where the last equality follows from the fact that \( s_t \) is independent of \( r_t \) and \( \tilde{u}_t \) for all \( t \) and \( \tau \). Since \( s_t \) follows a Markov chain, the first term in (3.9) is given by

\[
E(g_{s_{t+m}} | s_t, s_{t-1}, \ldots, s_{t-q+1}) = \sum_{j=1}^{K} g_j \cdot \text{Prob}(s_{t+m} = j | s_t) . \tag{3.10}
\]

The \( m \)-period-ahead transition probabilities can be calculated by multiplying the matrix in (3.1) by itself \( m \) times:

\[
\begin{bmatrix}
\text{Prob}(s_{t+m} = 1 | s_t = 1) & \text{Prob}(s_{t+m} = 1 | s_t = 2) & \cdots & \text{Prob}(s_{t+m} = 1 | s_t = K) \\
\text{Prob}(s_{t+m} = 2 | s_t = 1) & \text{Prob}(s_{t+m} = 2 | s_t = 2) & \cdots & \text{Prob}(s_{t+m} = 2 | s_t = K) \\
\vdots & \vdots & \ddots & \vdots \\
\text{Prob}(s_{t+m} = K | s_t = 1) & \text{Prob}(s_{t+m} = K | s_t = 2) & \cdots & \text{Prob}(s_{t+m} = K | s_t = K)
\end{bmatrix} = P^m .
\]

Thus, if the switching factors are collected in a \((1 \times K)\) vector \( g' \),

\[
g' = [g_1 \quad g_2 \quad \cdots \quad g_K] .
\]

then

\[
F(g_{s_{t+m}} | s_t = i) = g' P^m e_i . \tag{3.11}
\]

where \( e_i \) denotes the \( i \)th column of the \((K \times K)\) identity matrix.
The second term in (3.9) is equally simple to construct from the fact that 
\( \hat{u}_t \) follows a standard ARCH-L(q) process:

\[
E(\hat{u}_{t+m}^2 | \hat{u}_t, \hat{u}_{t-1}, \ldots, \hat{u}_{t-q+1}) = \alpha_0 + \alpha_1 \cdot \hat{u}_t^2 + \alpha_2 \cdot \hat{u}_{t-1}^2 + \cdots + \alpha_q \cdot \hat{u}_{t-q+1}^2 + \xi \cdot d_t \cdot \hat{u}_t^2 \quad \text{for } m = 1,
\]

\[
= \alpha_0 + (\alpha_1 + \xi/2) \cdot \hat{h}_{t+m-1|t}^2 + \alpha_2 \cdot \hat{h}_{t+m-2|t}^2 + \cdots + \alpha_q \cdot \hat{h}_{t+m-q|t}^2 \quad \text{for } m = 2, 3, \ldots,
\]

where

\[
\hat{h}_{t|t}^2 = \hat{u}_t^2 \quad \text{for } \tau \leq t,
\]

\[
E(\hat{u}_{t+m}^2 | \hat{u}_t, \hat{u}_{t-1}, \ldots) \quad \text{for } \tau > t.
\]

The sequence \( (\hat{h}_{t|t}^2) \) for \( \tau = t + 2, t + 3, \ldots \) is then calculated by iterating of (3.12). Recalling that \( \hat{u}_t = u_t/\sqrt{h_t} \), the forecast in (3.12) is a function of \( u_t, u_{t-1}, \ldots, u_{t-q+1} \) and the particular set of values for \( s_t, s_{t-1}, \ldots, s_{t-q+1} \) that were assumed:

\[
E(\hat{u}_{t+m}^2 | \hat{u}_t, \hat{u}_{t-1}, \ldots, \hat{u}_{t-q+1}) = \hat{h}_{t+m|t}^2 (s_t, s_{t-1}, \ldots, s_{t-q+1}, u_t, u_{t-1}, \ldots, u_{t-q+1}).
\]

The forecast in (3.9) can thus be written

\[
E(\hat{u}_{t+m}^2 | s_t, s_{t-1}, \ldots, s_{t-q+1}, u_t, u_{t-1}, \ldots, u_{t-q+1}) = (g^T P^m e_s) \cdot \hat{h}_{t+m|t}^2 (s_t, s_{t-1}, \ldots, s_{t-q+1}, u_t, u_{t-1}, \ldots, u_{t-q+1}). \quad (3.13)
\]

For given values of \( u_t, u_{t-1}, \ldots, u_{t-q+1} \), expression (3.13) describes a different forecast of \( u_{t+m}^2 \) for each possible configuration of \( s_t, s_{t-1}, \ldots, s_{t-q+1} \), which we might denote

\[
E(\hat{u}_{t+m}^2 | s_t, s_{t-1}, \ldots, s_{t-q+1}, u_t, u_{t-1}, \ldots, u_{t-q+1}) \equiv \kappa (s_t, s_{t-1}, \ldots, s_{t-q+1}, u_t, u_{t-1}, \ldots, u_{t-q+1}). \quad (3.14)
\]

In practice we do not know the value of \( s_t, s_{t-1}, \ldots, s_{t-q+1} \). However, by the law of iterated expectations,

\[
\sigma_{t+m|t}^2 = \mathbb{E}(u_{t+m}^2 | u_t, u_{t-1}, \ldots, u_{t-q+1})
\]

\[
= \sum_{s_t} \sum_{s_{t-1}} \cdots \sum_{s_{t-q+1}} \mathbb{E}(u_{t+m}^2 | s_t, s_{t-1}, \ldots, s_{t-q+1}, u_t, u_{t-1}, \ldots, u_{t-q+1})
\]

\[
\times p(s_t, s_{t-1}, \ldots, s_{t-q+1} | y_t, y_{t-1}, \ldots, y_{t-3}). \quad (3.15)
\]

That is, we simply weight each of the conditional forecasts in (3.13) by the filter probability of that particular configuration to calculate an \( m \)-period-ahead forecast of \( u_{t+m}^2 \) based on the actual data observed.
Our model thus provides a framework that could generate the kind of nonlinearity in stock return volatility documented by Friedman and Laibson (1989) and Friedman (1992). These authors argued that conventional ARCH models fail to forecast well because large and small shocks have different effects. In the context of our model, suppose that the analyst is confident that the market has been in state 1 for the past $q$ periods and that $p_{11}$ is close to unity. If the date $t$ residual is small, the analyst would continue to place a high probability on the event that $s_t = 1$, and the forecast would basically be

$$
E(u_{t+1}^2 | u_t, u_{t-1}, \ldots, u_{t-q+1})
$$

$$
= a_0 + a_1 \cdot (u_t^2 / g_1) + a_2 (u_{t-1}^2 / g_1) + \cdots + a_q (u_{t-q+1}^2 / g_1) + \xi \cdot d_t \cdot (u_t^2 / g_1).
$$

The marginal effect of $u_t^2$ on the forecast would then be given by $$(a_1 + \xi d_t) / g_1.$$ On the other hand, if the residual is sufficiently large that the analyst is persuaded that the regime has shifted to $s_t = 2$, the marginal effect of $u_t^2$ on the forecast would be given by $(a_1 + \xi d_t) / g_2$. The specification thus allows for nonlinearities of the type documented by Friedman and Laibson. In our model, the nonlinearity arises as a result of the analyst's inference about the current volatility regime.

### 3.2. Persistence

Finally, we comment on the characterization of the persistence of the ARCH component of a SWARCH process. From (3.12) the forecasts of $\hat{u}_{t+m}$ obey the following $q$th-order difference equation for $m \geq 2$:

$$
\hat{h}_{t+m}^2 = a_0 + (a_1 + \xi/2) \cdot \hat{h}_{t+m-1}^2 + a_2 \cdot \hat{h}_{t+m-2}^2 + \cdots + a_q \cdot \hat{h}_{t+m-q}^2.
$$

It is well known that the solution to this difference equation takes the form

$$
\hat{h}_{t+m}^2 = c_0 + c_1 \lambda_1^m + c_2 \lambda_2^m + \cdots + c_q \lambda_q^m,
$$

where $\lambda_1, \lambda_2, \ldots, \lambda_q$ are the eigenvalues of the matrix

$$
\begin{bmatrix}
(a_1 + \xi/2) & a_2 & \cdots & a_{q-1} & a_q \\
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots \\
0 & 0 & 1 & 0
\end{bmatrix}.
$$

We report the largest eigenvalue of this matrix as a measure of the persistence of the ARCH component of a SWARCH-L($K$, $q$) process.
4. Empirical results

We fit a variety of different SWARCH specifications to the weekly stock return data described in Section 2. We estimated models with $q = 1$ to 3 ARCH terms and $K = 2$ to 4 states, with Normal and Student $t$ innovations, and with and without the leverage parameter $\xi$. For each model, the negative log-likelihood was minimized numerically using the optimization program OPTIMUM from the GAUSS programming language, usually beginning with steepest ascent and then switching to the BFGS algorithm. For models with $K = 2$ we randomly generated over 200 different starting values; typically we found a single local maximum for the likelihood function. For the $K = 3$ specification we used 25 different starting values. The $K = 4$ specification proved extremely difficult to maximize, owing in part to a nearly singular Hessian, as described below.

In every specification we looked at, the second ARCH parameter $a_2$ and the leverage parameter $\xi$ were strongly statistically significant on the basis of both Wald tests and likelihood ratio tests. Further, both tests overwhelmingly rejected the Normal in favor of the Student $t$ formulation. We also investigated a GED specification, which only slightly outperformed the Normal. The third ARCH parameter $a_3$ was not statistically significant or only marginally significant in the specifications with $K = 2$, so we did not attempt to estimate this parameter for the SWARCH models with $K > 2$. For these reasons, Tables 1 through 3 primarily report the results from SWARCH-L($K$, 2) specifications driven by Student $t$ innovations.

Since the GARCH(1, 1) specifications are not strictly nested within the SWARCH specifications, rows five and seven in Tables 1 and 2 also report results for ARCH-L(2) models with Normal and Student $t$ innovations. Although an ARCH-L(2) process could be described as a special case of SWARCH-L($K$, 2) with $K = 1$, the usual regularity conditions justifying the $\chi^2$ approximation to the likelihood ratio test do not hold in this setting, since the parameter $q_2$ is unidentified under the null hypothesis that there is really only one state. Table 1 nevertheless reports critical values for the likelihood ratio tests as if the $\chi^2$ approximation were valid. At a minimum we regard these as a useful descriptive summary of the fit of alternative models. The $p$-values for tests of the null hypothesis of only one or two states are so tiny that we have little doubt that these hypotheses would be rejected by any more rigorous testing procedures; indeed these events are so remote that the probabilities are not reliably calculated by the numerical routines of the GAUSS programming language on which the table entries are based. On the other hand, the suggestion of Hansen (1991, 1992) has proposed asymptotically valid tests, though their application here would be quite difficult numerically.
in Table 1 that the three-state specification is rejected in favor of the four-state specification at the 0.01 level may be sensitive to the distributional assumptions.

In addition to conventional tests for statistical significance, Tables 2 and 3 compare models on the basis of forecasting performance. The SWARCH-L(4, 2) specification is the only model we investigated that has a better mean squared error in forecasting $u_t^2$ than the constant variance specification. The SWARCH-L(4, 2) model is also clearly the best in terms of minimizing the mean absolute error as well, and continues to give useful eight-week-ahead forecasts. This model also performs reasonably when judged by the difference between $\ln(u_t^2)$ and $\ln(\sigma_t^2)$, though it is not quite as good as the GARCH-L(1, 1) specifications.

Table 1 also reports the model selection statistics proposed by Akaike (1976) and Schwarz (1978), though the asymptotic justification for these statistics again assumes the same regularity conditions referred to earlier, which are not fulfilled for this application. Based on Akaike's criterion, the SWARCH-L(4, 2) is the best model among any we investigated, followed by the SWARCH-L(3, 2). Based on Schwarz's criterion, the GARCH-L(1, 1) is best, followed by the SWARCH-L(3, 2).

In estimating the SWARCH-L(3, 2) and SWARCH-L(4, 2) specifications, we initially imposed no constraints on any of the transition probabilities $p_{ij}$ other than the conditions that $0 \leq p_{ij} \leq 1$ and $\sum_{j=1}^{K} p_{ij} = 1$. Several of these unrestricted MLE's fell on the boundary $p_{ij} = 0$, which is another violation of the regularity conditions. To calculate standard errors we then imposed $p_{ij} = 0$ and treated this parameter as a known constant for purposes of calculating the second derivatives of the log-likelihood.

The estimated Student $t$ SWARCH-L(3, 2) specification is as follows, with standard errors in parentheses:

$$y_t = 0.35 + 0.25 y_{t-1} + u_t,$$

$$u_t = \sqrt{g_{s_t}},$$

$$\tilde{u}_t = h_t \cdot v_t,$$

$$v_t \sim \text{i.i.d. Student } t \text{ with unit variance and 7.2 d.f.},$$

$$h_t^2 = 0.57 + 0.03 \tilde{u}_{t-1}^2 + 0.12 \tilde{u}_{t-2}^2 + 0.42 d_{t-1} \tilde{u}_{t-1}^2,$$
\[ d_{t-1} = 1 \quad \text{if} \quad u_{t-1} \leq 0, \]
\[ = 0 \quad \text{if} \quad u_{t-1} > 0. \]
\[
\hat{\theta}_1 = 1, \quad \hat{\theta}_2 = 4.4, \quad \hat{\theta}_3 = 13.1, \\
(1.0) \quad (3.2)
\]
\[
\hat{P} = \begin{bmatrix}
0.9924 & 0 & 0.0026 \\
0.0076 & 0.9914 & 0.0144 \\
0 & 0.0086 & 0.9831 \\
\end{bmatrix}
\]

The row j, column i element of \( \hat{P} \) represents the probability of going from state \( i \) to state \( j \).

Note that the estimated autoregressive coefficient \( \hat{\phi} \) is clearly nonzero - the new distributional assumptions and description of heteroskedasticity do not alter the conclusion that weekly stock price returns exhibit positive serial correlation.

The variance in the medium-volatility state (\( s_t = 2 \)) is four times as great as that in the low-volatility state, while that in the high-volatility state (\( s_t = 3 \)) is thirteen times as large as in the low-volatility state.

The top panel of Fig. 2 plots the weekly stock return series \( y_t \), while the other three panels plot the smoothed probabilities \( \text{Prob}(s_t = i | y_{1T}, y_{T-1}, \ldots, y_{-3}) \).

The low-volatility state describes the long quiet period from January 1963 through the end of 1965. Most of the other observations come from the medium-volatility state, with high-volatility episodes characterizing the last half of 1962, May 1969 to January 1972, February 1973 to April 1976, the period following the October 1987 crash, and probably for a short period around November 1982 as well.

The U.S. economy experienced four economic recessions during this sample, whose beginning and ending dates are marked with vertical lines in the bottom panel of Fig. 2. The market is judged to have been in the high-volatility state throughout the 1969–70 and 1973–75 recessions and likely towards the end of the 1981–82 recession as well. Thus the episodes of high stock market volatility appear to be related to general business downturns.\(^7\) The brief recession in 1980 does not appear to have coincided with unusually high volatility, however.

The estimated transition probabilities describe each state as highly persistent. State 1 would be expected to last on average for \( (1 - \hat{\rho}_{11})^{-1} = 132 \) weeks, while

\(^7\) French and Sichel (1993) presented interesting related evidence that the volatility of economic activity is higher during recessions than during expansions.
states 2 and 3 typically last for 116 weeks and 59 weeks, respectively. The market was in the quiet state 1 for only a single episode in the sample, which episode was preceded by state 3 and followed by state 2. Hence the maximum likelihood estimate is that state 1 is never preceded by state 2 ($\hat{p}_{21} = 0$) and state 1 is never followed by state 3 ($\hat{p}_{13} = 0$).

Although the states are highly persistent, the underlying fundamental ARCH-L(2) process for $\tilde{u}_t$ is much less so, with decay parameter $\lambda$ estimated to be 0.48. Note that $\lambda^4 = 0.05$, meaning that the volatility effects captured by $\tilde{u}_t$ die out almost completely after a month. The bottom panel of Fig. 1 plots $\pm 2\hat{\sigma}_{t|t-1}$ for the second half of 1987. As with the GARCH specification appearing in the middle panel, the crash initially widens these bands by an order of magnitude. In
Fig. 3. Top panel: Weekly returns on the New York Stock Exchange from the week ended July 31, 1962 to the week ended December 29, 1987. Middle panel: $\pm 2 \cdot \hat{\sigma}_i$, where $\hat{\sigma}_i^2$ is calculated from (2.5), the variance process for the Student $t$ GARCH-L(1, 1) model. Bottom panel: $\pm 2 \cdot \hat{\sigma}_{i-1}^2$, where $\hat{\sigma}_{i-1}^2$ is calculated from expression (3.15) for $m = 1$ using the parameters for the Student $t$ SWARCH-L(3, 2) model.

In contrast to the subsequent slow decay implied by the GARCH specification, however, this dramatic effect dies out relatively quickly, though a more modest widening persists as long as the market remains in the high-volatility state 3.

Fig. 3 compares the $\pm 2 \cdot \sigma_i$ bands for the GARCH-L(1, 1) specification (middle panel) with those for the SWARCH-L(3, 2) model (bottom panel) for the entire sample of observations. Both models infer prolonged episodes of high variance in the early and middle 1970's. For the SWARCH model this reflects the long periods in which the market appeared to be in the high-volatility state 3, whereas for the GARCH model this results from the long moving average of many large squared residuals. Although the broad patterns are similar, the models differ significantly in the reaction to large outliers. The GARCH confidence intervals shrink very gradually in response to these events, while the
SWARCH intervals quickly return to the trend level associated with a particular regime. The improvement in forecasting of the SWARCH model over the GARCH appears to be due to the ability to track the year-long shifts in volatility without imputing a large degree of persistence to the effects of individual outliers.

To measure the sensitivity of the forecasts of our model to uncertainty about the population parameters, we conducted the following Monte Carlo experiment. Let $\hat{\theta} = (\hat{z}, \hat{\phi}, \hat{a}_0, \hat{a}_1, \hat{a}_2, \hat{\xi}, \hat{\theta}_{11}, \hat{\theta}_{22}, \hat{\theta}_{31}, \hat{\theta}_{32}, \hat{g}_2, \hat{\theta}_3, \hat{v})$ denote the maximum likelihood estimate of the population parameters for the SWARCH-L(3, 2) model described above with transition probabilities parametrized as in Footnote 5. Let $\Omega$ denote the asymptotic variance–covariance matrix of $\theta$ as estimated from second derivatives of the log-likelihood. We generated 500 values for the vector $\theta$ drawn from a $N(\bar{\theta}, \Omega)$ distribution. For each $\theta_i$ and each date $t$ in the sample we calculated what the historical one-period-ahead forecast of $u_{t+1}$ would have been for that value of $\theta_i$, based on the actual historical values for $y_t, y_{t-1}, \ldots, y_{t-3}$, with this forecast denoted $\sigma^2_{t+1}(\theta_i)$. The sample variance of $\sigma^2_{t+1}(\theta_i)$ across Monte Carlo draws,

$$
(1/500) \sum_{i=1}^{500} \left\{ \sigma^2_{t+1}(\theta_i) - (1/500) \sum_{j=1}^{500} \sigma^2_{t+1}(\theta_j) \right\}^2,
$$

then gives an indication of how sensitive the forecast variance for date $t + 1$ is to uncertainty about the true value of the parameter vector $\theta$. The mean value for the square root of (4.1) across the $t = 1, 2, \ldots, 1327$ observations was 1.1. This standard error compares with a mean forecast given by

$$
(1/T) \sum_{t=1}^{T} \left\{ (1/500) \sum_{j=1}^{500} \sigma^2_{t+1}(\theta_j) \right\} = 6.2.
$$

Hence the standard error arising from parameter uncertainty typically is modest relative to the size of the forecast itself.

Our maximum likelihood estimates for a Student t SWARCH-L(4, 2) specification were as follows:

$$
y_t = 0.35 + 0.25 y_{t-1} + u_t,
$$

$$
u_t = \sqrt{h_t} \cdot u_t,
$$

$$
u_t \sim \text{i.i.d. Student } t \text{ with unit variance and 8.7 d.f.,}
$$

$$
h_t^2 = 0.55 + 0.02 \hat{u}_{t-1}^2 + 0.13 \hat{u}_{t-2}^2 + 0.41 d_{t-1} \hat{u}_{t-1}^2,
$$
\[ d_{t-1} = 1 \quad \text{if} \quad u_{t-1} \leq 0, \]
\[ g_1 = 0 \quad \text{if} \quad u_{t-1} > 0, \]
\[ g_1 = 1, \quad \hat{g}_2 = 4.5, \quad \hat{g}_3 = 13.8, \quad \hat{g}_4 = 169, \]
\[ (0.9) \quad (3.3) \quad (168) \]

\[
\hat{P} = \begin{bmatrix}
0.9931 & 0 & 0 & 0.2758 \\
0.0069 & 0.9925 & 0.0153 & 0 \\
0 & 0.0034 & 0.9847 & 0.7242 \\
0 & 0.0042 & 0 & 0
\end{bmatrix}.
\]

It is interesting that the first three of these states are essentially the same as states 1 through 3 of the SWARCH-L(3, 2) specification, while the fourth state corresponds to an order of magnitude increase in the variance even beyond that predicted in the 'high-volatility' state 3. Fig. 4 plots the smoothed probabilities implied by this model. Only two observations in the sample are clearly generated by state 4. One is the October 1987 crash. The other is more surprising, and corresponds to the 5.8% surge in stock prices in the first week of January 1963. This move followed two very quiet weeks for which the squared residuals were essentially zero, forcing \( h_f \) nearly to its lower limit of \( a_0 = 2.5 \). The probability of this occurring is the probability that a \( t(8.7) \) variable exceeds

\[
\frac{5.37}{\sqrt{(2.5 \cdot (8.7 - 2) \div (8.7) = 3.87,}}
\]

which probability is 0.002. Even if the process were imputed to have switched from state 2 to state 3 for this date, the probability of generating so large a gain would still only be 0.03. Since some sort of shift must have occurred at this date, one is inclined to regard this observation as having come from the rare conditions associated with state 4.

It is also interesting to note that \( p_{34} \) is estimated to be zero – the extreme state 4 appears never to have developed from a high-volatility state 3 but rather always follows the moderate-volatility state 2. This is consistent with Bates's (1991) failure to find any indication in option prices that the market perceived an increase in risk in the two months prior to the October 1987 crash.

Although these results are intriguing, one should not read too much into them. It might appear that three parameters – \( p_{24}, p_{43}, \) and \( g_4 \) – are identified solely on the basis of two observations. It is not quite this bad, since there is a modest probability that some of the other observations also may have come from regime 4. It turns out that \( \sum_{t=1}^{T} \text{Prob}(s_t = 4 | y_{T}, y_{T-1}, \ldots, y_{-3}) = 3.6 \). These other observations also give some information about these parameters, and of course all of the observations from regime 2 contain information about \( P_{24} \) – one can say with considerable confidence that this probability must be quite low. Even so, the Hessian for this SWARCH-L(4, 2) specification is very
Fig. 4. Top panel: Weekly returns on the New York Stock Exchange from the week ended July 31, 1962 to the week ended December 29, 1987. Second panel: Smoothed probability that market was in regime 1 for each indicated week \([\text{Prob}(s = 1) | y_t, y_{t-1}, \ldots, y_{-3})]\), as calculated from the Student t SWARCH-L(4, 2) specification. Third panel: Smoothed probability for regime 2. Fourth panel: Smoothed probability for regime 3. Fifth panel: Smoothed probability for regime 4.

nearly singular, and ‘asymptotic’ standard errors for these parameters are not meaningful.

Although some of the individual parameters of the SWARCH-L(4, 2) model are not measured with much confidence, we nevertheless find the results of interest. It is worth emphasizing that nothing about the model specification forced the procedure to regard the October 1987 crash as an observation from a single extreme regime. Indeed, the likelihood function suggests special treatment of October 1987 only when a fourth state is allowed, the first three states being reserved for broader patterns common to hundreds of observations. Specifying a general probability law that allows a rich class of different possibilities and letting the data speak for themselves in this way seems preferable to imposing dummy variables or break points in an arbitrary fashion and
regarding the break dates as the outcome of a deterministic rather than a stochastic process. The improvement in the value of the likelihood function achieved by broadening the class of dynamic models in this dimension seems a valid and useful framework for deciding whether the quiet stock market of the early 1960's or the turbulence of October 1987 ought to be regarded as special episodes.

5. Conclusion

This paper introduced a class of Markov-switching ARCH models which were used to describe volatility of stock prices. Our SWARCH specification offers a better statistical fit to the data and better forecasts. Our estimates attribute most of the persistence in stock price volatility to the persistence of low-, moderate-, and high-volatility regimes, which typically last for several years. The high-volatility regime is to some degree associated with economic recessions. Our analysis also confirms the findings of earlier researchers that stock price decreases lead to a bigger increase in volatility than would a stock price increase of the same magnitude, that the fundamental innovations are much better described as coming from a Student $t$ distribution with low degrees of freedom than by a Normal distribution, and that weekly stock returns are positively serially correlated.

Appendix: Procedure for evaluating the likelihood function

Step $t$ of the iteration to calculate the likelihood function has as input

$$p(s_t, s_{t-1}, \ldots, s_{t-q} | y_t, y_{t-1}, \ldots, y_{-q}).$$  \hfill (A.1)

Each of the $K^{q+1}$ numbers represented by (A.1) is multiplied by $p(s_t, s_{t-1}, \ldots, s_{t-q} | x_t, y_{t-1}, \ldots, y_{-q})$ to yield the $K^{q+2}$ separate numbers

$$p(s_{t+1}, s_t, s_{t-1}, \ldots, s_{t-q} | y_{t+1}, y_t, y_{t-1}, \ldots, y_{-q}).$$  \hfill (A.2)

For the Gaussian specification the preceding calculation uses

$$f(y_{t+1} | s_{t+1}, s_t, \ldots, s_{t-q+1}, y_t, y_{t-1}, \ldots, y_{t-q+1})$$

$$= \frac{1}{\sqrt{2\pi \sigma_{t+1}(s_{t+1}, s_t, \ldots, s_{t-q+1})}} \cdot \exp \left\{ \frac{-(y_{t+1} - x - \phi_i)^2}{2\sigma_{t+1}^2(s_{t+1}, s_t, \ldots, s_{t-q+1})} \right\},$$
where \( \sigma_t^2(s_t, s_{t-1}, \ldots, s_{t-q}) \) is given by (3.5) with \( u_t \equiv y_t - \alpha - \phi y_{t-1} \). For the Student \( t \) version of the model we instead use

\[
\begin{align*}
&f(y_{t+1} \mid s_{t+1}, s_t, \ldots, s_{t-q+1}, y_t, y_{t-1}, \ldots, y_{t-q+1}) \\
&= \frac{\Gamma \left( \frac{v+1}{2} \right)}{\Gamma(v/2) \cdot \sqrt{\pi \cdot \sqrt{v - 2} \cdot \sigma_{t+1}^2(s_{t+1}, s_t, \ldots, s_{t-q+1})}} \\
&\times \left\{ 1 + \frac{(y_{t+1} - \alpha - \phi y_t)^2}{(v-2) \cdot \sigma_{t+1}^2(s_{t+1}, s_t, \ldots, s_{t-q+1})} \right\}^{-(v+1)/2}
\end{align*}
\]

The numbers in (A.2) sum to the conditional density of \( y_{t+1} \),

\[
f(y_{t+1} \mid y_t, y_{t-1}, \ldots, y_{-3})
\]

from which the sample log-likelihood (3.6) can be calculated. If for any given \( s_{t+1}, s_t, \ldots, s_{t-q+1} \) the numbers in (A.2) are summed over the \( K \) possible values for \( s_t \) and the result is then divided by (A.3), one obtains

\[
p(s_{t+1}, s_t, \ldots, s_{t-q+1} \mid y_{t+1}, y_t, \ldots, y_{-3}),
\]

which is the input for step \( t + 1 \) of the iteration.

The iteration was started with \( p(s_0, s_{-1}, \ldots, s_{-q} \mid y_0, y_{-1}, \ldots, y_{-3}) \) set equal to the ergodic probabilities implied by the Markov chain as described in Eq. (22.2.26) in Hamilton (1994). Kim’s (1994) algorithm for calculating the smoothed probabilities \( p(s_t \mid y_T, y_{T-1}, \ldots, y_{-3}) \) is described in Eq. (22.4.14) in Hamilton (1994).

References


Engle, Robert F. and Victor K. Ng, 1991, Measuring and testing the impact of news on volatility, Mimeo. (University of California, San Diego, CA).


Friedman, Benjamin M., 1992, Big shocks and little shocks: Security returns with nonlinear persistence of volatility, Mimeo. (Harvard University, Cambridge, MA).


Glosten, Lawrence R., Ravi Jagannathan, and David Runkle, 1989, Relationship between the expected value and the volatility of the nominal excess return on stocks, Mimeo. (Northwestern University, Evanston, IL).


Hansen, Bruce E., 1991, Inference when a nuisance parameter is not identified under the null hypothesis, Mimeo. (University of Rochester, Rochester, NY).


