1 Conditional Heteroskedasticity Models with Pearson Family Disturbances

*Michael A. Magdalinos and George P. Mitsopoulos**

1 Introduction

The Autoregressive Conditional Heteroskedasticity (ARCH) model was introduced by Engle (1982). In this model the conditional variance of the errors is assumed to be a function of the squared past errors. Engle derives the Maximum Likelihood (ML) estimator for the ARCH model under the assumption that the conditional density of the error term is normal. Bollerslev (1986), suggested the Generalized Autoregressive Conditional Heteroskedasticity model (GARCH) in which the conditional variance of the errors is assumed to be a function of its lagged values and the squared past errors. Bollerslev derives the Maximum Likelihood (ML) estimator for the GARCH model under the assumption that the conditional density of the error term is normal. The ARCH and GARCH models are useful in modelling economic phenomena, mainly in the theory of finance (see e.g., Bollerslev *et al.* 1992 and Engle, 2002). In the above models the conditional density of the error term is assumed to be normal but in the applications with actual data, distributions other than the normal have been observed with fatter tails or with skewness significantly different from zero. For this reason, in particular applications with real data, other distributions have been used. Bollerslev (1987) used the Student's t distribution to model the monthly returns composite index. Baillie and Bollerslev (1989) also used the Student's t distribution while Hsieh (1989) chose the mixture Normal–Lognormal to model daily foreign-exchange rates. Jorion (1988) employed a mixture distribution of Normal–Poisson to model the foreign exchange and stockmarkets. Hansen (1993) used the skewed Student's t with a shape parameter, which may vary over time, to model exchange rates. Peruga (1988)

* In memory of my coauthor Professor Michael Magdalinos. The author G. Mitsopoulos is grateful to G. Phillips, E. Tzavalis and S. Symeonides for helpful comments on the first draft of the paper.

found empirical evidence of high skewness and kurtosis in exchange rates, and so on.

As an alternative approach, instead of using a specific distribution for the error term structure, non-parametric and semi-parametric techniques have been used to approximate the true error density. A Gaussian kernel was used in Pagan and Hong (1991) to model the risk premium. Pagan and Schwert (1990) again use a Gaussian kernel and include in the variance specification a Fourier transformation to model the stockmarket volatility. A Gaussian kernel was used in Sentana and Wadhwani (1991) to model stockmarket returns. Engle and Gonzalez-Rivera (1991) use a semi-parametric technique developed by Tapia and Thompson (1978) to model the exchange rate between the British pound and the US dollar. Tzavalis and Wickens (1995) use Cram–Charlier polynomials in order to incorporate in the model additional information for skewness and kurtosis, and so on. For a review of the Autoregressive Conditional Heteroskedasticity models, theory and applications, see Bollerslev *et al.* (1992), Bollerslev *et al.* (1994), Engle (2002).

Moreover, Magdalinos and Mitsopoulos (2003) use the Pearson System of Distributions (PSD) (see Eldetron and Johnson, 1969; Kendall and Stuart, 1977 and Johnson and Kotz, 1970) in order to approximate the error density in the case of the linear regression model. Here, we extend the use of the PSD for the case of the ARCH and GARCH models. The PSD includes a wide range of distributional shapes (such as Normal, Beta, Gamma, etc.), and is parsimoniously parameterized in terms of its first four cummulants. This is very convenient, as in practical cases it is unlikely to obtain reliable sample information for the higher-order cummulants. The definition of PSD is given in terms of the derivative of the log-density function. This implies that the score vector corresponding to the ARCH or GARCH models can be derived without the explicit identification of the error distribution.

The rest of the chapter is organized as follows. In section 2 we assume that the true error density belongs to the PSD and derive the one-step scoring estimator for the unknown parameters of the GARCH model. In section 3 we present an experimental study of the properties of the proposed estimator, while remarks and conclusions are presented in the concluding section 4.

2 GARCH with Pearson Family Disturbances

Let y_t, $(t = 1, \ldots, T)$ be an observable random variable, $x_t' = (1, x_{1t}, \ldots, x_{nt})$, $(t = 1, \ldots, T)$, be a vector of the regressors and b be the $n \times 1$ vector of unknown parameters. Also, let D_t be the information set available at time t and $\sigma(x_t)$ be the information set containing the information for the contemporaneous regressors. Using these information sets, define $\Delta_t = \sigma\{\sigma(x_t) \cup D_{t-1}\}$ as the information set containing the information from the contemporaneous regressors and all the past information. Moreover, let h_t be the conditional variance of the error term, which is a stochastic process and $\varphi(\cdot)$ the conditional density of y_t on Δ_t.

The above model may be written as

$$y_t \mid \Delta_t \sim \varphi(x_t' b, h_t). \tag{1}$$

Given the model (1), we assume that the conditional variance h_t is a stationary process. Then the Wold Decomposition Theorem (see e.g., Priestley, 1981, p. 756) implies that h_t can be expressed as a summable $MA(\infty)$ process, that is, in the form

$$h_t = \alpha_0 + \sum_{i=1}^{\infty} \alpha_i \varepsilon_{t-i}^2, \ \sum_{i=1}^{\infty} \alpha_i^2 < \infty \tag{2}$$

where ε_t is a white-noise process. This can be seen as a direct extension of the fact that an element of a linear space can be expressed in terms of an orthogonal basis. The choice of the basis, however, is arbitrary. Here we choose a positive basis to emphasize the fact that the conditional variance h_t is non-negative.

If we impose in (2) the following restrictions

$$\alpha_{p+1} = \alpha_{p+2} = \alpha_{p+3} = \cdots = 0 \tag{3}$$

then the representation (2) provides the theoretical foundation for the ARCH(p) model. Assumption (3) means that the conditional variance h_t is independent of the information $p+i$, $(i = 1, 2, \ldots)$, periods in the past.

If assumption (3) cannot be made, or if p is relatively large, then a more parsimonious parameterization can be obtained as follows. The representation (2) can be written as

$$h_t = A(L)\varepsilon_t^2, \ A(L) = \sum_{i=0}^{\infty} \alpha_i L^i \tag{4}$$

where L is the lag operator. Dhrymes (1971) shows that the linear space of the lag operators of the form

$$A(L) = \sum_{i=0}^{\infty} \alpha_i L^i, \ \sum_{i=0}^{\infty} \alpha_i^2 < \infty \tag{5}$$

is isomorphic to the space of real polynomials

$$A(x) = \sum_{i=0}^{\infty} \alpha_i x^i, \; x \in \mathbb{R}, \; \sum_{i=0}^{\infty} \alpha_i^2 < \infty. \tag{6}$$

This means that the properties of the two spaces are the same, so that a theorem that holds in one space is also valid in the other. It is well-known (see e.g., Bultheel, 1987, p. 36) that the polynomial (6) is approximated equally well by a Pade approximation of the form

$$R(x) = \frac{A_p(x)}{1 - B_q(x)}, \; A_{\mathrm{p}}(x) = \sum_{i=1}^{p} \alpha_i x^i, \; B_{\mathrm{q}}(x) = \sum_{j=1}^{q} \beta_j x^j \tag{7}$$

for finite values of p and q.

Hence we can approximate (2) by

$$h_t = \delta + R(L)\varepsilon_t^2 = \delta + \frac{A_p(L)}{1 - B_q(L)}\varepsilon_t^2$$

or

$$\left[1 - B_q(L)\right] h_t = \delta \left[1 - B_q(L)\right] + A_p(L)\varepsilon_t^2 \tag{8}$$

that is, the GARCH(p, q) representation

$$h_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^{q} \beta_j h_{t-j} \tag{9}$$

where $\alpha_0 = \delta\left[1 - B_q(L)\right] = \delta(1 - \beta_1 - \beta_2 - \cdots - \beta_q)$.

Now consider the GARCH(p, q) model that is defined by (1) and (9). The error term of this model is

$$\varepsilon_t = y_t - x_t' b \tag{10}$$

where it is assumed that the density of ε_t is $g(\varepsilon_t)$.

It is more convenient to work with the standardized residuals

$$u_t = h_t^{-1/2} \varepsilon_t \tag{11}$$

and assume that the density of u_t is $f(u_t)$.

The conditional variance h_t that is defined in (9) may be written as

$$h_t = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \cdots + \alpha_p \varepsilon_{t-p}^2 + \beta_1 h_{t-1} + \cdots + \beta_q h_{t-q} = \xi_t' \omega \qquad (12)$$

where

$$\xi_t' = \left(1, \varepsilon_{t-1}^2, \ldots, \varepsilon_{t-p}^2, h_{t-1}, \ldots, h_{t-q}\right)$$

$$\omega' = \left(\alpha_0, \alpha_1, \ldots, \alpha_p, \beta_1, \ldots, \beta_q\right) = \left(\alpha', \beta'\right).$$

Moreover let the shape (nuisance) parameters of the density $f(u)$ of u_t be $\gamma' = (\gamma_1, \gamma_2)$ where γ_1 and γ_2 are the skewness and kurtosis coefficients respectively.

We assume that the unknown density $f(u_t)$, of the disturbances u_t, belongs to the PSD. Since u_t are standardized, the equation defining the PSD can be written as

$$\frac{d \log f(u)}{du} = \frac{u - c_1}{c_1 u + c_2(u^2 - 3) - 1} \equiv \eta(u) \qquad (13)$$

where

$$c_1 = -\gamma_1(\gamma_2 + 6)/A, \quad c_2 = -\left(2\gamma_2 - 3\gamma_1^2\right)/A, \quad A = 10\gamma_2 - 12\gamma_1^2 + 12.$$

Substituting in (13) for c_1, c_2 we find

$$\eta(u) = \frac{\gamma_1(\gamma_2 + 6) + 2(5\gamma_2 + 6\gamma_1^2 + 6)u}{3\gamma_1^2 - 4(\gamma_2 + 3) - \gamma_1(\gamma_2 + 6)u - (2\gamma_2 - 3\gamma_1^2)u^2}. \qquad (14)$$

For the GARCH(p, q) model (1), (9), (11), the conditional density of y_t is defined as

$$\varphi(y_t \mid \Delta_t) = \frac{1}{h_t^{1/2}} f(u_t) \qquad (15)$$

and the log-likelihood function for the tth observation is

$$\ell_t(\theta) = -\frac{1}{2} \log h_t + \log f(u_t) \qquad (16)$$

where $\theta' = (b', \omega') = (b', \alpha', \beta')$ the set of the unknown parameters.

So that the conditional log-likelihood function for all observation is

$$L(\theta) = \sum_{t=1}^{T} \left(-\frac{1}{2} \, log\, (h_t) + log\, f(u_t) \right). \tag{17}$$

The log-likelihood function (16) depends on the functional form of the density $f(u)$ which includes the unknown shape parameters γ. However, if we assume that the shape parameters γ are known, then using (16) we can define the score vector, corresponding to the mean and variance parameters, for the tth observation, $s_t(\theta)$, as

$$\begin{aligned} s_t(\theta) &= \frac{\partial \ell_t(\theta)}{\partial \theta} = -\frac{1}{2h_t}\frac{\partial h_t}{\partial \theta} + \frac{\partial\, log\, f(u_t)}{\partial u_t}\frac{\partial u_t}{\partial \theta} \\ &= -\frac{1}{2h_t}\frac{\partial h_t}{\partial \theta} + \eta(u_t)\frac{\partial u_t}{\partial \theta}. \end{aligned} \tag{18}$$

Since $\theta' = (b', \omega')$ we obtain from (18)

$$s_t(b) = \frac{\partial \ell_t(\theta)}{\partial b} = -\frac{1}{2h_t}\frac{\partial h_t}{\partial b} + \eta(u_t)\frac{\partial u_t}{\partial b} \tag{19}$$

$$s_t(\omega) = \frac{\partial \ell_t(\theta)}{\partial \omega} = -\frac{1}{2h_t}\frac{\partial h_t}{\partial \omega} + \eta(u_t)\frac{\partial u_t}{\partial \omega} \tag{20}$$

where the derivatives of u_t with respect to b and ω are

$$\frac{\partial u_t}{\partial b} = -\frac{1}{h_t^{1/2}} x_t - \frac{u_t}{2h_t}\frac{\partial h_t}{\partial b} \tag{21}$$

$$\frac{\partial u_t}{\partial \omega} = -\frac{u_t}{2h_t}\frac{\partial h_t}{\partial \omega}. \tag{22}$$

Substituting (21) in (19) and (22) in (20) we obtain for the score vector

$$s_t(b) = -\frac{\eta(u_t)}{2h_t} x_t - \frac{1}{2h_t}(\eta(u_t)u_t + 1)\frac{\partial h_t}{\partial b} \tag{23}$$

$$s_t(\omega) = -\frac{1}{2h_t}(\eta(u_t)u_t + 1)\frac{\partial h_t}{\partial \omega} \tag{24}$$

where

$$\frac{\partial h_t}{\partial b} = -\sum_{i=1}^{p} \alpha_i \varepsilon_{t-i} x_{t-i} + \sum_{i=1}^{q} \beta_i \frac{\partial h_{t-i}}{\partial b} \tag{25}$$

$$\frac{\partial h_t}{\partial \omega} = z_t + \sum_{i=1}^{q} \beta_i \frac{\partial h_{t-i}}{\partial \omega}. \tag{26}$$

The Full Information Maximum Likelihood estimation of all the unknown parameters θ and γ is impossible, since the exact functional form of the density $f(u)$ is unknown. For this reason the following estimation procedure is used.

Assume some initial consistent estimates $\hat{\theta}$ for the set of the unknown parameters θ; these estimates may come from the application of the QML as defined in Bollerslev and Wooldridge (1992). Using these estimates we can estimate the residuals $\hat{\varepsilon}_t$, the variance $\hat{h}_t$ and the standardized residuals $\hat{u}_t = \hat{\varepsilon}_t / \hat{h}_t^{1/2}$. The shape parameters $\gamma' = (\gamma_1, \gamma_2)$ may be estimated consistently from the standardized residuals $\hat{u}_t$ as

$$\hat{\gamma}_1 = \frac{1}{T} \sum_{t=1}^{T} \hat{u}_t^3, \ \hat{\gamma}_2 = \frac{1}{T} \sum_{t=1}^{T} \hat{u}_t^4 - 3.$$

Lastly, by using $\hat{\theta}$, $\hat{\gamma}$, $\hat{\varepsilon}_t$, $\hat{u}_t$, $\hat{h}_t$ we can define the one-step scoring estimator for the unknown parameters θ as

$$\tilde{\theta} = \hat{\theta} + k\hat{c} \tag{27}$$

where

$$\hat{c} = \left(\sum_{t=1}^{T} s_t(\hat{\theta}) s_t(\hat{\theta})' \right)^{-1} \sum_{t=1}^{T} s_t(\hat{\theta}) \tag{28}$$

is the correction vector and the constant $k \in (0, 1)$.

The correction vector (28) may be easily calculated from the regression of a vector m with all elements equal to one on a matrix Z with rows

$$z_t = \left(s_t(\hat{b}), s_t(\hat{\alpha}), s_t(\hat{\beta}) \right)$$

so will be

$$\hat{c} = (Z'Z)^{-1}Z'm. \tag{29}$$

We shall refer to $\tilde{\theta}$ as the Pearson Improved (PI) estimator for the case of the GARCH(p, q) model (1), (12). If $q = 0$ then the GARCH(p, q) model reduces to the ARCH(p) model and the estimator (27) will be the Pearson Improved (PI) estimator for the ARCH(p) model.

The use of $\hat{\gamma}$ in the estimation procedure can affect the efficiency of the estimator. The PI estimator would be fully efficient only in the special case of full adaptation. In the case considered here the PI estimator is not fully adaptive, since the conditions for adaptivity as they are cited in Bickel (1982) and Manski (1984) are not valid. On the other hand the proposed estimator is adaptive in the case of the simple GARCH model introduced by Gonzalez-Rivera and Racine (1995), since the conditions for successful adaptation cited therein are valid for our case.

3 Sampling Experiments

To examine the relative efficiency of the proposed estimator Monte Carlo experiments were carried out. To generate the data we used the same model as in Engle and Gonzalez-Rivera (1991), that is

$$\begin{aligned} y_t &= b_1 y_{t-1} + b_2 y_{t-2} + e_t, \\ e_t &= h_t^{1/2} u_t, \\ h_t &= \alpha_0 + \alpha_1 e_{t-1}^2 + \alpha_2 e_{t-2}^2 + \beta_1 h_{t-1}, \\ b_1 &= 0.5, b_2 = 0.15, \\ \alpha_0 &= 0.1, \alpha_1 = 0.1, \alpha_2 = 0.2, \beta_1 = 0.6. \end{aligned} \tag{30}$$

The disturbance term u_t was generated from one of the distributions that are given in Tables 1.1 and 1.2.

First, we generate an independent sample of random numbers, say v_t, for each of the distributions in Tables 1.1 and 1.2. Second, we transform each random sample so that it has zero mean and variance one (where it is needed), that is

$$u_t = (v_t - \mu_d)/\sigma_d$$

where μ_d is the mean and σ_d is the standard deviation of each distribution in Tables 1.1 and 1.2.

Table 1.1. *Distributions that belong to the PSD, with finite γ's*

		Distribution	μ_d	σ_d	γ_1	γ_2
1.	$N(0,1)$	Standard Normal	0.000	1.000	0.000	0.000
2.	$t(5)$	Student t with $v = 5$ d.f.	0.000	1.291	0.000	6.000
3.	$B(0.5,3)$	Beta with parameters $a = 0.5$, $b = 3$	0.143	0.165	1.575	2.224
4.	$B(4,2)$	Beta with parameters $a = 4$, $b = 2$	0.667	0.178	−0.468	−0.375
5.	$G(0.5,1)$	Gamma with parameters $a = 0.5$, $b = 1$	0.500	0.707	2.828	12.000
6.	$G(5,2)$	X^2, Chi square with $v = 10$ d.f.	10.000	4.472	0.894	1.200
7.	$F(2,9)$	Fisher F with $v_1 = 2$ and $v_2 = 9$ d.f.	1.286	1.725	5.466	146.444

Table 1.2. *Distributions that belong to the PSD, with no finite γ's or do not belong to the PSD*

		Distribution	μ_d	σ_d	γ_1	γ_2
1.	$t(5)$	Student t with $v = 3$ d.f.	0.000	1.732	—	—
2.	$LN(0,1)$	LogNormal, the distribution of $\exp(z)$, where z is distributed as $N(0, 1)$	1.649	2.161	6.185	110.936
3.	$W(2,1)$	Weibull with parameters $a = 2$, $b = 1$	0.886	0.463	0.631	0.245
4.	$W(8,2)$	Weibull with parameters $a = 8$, $b = 2$	1.027	0.152	−0.534	0.328
5.	VCN	Variance Contaminated Normal $0.9N(0, 0.1) + 0.1N(0, 9)$	0.000	0.995	0.000	21.821
6.	BSM	Bimodal Symmetric Mixture of two Normals, $0.5N(-3, 1) + 0.5N(3, 1)$	0.000	3.162	0.000	−1.620

Then using the model described by relations (30), we recursively generate the variable y_t. To avoid starting problems we generate for y_t 10 percent more observations than are required for each sample size and then reject the first 10 percent of the generated observations.

For each distribution consider samples of $T = 500, 1000, 2000$ which gives a total of 39 experiments. Each experiment consists of 5,000 replications and is executed by a double precision Fortran program. The pseudo-random numbers were generated by NAG/WKSTN subroutines and by (tested) subroutines written by the authors.

The estimation procedure is the following.

First, estimate the parameters $\theta' = (b_1, b_2, \alpha_0, \alpha_1, \alpha_2, \beta_1)$ by using the QML procedure, as it is described in Bollerslev and Wooldridge (1992).

That is, maximize the log-likelihood function (17) by using the iterative procedure based on the Berndt *et al.* (1974) (BHHH) algorithm. The convergence criterion used in the BHHH updating regression was $R^2 < 0.001$. Second, use the QML estimates for the parameters θ to calculate the residuals $\hat{\varepsilon}_t$, the variance $\hat{h}_t$, the standardized residuals $\hat{u}_t$ and the shape parameters $\hat{\gamma}_1$, $\hat{\gamma}_2$. Using these estimates construct the matrix Z and apply (29) to obtain the improvement vector $\hat{c}$. Third, as in Magdalinos and Mitsopoulos (2003), calculate the constant $k = (1/exp(\|c\|))^2$, where $\|c\|$ is the Euclidian norm of the improvement vector (29). Lastly, by applying (27), obtain the PI estimator for the parameters θ.

Moreover, as in Engle and Gonzalez-Rivera (1991), we transform the residuals $\hat{u}_t$ in order to have mean 0 and variance 1 (where it is needed) and following the above steps produce the PIs estimator.

For each distribution we produce Monte Carlo estimates of the Bias and the Standard Deviation (StDev) of the QML estimator and of the estimators PI and PIs. Moreover, the efficiency gains of the proposed estimators were calculated by using the following formula

$$Efficiency\ Gains = 1 - \frac{StDev_i}{StDev_{QML}}, \quad i = PI, PIs$$

The results for the parameter Bias, Standard Deviation and the estimates of the efficiency gains are presented in the Tables 1.3 through 1.8. Moreover, Figures 1.1 through 1.6 illustrate the estimates for the efficiency gains.

As expected, the variability of the estimators in terms of standard deviation decreases as the sample size increases from $T = 500$ to $T = 2000$. More analytically, the performance of the QML estimator varies across error distributions and seems to depend on the kurtosis coefficient of the error distribution. The highest standard deviation of the QML estimator is observed in the cases of the distributions $LN(0,1)$ and $F(2,9)$; these distributions also present the highest kurtosis (see Tables 1.1 and 1.2). In the case of the distributions VCN, $G(0.5,1)$ $t(5)$,$B(0.5,3)$ and $G(5,2)$, where $\gamma_2 > 1$, the QML estimator shows a higher standard deviation than in the case of the distributions $W(8,2)$, $W(2,1)$ and $B(4,2)$, where $\gamma_2 < 1$. The standard deviation of the last distributions varies around that of the $N(0,1)$ distribution which is the ideal case for the QML estimator. Moreover, the QML estimator performs well in the case of the $t(3)$distribution which has no definite skewness and kurtosis and in the case of the BSM distribution which is bimodal.