3-1-1993

Adaptive Estimation in ARCH Models

Oliver B. Linton

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
AT YALE UNIVERSITY

Box 2125 Yale Station
New Haven, Connecticut 06520

COWLES FOUNDATION DISCUSSION PAPER NO. 1054

Adaptive Estimation in ARCH Models

by

Oliver Linton

March 1993

ADAPTIVE ESTIMATION IN ARCH MODELS

Oliver Linton[1]
Cowles Foundation for Research in Economics
Yale University

December 1991
Revised: March 1993

Abstract

We construct efficient estimators of the identifiable parameters in a regression model when the errors follow a stationary parametric ARCH(P) process. We do not assume a functional form for the conditional density of the errors, but do require that it be symmetric about zero. The estimators of the mean parameters are adaptive in the sense of Bickel [2]. The ARCH parameters are not jointly identifiable with the error density. We consider a reparameterisation of the variance process and show that the identifiable parameters of this process are adaptively estimable.

1. INTRODUCTION

We consider the problem of obtaining efficient estimators of the identifiable parameters in the following linear regression model where the errors are conditionally heteroskedastic according to an ARCH(P) process:

$$y_t = \beta^T x_{t_P} + u_t \; ; \; u_t = \varepsilon_t \sigma_t \; ; \tag{1.1}$$

$$\sigma_t^2 = a + \sum_{j=1}^{} c_j u_{t-j}^2, \; t=1,2,..,T. \tag{1.2}$$

This specification of the error process was originally suggested in Engle [10], and was employed there to model UK inflation rates. It has been used in countless empirical studies – see the survey papers of Engle and Bollerslev [12] and Bollerslev, Chou and Kroner [7] for references.

The ARCH specification rationalizes two well established empirical regularities about financial and macroeconomic time series. When c_j, $j=1,2,..,P$ are all positive, the process σ_t^2 is positively serially dependent. This is an important feature: many financial and macroeconomic time series are characterized by episodic bursts of volatility followed by more tranquil periods. Uncertainty about future events – and the consequent risk to investors – varies over time yet typically is closely related to previous assessments of uncertainty.

In addition, even when ε_t are i.i.d normal, the unconditional distribution of the innovation u_t will be leptokurtic by virtue of the mixing that random σ induces. This is consistent with the findings of

many researchers – e.g. Mandelbrot [28] and Gallant et al [17] – who have found that the distribution of stock returns tends to have heavier tails than the normal.

A number of generalizations of the ARCH specification are given in Engle and Bollerslev [12] and Bollerslev, Chou, and Kroner [7]. Recent research has focused on Generalized ARCH and Integrated GARCH models – see Bollerslev [5] and Lumsdaine [27], on exponential ARCH/GARCH models – see Geweke [18] and Nelson [30], and on semiparametric ARCH/GARCH modelling – see Engle and Gonzalez-Rivera [13] and Whistler [45]. A number of authors have proposed alternatives to the ARCH paradigm. In particular, Spanos [39] and Shephard [38] highlight deficiencies in the ARCH modelling approach and suggest alternatives.

In this paper we are concerned with the semiparametric approach. Specifying a parametric model for the density of ε_t imposes strong restrictions on the data generation process, especially when the normal distribution is used. Although Bollerslev [5] and Weiss [44] show that the Gaussian (Pseudo) Maximum Likelihood Estimator (hereafter the PMLE) of $\theta \equiv (\beta^T, a, c^T)^T$, where $c \equiv (c_1, c_2, .., c_p)^T$, is $\sqrt{T}$ consistent asymptotically normal under quite general conditions on the error density f, this estimator is inefficient for non-normal f's. With the large sample sizes available for financial data one ought to be able to do better.

Engle and Gonzalez-Rivera [13] consider a semiparametric extension of (1.1) and (1.2). They retained the linear relationship (1.2) yet allowed f(.) to be of unknown form. They used nonparametric estimates of the score function of ε to estimate the parameters of a GARCH

process. They report monte carlo simulation results that suggest improvements for this method over the Gaussian PMLE when the true error distribution was non-normal.

We examine further this semiparametric model. The issues we address here are twofold. Firstly, what is the information bound for estimation of θ when no parametric structure is assumed for the error density? In particular, is this an *adaptive* situation – can one in principle estimate θ as well when f is unknown as when it is known? Secondly, is it possible to construct an estimate of θ that achieves the information bound asymptotically?

Bickel [2] gives a necessary condition for adaptation in the context of a semiparametric model $P_{\theta,G}$, where θ is a finite dimensional parameter and G is an infinite dimensional nuisance parameter: the scores for θ must be orthogonal to the tangent space for G. In particular, the scores for θ must be orthogonal to the scores for the scalar parameter τ for each parameterization $G(.;\tau)$ of $G(.)$.

This orthogonality condition is satisfied in a number of semiparametric models. In particular, Bickel [2] shows that the information bound for estimating the slope parameters in a linear regression is the same whether or not the error density is known. Kreiss [21,22] extends these results to a time series context. He shows that it is possible to estimate the identifiable parameters of a stationary invertible ARMA model adaptively in the presence of an unknown error density. His results are considerably easier to derive when the error distribution is symmetric.

Engle and Gonzalez-Rivera [13] examined the performance of their semiparametric estimator when the true error density was either t_5 or gamma distributed. They found that although the estimator generally outperformed the Gaussian PMLE, it appeared to be considerably less efficient than the MLE. They suggested that the semiparametric estimator was not adaptive even when the error density was symmetric.

In this paper, we consider only the situation where the unknown error density is symmetric about zero. We find that under the specification (1.1) and (1.2) the mean parameters β can be estimated adaptively when the error density is unknown, while the parameters (a,c,f) are not jointly identifiable. To deal with the identifiability problem we reparameterise (1.2) as

$$\sigma_t^2(\theta) = e^{\alpha}[\ 1 + \sum_{j=1}^{P} \gamma_j u_{t-j}^2]. \qquad (1.3)$$

This parameterization separates the overall scale effect (e^{α}) from the relative effects measured by γ_j, j=1,2,..,P. We find that the parameter α has zero information when f is unknown, while the scores for $\gamma \equiv (\gamma_1, \gamma_2, .., \gamma_P)^T$ are orthogonal to the tangent space for the unknown error density – i.e. γ is in principle adaptively estimable. We construct an estimator of the identifiable parameters and show that our estimator is asymptotically equivalent to the MLE, and hence is adaptive. To establish the asymptotic properties of our estimator we use the methodology developed in Le Cam [25], Bickel [2], Kreiss [21,22] and Swensen [43].

Engle and Gonzalez-Rivera [13] employ a different approach to

ensure semiparametric identifiability – they normalize the variance of f to be one. This introduces a nonlinear constraint on the class of allowable error densities which is difficult to incorporate in information bound calculations. It is not clear whether this approach will give rise to adaptively estimable parameters other than β, although the preliminary calculations presented in Steigerwald [41] suggest it may not.

The paper is structured as follows. In Section 2 we examine whether Bickel's orthogonality condition holds in the original ARCH model and in our reparameterisation. In Section 3 we state our assumptions. In Section 4 we establish the fundamental LAN property for our reparameterised ARCH model. In Section 5 we establish properties of linearised MLE's of the unknown parameters when the error density is known. In Section 6 we construct an estimator that does not require knowledge of the error density and is adaptive. Section 7 concludes.

2. IS ADAPTIVE ESTIMATION POSSIBLE?

2.1 The Location Scale Model

We first review the theory developed in Bickel et al [3] concerning information bounds in semiparametric models in the context of the location scale model

$$y = \mu + \varepsilon\sigma,$$

where ε is distributed symmetrically about zero with density f. When f is known, the scores for the unknown parameters (μ,σ) are

$$\dot{\ell}_{\mu}(\varepsilon) = -\sigma^{-1}\frac{f'}{f}(\varepsilon) \equiv -\sigma^{-1}\psi_1(\varepsilon) \; ; \; \dot{\ell}_{\sigma}(\varepsilon) = -\sigma^{-1}[\varepsilon\frac{f'}{f}(\varepsilon)+1] \equiv -\sigma^{-1}\psi_2(\varepsilon).$$

These scores are mutually orthogonal when f is symmetric about zero. In this case, the information bounds for estimating μ and σ when f is known are given by $\sigma^2 I_1(f)^{-1}$ and $\sigma^2 I_2(f)^{-1}$ respectively, where

$$I_1(f) = E[\psi_1(\varepsilon)^2], \; I_2(f) = E[\psi_2(\varepsilon)^2],$$

We verify Bickel's orthogonality condition using the following heuristic argument. Suppose that f is parameterized by a scalar parameter τ such that $f(.;\tau)$ is symmetric about zero for all τ, and let $f_{\tau}(.;\tau)$ denote the partial derivative of $f(.;\tau)$ with respect to τ. Since $f(.;\tau)$ is symmetric about zero for all τ,

$$f_{\tau}(.;\tau) = \operatorname*{Lim}_{\delta \to 0} \frac{f(.;\tau+\delta)-f(.;\tau)}{\delta}$$

is also symmetric about zero. The score function for τ in the parametric model P_{θ}, where $\theta=(\mu,\sigma,\tau)^T$, is $\dot{\ell}_{\tau}$, where

$$\dot{\ell}_{\tau}(\varepsilon) = \frac{f_{\tau}}{f}(\varepsilon),$$

and is therefore symmetric about zero. Furthermore, $\dot{\ell}_{\tau}$ is orthogonal to $\dot{\ell}_{\mu}$, and the information bound for estimating μ in the presence of

the unknown parameter τ is $I_1(f)$ – knowledge of τ provides no useful information about μ. Since the parameterization was arbitrary, we conclude that knowledge of the error density f is irrelevant as far as estimation of μ is concerned. However, the parameter σ is not identifiable without further restrictions on f.

Suppose that instead of the natural parameterization (μ,σ) one had parameterized the model by (ζ,η), where $\zeta=(\mu+\sigma)/2$ and $\eta=(\mu-\sigma)/2$. In this case, the scores for ζ and for η are both correlated with those for τ, and one might conclude that this was not an adaptive situation[2]. We suggest that the ARCH model manifests this phenomenon – in (1.2) each of the parameters a and c_j, j=1,2,..,p gives information about scale.

2.2 Engle's ARCH Model

We now examine the ARCH model defined by (1.1) and (1.2). Let x_t be a K by 1 vector of fixed regressors, and suppose that ε_t is i.i.d, zero mean, with density f symmetric about zero. Let $(\beta^T,a,c^T)^T\equiv\phi$, where $c=(c_1,c_2,..,c_p)^T$ and let the zero subscript denote the true parameter value where necessary. Furthermore, suppose that the initial conditions $Y_0=(\varepsilon_0,..,\varepsilon_{1-P},\sigma_0^2,..,\sigma_{1-P}^2)$ are observed, and let $f_0(Y_0;\phi)$ denote the unconditional density of Y_0.

The sample log likelihood $\ell(Y_0,Y_1,..,Y_T;\phi)$ for the ARCH model (1.1) and (1.2) can be written as

$$\ell = \log f_0(Y_0;\phi) + \sum_{t=1}^{T} \log f(\varepsilon_t(\phi)) - \frac{1}{2}\sum_{t=1}^{T} \log[\sigma_t^2(\phi)],$$

where

$$\varepsilon_t(\phi) = (y_t - \beta^T x_t)/\sigma_t(\phi) \;\; ; \;\; \sigma_t^2(\phi) = a + \sum_{j=1}^{P} c_j (y_{t-j} - \beta^T x_{t-j})^2 .$$

We shall assume that the process $\sigma_t^2(\phi_0)$ is stationary, and that f_0 makes a vanishingly small contribution to the asymptotic properties of the MLE. We focus our attention on the conditional likelihood that drops f_0. Let $\Delta_{t\phi}(\phi) = (\Delta_{t\beta}^T, \Delta_{ta}, \Delta_{tC}^T)^T$ denote the K+P+1 vector of period t contributions to the sample scores of the conditional likelihood, and let $\dot{\ell}_{T\phi} = (\dot{\ell}_{T\beta}^T, \dot{\ell}_{Ta}, \dot{\ell}_{TC}^T)^T$, where $\dot{\ell}_{T\phi}(\phi) \equiv \sum_{t=1}^{T} \Delta_{t\phi}(\phi)$. Then

$$\dot{\ell}_{T\beta}(\phi) = -\sum_{t=1}^{T} \{\sigma_t(\phi)^{-1} x_t \psi_1(\varepsilon_t(\phi)) + W_t(\phi)\psi_2(\varepsilon_t(\phi))\} \equiv \dot{\ell}_{T\beta 1}(\phi) + \dot{\ell}_{T\beta 2}(\phi),$$

where

$$W_t(\phi) = -\sum_{j=1}^{P} c_j x_{t-j} (y_{t-j} - \beta^T x_{t-j})/\sigma_t^2(\phi) = -\sum_{j=1}^{P} c_j x_{t-j} w_{t-j}(\phi),$$

and

$$w_{t-j}(\phi) = (y_{t-j} - \beta^T x_{t-j})/\sigma_t^2(\phi).$$

Both $W_t(\phi)$ and $\sigma_t^2(\phi)$ depend only on the past. Similarly,

$$\dot{\ell}_{TC}(\phi) = -\frac{1}{2}\sum_{t=1}^{T} \psi_2(\varepsilon_t(\phi)) v_t(\phi) \;\; ; \;\; \dot{\ell}_{Ta}(\phi) = -\frac{1}{2}\sum_{t=1}^{T} \psi_2(\varepsilon_t(\phi)) \sigma_t(\phi)^{-2},$$

where $v_t = (v_{1t}, v_{2t}, .., v_{Pt})^T$, and

$$v_{jt}(\phi) = (y_{t-j} - \beta^T x_{t-j})^2/\sigma_t^2(\phi), \quad j=1,2,..,P$$

which also depends only on the past.

To investigate whether Bickel's orthogonality conditions are met in the ARCH model we proceed heuristically as in Section 2.1. Suppose that f is parameterized by a scalar parameter τ. In this case

$$\dot{\ell}_{T\tau} = \sum_{t=1}^{T} \frac{f_\tau}{f}(\varepsilon_t),$$

where $\frac{f_\tau}{f}(.)$ is symmetric about zero. Although $\Delta_{t\beta 2}$ is an even function of ε_t, it is orthogonal to $\Delta_{t\tau}$: since W_t is independent of ε_t and is mean zero, $\dot{\ell}_{T\tau}$ and $\dot{\ell}_{T\beta}$ are mutually orthogonal. Thus there is no efficiency loss from not knowing τ. Since we only exploit symmetry in obtaining this orthogonality, this result carries over to the semiparametric model. One should be able to construct adaptive estimates of β, provided one can estimate the score functions ψ_i suitably well.

This orthogonality does not hold for the remaining parameters, since $\dot{\ell}_{Ta}$, $\dot{\ell}_{Tc}$, and $\dot{\ell}_{T\tau}$ are in general correlated. We argue that this correlation is a manifestation of the fact that we cannot separately identify (a,c,f). Before discussing information bounds in this model we must deal with this issue.

We reparameterize the ARCH process according to (1.3), i.e.

$$\sigma_t^2(\theta) = e^{\alpha}[1 + \sum_{j=1}^{P} \gamma_j (y_{t-j} - \beta^T x_{t-j})^2],$$

where $\theta \equiv (\beta^T, \alpha, \gamma^T)^T$, and $\gamma = (\gamma_1, \gamma_2, .., \gamma_P)^T$. Now let $\dot{\ell}_{T\theta}(\theta) \equiv \sum_{t=1}^{T} \Delta_{t\theta}(\theta)$, where $\Delta_{t\theta}(\theta) = (\Delta_{t\beta}^T, \Delta_{t\alpha}, \Delta_{t\gamma}^T)^T$ and $\dot{\ell}_{T\theta} = (\dot{\ell}_{T\beta}^T, \dot{\ell}_{T\alpha}, \dot{\ell}_{T\gamma}^T)^T$. Then

$$\dot{\ell}_{T\beta}(\theta) = -\sum_{t=1}^{T} \{\sigma_t(\theta)^{-1} x_t \psi_1(\varepsilon_t(\theta)) + \tilde{W}_t(\theta)\psi_2(\varepsilon_t(\theta))\} \qquad (2.1)$$

$$\equiv \dot{\ell}_{T\beta 1}(\theta) + \dot{\ell}_{T\beta 2}(\theta),$$

where

$$\tilde{W}_t(\theta) = -\sum_{j=1}^{P} \gamma_j x_{t-j}(y_{t-j} - \beta^T x_{t-j})/\tilde{\sigma}_t^2(\theta) = -\sum_{j=1}^{P} \gamma_j x_{t-j} \tilde{w}_{t-j}(\theta),$$

$$\tilde{\sigma}_t^2(\theta) = 1 + \sum_{j=1}^{P} \gamma_j (y_{t-j} - \beta^T x_{t-j})^2 \; ; \; \tilde{w}_{t-j}(\theta) = (y_{t-j} - \beta^T x_{t-j})/\tilde{\sigma}_t^2(\theta) \; ;$$

while

$$\dot{\ell}_{T\alpha}(\theta) = -\tfrac{1}{2}\sum_{t=1}^{T} \psi_2(\varepsilon_t(\theta)) \; ; \; \dot{\ell}_{T\gamma}(\theta) = -\tfrac{1}{2}\sum_{t=1}^{T} \psi_2(\varepsilon_t(\theta))\tilde{v}_t(\theta), \qquad (2.2)$$

where

$$\tilde{v}_{jt}(\theta) = (y_{t-j} - \beta^T x_{t-j})^2/\tilde{\sigma}_t^2(\theta), \; j=1,2,..,P.$$

Notice that $\sigma_t(\theta_0)^{-1}$, $\tilde{w}_t(\theta_0)$, and $\tilde{v}_t(\theta_0)$ are all stationary ergodic processes, and are all bounded from above when $\alpha_0 > -\infty$ and $\gamma_j \geq 0$, $j=1,2,..,P$.

The efficient score function for γ in the presence of unknown α, obtained by projecting $\dot{\ell}_{T\gamma}$ onto $\dot{\ell}_{T\alpha}$, is

$$\dot{\ell}^{*}_{T\gamma} = \dot{\ell}_{T\gamma} - E[\dot{\ell}_{T\gamma}\dot{\ell}_{T\alpha}]\{E[\dot{\ell}^{2}_{T\alpha}]\}^{-1}\dot{\ell}_{T\alpha} = -\frac{1}{2}\sum_{t=1}^{T}\psi_{2}(\varepsilon_{t})(\tilde{v}_{t}-\bar{v}) \quad ;$$

where $\bar{v}=E[\tilde{v}_{t}]$. Now $\dot{\ell}^{*}_{T\gamma}$ is orthogonal to any score function $\dot{\ell}_{T\tau}$, where

$$\dot{\ell}_{T\tau} = \sum_{t=1}^{T}\frac{f_{\tau}}{f}(\varepsilon_{t}),$$

for $f(.;\tau)$ any parameterization of the symmetric function $f(.)$. Therefore, the information bound for estimating γ is the same whether or not f is known. Under suitable regularity conditions we should be able to estimate γ adaptively.

In the sequel we construct an estimator of $(\beta^{T},\gamma^{T})^{T}$ that achieves the information bound provided that f is symmetric about zero.

REMARK: Consider the exponential ARCH model:

$$\log[\sigma_{t}^{2}] = \alpha + \sum_{j=1}^{P}\gamma_{j}r(\varepsilon_{t-j}\sigma_{t-j}), \qquad (2.1)$$

where $r(.)$ is a known function, see Nelson [30]. If $E[r'(\varepsilon_{t}\sigma_{t})]=0$, the scores for β are orthogonal to those for f. In this case, we argue in Appendix I that both β and γ are in principle adaptively estimable, although see Bickel and Ritov [3] for a cautionary tale in this regard.

3. ASSUMPTIONS

Although (1.2) and (1.3) generate the same family of probability measures, the relevant parameter spaces differ. To avoid any ambiguity we shall restrict our attention to parameterization (1.3). We use the following conditions.

A1. The random variables $\{\varepsilon_t\}$ are i.i.d., with absolutely continuous Lebesgue density f, where $f(x)>0$ $\forall x\in\mathbb{R}$.

Let the score functions, ψ_1 and ψ_2 be defined as follows:

$$\psi_1(x) = \frac{f'}{f}(x) \; ; \; \psi_2(x) = x\,\frac{f'}{f}(x)+1,$$

where $\frac{1}{2}f^{-1/2}(x)f'(x)$ is the quadratic mean derivative of $f(x)^{1/2}$, i.e.

$$\lim_{m\to 0}\frac{1}{m^2}\int\left[\, f(x+m)^{1/2} - f(x)^{1/2} - \frac{m}{2}\, f^{-1/2}(x)f'(x)\,\right]^2 dx = 0. \quad (3.1)$$

We do not assume that f is necessarily differentiable everywhere in the usual sense, although the following assumptions restrict the lack of smoothness that can be permitted.

A2. The density f has finite Fisher information for both scale and location parameters,

$$0 < I_1(f) = \int \psi_1(x)^2 f(x)dx < \infty \; ; \; 0 < I_2(f) = \int \psi_2(x)^2 f(x)dx < \infty.$$

A3. The score functions, ψ_i i=1,2, satisfy the following conditions:

$$(1)\ \int \{\psi_1((x+m)/(1+s))-\psi_1(x)\}^2 f(x)dx \Rightarrow 0 \quad as\ m,s \Rightarrow 0,\ i=1,2,$$

$$(2)\ \int m^{-1}\psi_1((x+m)/(1+s))f(x)dx \Rightarrow -I_1(f)\ as\ m,s \Rightarrow 0,$$

$$(3)\ \int s^{-1}\psi_2((x+m)/(1+s))f(x)dx \Rightarrow -I_2(f)\ as\ m,s \Rightarrow 0,$$

$$(4)\ \int s^{-1}\psi_1((x+m)/(1+s))f(x)dx \Rightarrow 0\ as\ m,s \Rightarrow 0,$$

$$(5)\ \int m^{-1}\psi_2((x+m)/(1+s))f(x)dx \Rightarrow 0\ as\ m,s \Rightarrow 0.$$

REMARK: These are essentially second derivative conditions, and are satisfied by a large class of densities: for example, the normal, the GED distribution considered in Nelson [30], and the Laplace distribution. Condition A3(1) is an obvious extension of condition A5(1) in Kreiss [21], while condition A3(2) is condition A5(ii) of Kreiss [21].

REMARK: Lind and Roussas [26] establish in a more general context that quadratic mean differentiability assumptions such as (3.1) and A3 imply Cramer's conditions (see Cramer [9], p500).

A4: The error density also satisfies:

1) The density f is symmetric about zero,

2) $\int x^4 f(x)dx<\infty$,

3) $\int \psi_j(x)^4 f(x)dx<\infty$, $j=1,2$.

A5. The parameter space Θ is an open subset of $\mathbb{R}^{K+P+1}$ that satisfies various restrictions such that

(1) The process $\{\sigma_t^2\}_{t=1}^{\infty}$ is bounded below by a constant $\underline{\sigma}>0$.

(2) The process $\{\sigma_t^2\}_{t=1}^{\infty}$ is strictly stationary and ergodic.

(3) The process $\{\sigma_t^2\}_{t=1}^{\infty}$ satisfies $E[\sigma_t^4]<\infty$.

REMARK: A sufficient condition for A5(1) to hold is that

$$\gamma_1 \geq 0, \ \gamma_2 \geq 0, \ .., \ \gamma_P \geq 0.$$

Nelson and Cao [32] show that these conditions can be weakened somewhat. Primitive conditions on α and γ and on the distribution of the white noise error that imply assumption A5(2) are given in Nemec and Linnell [33]. Similar conditions are given in Nelson [31], Sampson [37] and Bougerol and Picard [8] for the GARCH(1,1) model. Condition A5(3) also requires substantial restrictions on the parameter space as discussed in Bollerslev [5] and Milhoj [29].

For any $\theta\epsilon\Theta$, let $P_{T,\theta}$ be the joint probability measure of a sample $\{y_t,x_t\}_{t=1}^{T}$. In the sequel, unless otherwise stated, we let $\overset{P}{\Rightarrow}$ denote convergence in probability under P_{T,θ_0}, while $o_P(.)$ and $O_P(.)$ will also hold under P_{T,θ_0}. Likewise, $\Rightarrow$ denotes weak convergence of the associated probability measure under P_{T,θ_0}. We make an additional assumption:

A6. The density $f_0(Y_0;\theta)$ is continuous in probability: let $\theta_T=\theta_0+T^{-1/2}h$, and assume that for any $h\in\mathbb{R}^{K+P+1}$ and $\forall\theta_0\in\Theta$

$$f_0(Y_0;\theta_T) \xrightarrow{P} f_0(Y_0;\theta_0) \text{ as } T\to\infty.$$

We assume throughout that the K by 1 vector of explanatory variables x_t are strictly exogenous, and we therefore condition our inference on $\{x_t\}_{t=1}^T$. Define the sequence of K by K matrices

$$M_T(s) = \{m_T(s)_{jk}\} = T^{-1}\sum_{t=s+1}^{T} x_t x_{t-s}^T, \quad s = 0,1,2,..,P.$$

We make the following assumption about the regressors:

B1. The matrix $M_T(0)$ converges to a finite limit $M(0)$, where $M(0)$ is strictly positive definite.

REMARK: This assumption on the regressors could be relaxed to allow trending regressors, for example, by assuming Grenander's conditions. In this case, we must replace the $\sqrt{T}$ norming of our estimator by a suitable matrix as in Swensen [43].

Finally, we shall assume that there exists a $\sqrt{T}$ consistent estimator $\tilde{\theta}_T$ of θ. Recall that $a=e^{\alpha}$ and $c_j=e^{\alpha}\gamma_j$, $j=1,2,..,P$. Weiss [44] and Lumsdaine [27] give conditions under which least squares estimators and Gaussian PMLE's of the parameters a and c are $\sqrt{T}$ consistent. A delta method argument can then be used to establish the $\sqrt{T}$ consistency of the resulting estimators of θ. These authors impose

additional moment conditions of various types.

4. LOCAL ASYMPTOTIC NORMALITY

In this section, we establish that the log-likelihood ratio of the ARCH model (1.1) and (1.3) satisfies the *Local Asymptotic Normality* (LAN) condition defined in Theorem 1 below. This condition, introduced in Le Cam [23], controls the behavior of the log-likelihood ratio in a small neighborhood of the true value, requiring that in large samples it be approximately quadratic in a neighborhood of the true parameter. This regularity is essential when establishing the properties of the Newton-Raphson estimators we consider in later sections.

Le Cam [23], Swensen [42,43], and Roussas [36] give conditions under which the log-likelihood ratio of a general stochastic process satisfies the LAN condition. These conditions have been verified for stationary invertible ARMA processes in Kreiss [21], and for linear regression models with autoregressive errors in Swensen [43]. This latter result was extended by Steigerwald [40] to linear regression models with ARMA errors. Generalizations of this concept to *Locally Asymptotically Mixed Normal* (LAMN) considered in Swenson [42] have found applications in the theory of nonstationary processes – see Phillips [34].

The parameters of interest in the above examples are all location parameters. In the ARCH model, parameters that determine the scale of the process are also of interest. We verify the conditions of Swensen [42] below, using some modifications of the argument presented in

Swensen [43].

We first establish some notation. Define the square root of the likelihood ratio λ to be

$$\lambda(\theta_0,\theta) = \left[\frac{f_0(Y_0;\theta)}{f_0(Y_0;\theta_0)}\right]^{1/2} \prod_{t=1}^{T} \left[\frac{f(\varepsilon_t(\theta))\sigma_t(\theta_0)}{f(\varepsilon_t(\theta_0))\sigma_t(\theta)}\right]^{1/2}.$$

The log likelihood ratio, is defined as

$$\Lambda_T(\theta_0,\theta) = \log\{f_0(Y_0;\theta)/f_0(Y_0;\theta_0)\} + 2\sum_{t=1}^{T} \log \phi_t(\theta_0,\theta)$$

where

$$\phi_t(\theta_0,\theta) = \left[\frac{f(\varepsilon_t(\theta))\sigma_t(\theta_0)}{f(\varepsilon_t(\theta_0))\sigma_t(\theta)}\right]^{1/2}.$$

Let $S_{T\theta}(\theta)=T^{-1/2}\dot{\ell}_{T\theta}=T^{-1/2}\sum_{t=1}^{T}\Delta_{t\theta}$, where $\dot{\ell}_{T\theta}$ are defined in (2.1) and (2.2). Furthermore, let $S_{T\beta}\equiv S_{T\beta_1}+S_{T\beta_2}$, $S_{T\alpha}$, and $S_{T\gamma}$ denote the corresponding subvectors. We now define the information matrix.

Definition: Let the information matrix $J_{\theta\theta}(\theta_0)$ be the probability limit under P_{T,θ_0} of the observed information matrix

$$J_{T\theta\theta}(\theta_0) = T^{-1}\sum_{t=1}^{T} \Delta_{t\theta}(\theta_0)\Delta_{t\theta}(\theta_0)^T.$$

The matrix $J_{\theta\theta}(\theta_0)$ exists by A1, A2, A4, A5, and B1; it is strictly positive definite[3] under A4(b) and A5(3) – see Weiss [44], Lemma 3.2. It has the following structure:

$$J_{\theta\theta} = \begin{bmatrix} J_{\beta\beta} & 0 & 0 \\ 0 & J_{\alpha\alpha} & J_{\gamma\alpha} \\ 0 & J_{\alpha\gamma} & J_{\gamma\gamma} \end{bmatrix} \tag{4.1}$$

where

$$J_{\beta\beta} = \{ \bar{g}_{\beta\beta_1} I_1(f) + \bar{g}_{\beta\beta_2} I_2(f) \} M(0) \equiv J_{\beta\beta_1} + J_{\beta\beta_2},$$

$$J_{\alpha\alpha} = \tfrac{1}{4} \bar{g}_{\alpha\alpha} I_2(f) \; ; \; J_{\gamma\gamma} = \tfrac{1}{4} \bar{g}_{\gamma\gamma} I_2(f) \; ; \; J_{\alpha\gamma} = \tfrac{1}{4} \bar{g}_{\alpha\gamma} I_2(f),$$

while

$$g_{t\beta\beta_1}(\theta) = \sigma_t(\theta)^{-2} \; ; \; g_{t\beta\beta_2}(\theta) = \sum_{j=1}^{P} \gamma_j^2 \tilde{w}_{t-j}(\theta)^2 \; ;$$

$$g_{t\alpha\alpha}(\theta) = 1 \; ; \; g_{t\gamma\gamma} = \tilde{v}_t(\theta)\tilde{v}_t(\theta)^T \; ; \; g_{t\alpha\gamma}(\theta) = \tilde{v}_t(\theta),$$

and $\bar{g}_j = E_{\theta_0}[g_{tj}(\theta_0)]$ for each j.

With these definitions we now state the main theorem of this section.

Theorem 1 (Local Asymptotic Normality) Assume that A1-A6 and B1 hold, and let $\theta_T = \theta_0 + T^{-1/2}h$ *for any* $h \in \mathbb{R}^{K+P+1}$. *Then*

1) $\Lambda_T(\theta_0, \theta_T) - h^T S_{T\theta}(\theta_0) + \frac{1}{2} h^T J_{\theta\theta}(\theta_0, f) h \xrightarrow{P} 0$, *as* $T \to \infty$.

2) $S_T(\theta_0) \to N(0, J_{\theta\theta}(\theta_0, f))$,

REMARK: The asymptotic normality of $S_T(\theta_0)$ under P_{T,θ_0} is easy to establish because $\{\Delta_{t\theta}(\theta_0)\}_{t=1}^{\infty}$ is a sequence of martingale differences

with uniformly bounded variances – Bollerslev and Wooldridge [6] establish a similar result when the Gaussian likelihood is employed.

REMARK: The LAN condition is straightforward to verify when Cramer like differentiability conditions are assumed on the log-likelihood function – see Lind and Roussas [26]. We establish this result under weaker smoothness conditions.

We now outline how Theorem 1 is proved. The iid location scale model considered in Section 2.1 satisfies the LAN condition. In this case, joint quadratic mean differentiability of the square root of the likelihood ratio is sufficient for the LAN condition to hold, as is discussed in Appendix II. To show that this condition holds for the ARCH model requires a conditioning argument.

Let $h=(h_\beta^T, h_\alpha, h_\gamma^T)^T$, then

$$\varepsilon_t(\theta_T) = \frac{(\varepsilon_t + \delta_{T,t})}{(1+\eta_{T,t})^{1/2}},$$

where

$$\delta_{T,t} = (\beta_T - \beta_0)^T x_t/\sigma_t(\theta_0) \;;\; \eta_{T,t} = (\sigma_t^2(\theta_T) - \sigma_t^2(\theta_0))/\sigma_t^2(\theta_0).$$

Substituting for $\sigma_t^2(\theta_T)$ we obtain

$$\delta_{T,t} = T^{-1/2}\sigma_t^{-1} h_\beta^T x_t,$$

$$\eta_{T,t} = \sigma_t^{-2}\{ a_T - a_0 + \sum_{j=1}^{P} [(c_{jT}-c_{j0})\varepsilon_{t-j}^2\sigma_{t-j}^2 - 2c_{j0}x_{t-j}^T(\beta_T-\beta_0)\varepsilon_{t-j}\sigma_{t-j}$$

$$+ c_{j0}(\beta_T-\beta)^T x_{t-j}x_{t-j}^T(\beta_T-\beta_0) - 2(c_{jT}-c_{j0})x_{t-j}^T(\beta_T-\beta_0)\varepsilon_{t-j}\sigma_{t-j}$$

$$+ (c_{jT}-c_{j0})(\beta_T-\beta_0)^T x_{t-j}x_{t-j}^T(\beta_T-\beta_0) \ ] \ \},$$

where $\sigma_t^2=\sigma_t^2(\theta_0)$, $\varepsilon_t=\varepsilon_t(\theta_0)$.

Both $\delta_{T,t}$ and $\eta_{T,t}$ depend only on the regressors and on the past. In addition, we show in Lemma 1.2 of Appendix II that

$$\sum_{t=1}^{T}(\eta_{t,T}^2+\delta_{t,T}^2) < c < \infty \ ; \ \underset{1\le t\le T}{\text{Max}}(\eta_{t,T}^2+\delta_{t,T}^2) \le k(T) \to 0,$$

where k(T) is a deterministic sequence. Thus $\varepsilon_t(\theta_T)$ is close to ε_t. Therefore, the log-likelihood ratio should be well behaved in a neighborhood of the true parameter value.

Let $\{\mathfrak{F}_s: 1\le s\le\infty\}$ be the increasing family of sigma fields such that $\mathfrak{F}_t=\{x_t,d_{t-1},d_{t-2},..,d_0\}$, where $d_t=(y_t,x_t^T)^T$. For convenience sake we define the following quantities

$$X_{T,t} = \left[\frac{f(\varepsilon_t(\theta_T))\sigma_t(\theta_0)}{f(\varepsilon_t(\theta_0))\sigma_t(\theta_T)}\right]^{1/2} - 1 \ ; \ Z_{T,t} = -\tfrac{1}{2}T^{-1/2}h^T\Delta_t\psi(\varepsilon_t),$$

where $\Delta_{1t}=(\tilde{\sigma}_t^{-1}x_t^T,0,..,0)^T$, $\Delta_{2t}=(\tilde{W}_t^T,1,\tilde{v}_t^T)^T$, $\Delta_t=(\Delta_{1t},\Delta_{2t})$, and $\psi=(\psi_1,\psi_2)^T$. The iid vector $\psi(\varepsilon)$ has diagonal covariance matrix I, where $I=\text{diag}\{I_1(f),I_2(f)\}$, while the uniformly bounded K+P+1 by 2 random matrix Δ_t depends only on the past and on the nonstochastic x's, and therefore is independent of $\psi(\varepsilon_t)$. The random variable $Z_{T,t}$ is the (total) quadratic mean derivative of $X_{T,t}$.

The following proposition is given in Swensen [43], we verify these

conditions in Appendix II.

Proposition 1: Assume that the following conditions are satisfied. Then the conclusions of Theorem 1 hold.

$$1)\ \sum_{t=1}^{T} E(X_{T,t}-Z_{T,t})^2 \to 0\ ;\ 2)\ \operatorname{Sup}_T E\{\sum_{t=1}^{T} Z_{T,t}^2\} < \infty\ ;$$

$$3)\ \operatorname{Max}_{1\le t\le T} |Z_{T,t}| \xrightarrow{P} 0\ ;\ 4)\ \sum_{t=1}^{T} Z_{T,t}^2 \xrightarrow{P} h^T J_{\theta\theta}(\theta_0)h > 0\ ;$$

$$5)\ \sum_{t=1} E[Z_{T,t}^2 1(|Z_{T,t}| > \tfrac{1}{2}|\mathfrak{F}_{t-1}] \xrightarrow{P} 0;\ 6)\ E[Z_{T,t}|\mathfrak{F}_{t-1}] = 0.$$

Therefore, the LAN property holds for the ARCH model (1.1) and (1.3). This property has two consequences. Fabian and Hannan [14] show that if the log likelihood ratio satisfies the LAN condition, the *Local Asymptotic Minimax* bound, is achieved by estimators equivalent to the MLE. We discuss this further in the next section.

A second consequence of Theorem 1 is that the sequence of probability measures P_{T,θ_T} and P_{T,θ_0} are contiguous in the sense of Roussas [35] definition 2.1, p7. This means that we can interchange the two measures when we make statements about convergence to zero in probability: for any event A, we have $P_{T,\theta_T}(A)\to 0$ if and only if $P_{T,\theta_0}(A)\to 0$. The estimators we consider are constructed from OLS residuals. The significance of the contiguity property is that it enables us to proceed, in many respects, as if we had the true errors instead of these residuals.

5. ESTIMATION OF θ WHEN THE ERROR DENSITY IS KNOWN.

Subject to regularity conditions, the MLE of θ_0 when f is known is $\sqrt{T}$ consistent asymptotically normal with covariance matrix $J_{\theta\theta}(\theta_0)^{-1}$. In this section we verify that a two step estimator based on an initial $\sqrt{T}$ consistent estimator is asymptotically equivalent to the MLE, and is therefore efficient. The precise notion of efficiency that is appropriate here is the *Locally Asymptotically Minimax* (LAM) criterion of Fabian and Hannan [14] to which paper we refer the reader for a proper definition of this concept. This property is not violated by 'superefficient' estimators unlike the Cramer-Rao lower bound, see Hajek [16]. An alternative efficiency property is that the MLE has the minimal covariance matrix amongst all uniformly asymptotically normal estimators.

We make the following definition:

Definition: A sequence of estimates, $\tilde{\theta}_T$, of θ_0 is asymptotically efficient if it is asymptotically equivalent to the MLE, i.e.

$$\sqrt{T}(\tilde{\theta}_T-\theta_0) = J_{\theta\theta}(\theta_0,f)^{-1}S_{T\theta}(\theta_0) + o_P(1).$$

For technical reasons we shall restrict ourselves to discretised estimators:

Definition: For any sequence of estimators $\tilde{\theta}_T$ define the discretised estimator $\bar{\theta}_T$ to be the nearest vertex of $\{\theta : \theta=n^{-1/2}(i_1,i_2,..,i_p)$, i_j

integers}.

This restriction was employed by Le Cam [25], Bickel [2], and Kreiss [21]. The reason for introducing this concept is that using discretised estimators we can establish the validity of the Newton-Raphson type estimators without introducing additional differentiability or boundedness assumptions. Kreiss [21] Lemma 4.4 establishes that for any sequence of random variables $q_T(\theta)$ if $q_T(\theta_T)=o_p(1)$, where $|\sqrt{T}(\theta_T-\theta_0)|\leq c$ for some constant $c>0$, then $q_T(\bar{\theta}_T)=o_p(1)$ for any discrete and $\sqrt{T}$ consistent estimator $\bar{\theta}_T$. Therefore, we can restrict our attention to nonstochastic sequences θ_T.

We now consider estimation of $J_{\theta\theta}$. There are a number of possible consistent estimators: for example, the outer product of the sample scores $J_{T\theta\theta}(\bar{\theta}_T)$, where $\bar{\theta}_T$ is any discrete and $\sqrt{T}$ consistent estimator of θ. Alternatively, we can exploit the known structure of $J_{\theta\theta}$. Let $\bar{J}_{T\theta\theta}$ be given by (4.1) with M(0) replaced by $M_T(0)$, γ replaced by $\bar{\gamma}_T$, and $\bar{g}_j$ replaced by $\hat{\bar{g}}_j$, where

$$\hat{\bar{g}}_j = T^{-1}\sum_{t=1}^{T} g_{tj}(\bar{\theta}_T),$$

for $\bar{\theta}_T$ a $\sqrt{T}$ consistent discrete estimator of θ_0.

To establish the efficiency of our Newton-Raphson estimator defined below we need to establish that our estimator of $J_{\theta\theta}$ is consistent. This is the content of the following theorem which is proved in Appendix II:

Theorem 2: Let $\bar{\theta}_T$ be a discrete $\sqrt{T}$ consistent estimator of θ. Then $\bar{J}_{T\theta\theta}(\bar{\theta}_T, f)$ is consistent.

We also establish that the following asymptotic linearity holds so that we can approximate the estimator by a function linear in iid random variables.

Theorem 3: (Asymptotic Linearity): Assume that A1-A6 and B1 hold and let $\theta_T=\theta_0+T^{-1/2}h$, for any $h\in\mathbb{R}^{K+P+1}$. Then

$$S_{T\theta}(\theta_T) - S_{T\theta}(\theta_0) = - J_{\theta\theta}(\theta_0, f)\sqrt{T}(\theta_T-\theta_0) + o_P(1).$$

This is proved in Appendix II. We are now able to establish the main result of this section.

Theorem 4: Let $\bar{\theta}_T$ be a discrete and $\sqrt{T}$ consistent estimator of θ_0, and assume that A1-A6 and B1 hold. Let

$$\hat{\theta}_T = \bar{\theta}_T + T^{-1/2}\bar{J}_{T\theta\theta}(\bar{\theta}_T, f)^{-1}S_{T\theta}(\bar{\theta}_T).$$

Then $\hat{\theta}_T$ is efficient.

Therefore, the linearised MLE of θ is asymptotically efficient for a very broad class of densities f. Theorem 4 follows because

$$\sqrt{T}(\hat{\theta}_T-\theta_0) = \sqrt{T}(\bar{\theta}_T-\theta_0) + \bar{J}_{T\theta\theta}(\bar{\theta}_T, f)^{-1}S_{T\theta}(\bar{\theta}_T)$$

$$= \sqrt{T}(\bar{\theta}_T - \theta_0) + J_{\theta\theta}(\theta_0)^{-1} S_{T\theta}(\bar{\theta}_T) + o_p(1)$$

by Theorem 2

$$= \sqrt{T}(\bar{\theta}_T - \theta_0) + J_{\theta\theta}(\theta_0)^{-1}[\, S_{T\theta}(\theta_0) - J_{\theta\theta}(\theta_0)\sqrt{T}(\bar{\theta}_T - \theta_0)\,] + o_p(1)$$

by Theorem 3

$$= J_{\theta\theta}(\theta_0)^{-1} S_{T\theta}(\theta_0) + o_p(1).$$

The results of Theorem 4 complement the existing asymptotic theory for parametric GARCH models described in Weiss [44], Bollerslev and Wooldridge [6] and Lumsdaine [27]. These authors establish asymptotic theory for estimators derived from Gaussian PML and least squares criteria.

6. ESTIMATION OF θ WHEN THE ERROR DENSITY IS UNKNOWN.

We have assumed up to now that the error density is known. We now relax this assumption and construct an estimator that utilizes a consistent estimator of the unknown density.

The first problem we must face is that α and f cannot be separately identified. We can either fix α and let f be unrestricted, or we can estimate α and rescale our estimate of f so that it has unit variance. We assume that α is 0, and is therefore not estimated. We redefine θ so that $\theta \equiv (\beta^T, \gamma^T)^T \in \mathbb{R}^{P+K}$.

For convenience we estimate the unknown score function using the kernel method with a normal density function. Undoubtedly, other kernels could be used, and indeed other nonparametric estimation techniques – such as nearest neighbor, splines, or penalized

likelihood – see Bickel et al [3].

For any b=b(T) let $\phi(x;b)$ denote the density function of a N(0,b(T)) random variable evaluated at x. For any θ we estimate the symmetric error density f by the leave-one-out estimate

$$\hat{f}_{b,t}(x;\theta) = \frac{1}{2(T-1)} \sum_{\substack{s=1 \\ s\neq t}}^{T} \{\phi(x+\varepsilon_s(\theta);b) + \phi(x-\varepsilon_s(\theta);b)\} \quad t=1,2,..,T$$

This estimator of f is symmetric by construction. As in Bickel [2] and Kreiss [22] we trim out excessive contributions to our estimator. We estimate ψ_1 by $\hat{\psi}_{T,t}$, where

$$\hat{\psi}_{T,t}(x;\theta) = \frac{\hat{f}'_{b,t}}{\hat{f}_{b,t}}(x;\theta) \quad \text{if} \begin{cases} \hat{f}_{b(T),t}(x;\theta) \geq d_T \\ |\hat{f}'_{b(T),t}(x;\theta)| \leq c_T \hat{f}_{b(T),t}(x;\theta) \\ |x| \leq e_T \end{cases}$$

$$= 0 \qquad \text{else.}$$

We also define $\hat{I}_{Tj}(\theta,\hat{f})$, where

$$\hat{I}_{T1}(\theta,\hat{f})=T^{-1}\sum_{t=1}^{T}\hat{\psi}_{T,t}(\varepsilon_t(\theta);\theta)^2;\hat{I}_{T2}(\theta,\hat{f})=T^{-1}\sum_{t=1}^{T}[\varepsilon_t(\theta)\hat{\psi}_{T,t}(\varepsilon_t(\theta);\theta)+1]^2.$$

The sample scores are estimated by $\hat{S}_{T\beta}(\theta)=\hat{S}_{T\beta1}(\theta)+\hat{S}_{T\beta2}(\theta)$, and $\hat{S}_{T\gamma}(\theta)$, where

$$\hat{S}_{T\beta1}(\theta) = - T^{-1/2}\sum_{t=1}^{T} \sigma_t(\theta)^{-1}x_t\hat{\psi}_{T,t}(\varepsilon_t(\theta);\theta) \ ;$$

$$\hat{S}_{T\beta2}(\theta) = - T^{-1/2}\sum_{t=1}^{T} [\varepsilon_t(\theta)\hat{\psi}_{T,t}(\varepsilon_t(\theta);\theta)+1]\tilde{W}_t(\theta) \ ;$$

$$\hat{S}_{T\gamma}(\theta) = -\tfrac{1}{2}T^{-1/2}\sum_{t=1}^{T}[\varepsilon_t(\theta)\hat{\psi}_{T,t}(\varepsilon_t(\theta);\theta)+1][\tilde{v}_t(\theta)-\hat{\bar{v}}(\theta)],$$

while $\bar{v}(\theta)$ is estimated by $\hat{\bar{v}}(\theta)=T^{-1}\sum_{t=1}^{T}\tilde{v}_t(\theta)$. In proving Theorem 5 and 6 below, we also utilize a form of sample splitting for $\hat{S}_{T\gamma}(\theta)$ similar to that contained in Bickel [2]. This is purely for technical convenience and is not recommended for applications.

We require the bandwidth and trimming sequences to satisfy the following condition:

Condition C: Assume that b(T),c(T),d(T), and e(T) satisfy*

1) $b(T),\ d(T) \to 0,\ c(T),\ e(T) \to \infty,$

2) $b(T)c(T) \to 0,\ Tb(T)^3c(T)^{-2}e(T)^{-2} \to \infty.$

The additional restriction on the bandwidth sequence is required when estimating the score function $\psi_2(.)$. With these conditions we establish the following theorem in Appendix II.

Theorem 5: Let $\theta_T=\theta_0+T^{-1/2}h$, for any $h\in\mathbb{R}^{K+P}$, and assume that A1-A6 B1 and C hold. Then*

$$\hat{S}_{T\theta}(\theta_T) - S_{T\theta}(\theta_T) = o_p(1).$$

Furthermore, $\hat{I}_{Tj}(\bar{\theta}_T,\hat{f})$ are consistent estimators of $I_j(f)$ for j=1,2.

The information bound for γ in the presence of unknown α is $J^{*-1}_{\gamma\gamma}$,

where

$$J^{*}_{\gamma\gamma} = J_{\gamma\gamma} - J_{\alpha\gamma}J^{-1}_{\alpha\alpha}J_{\gamma\alpha} = \tfrac{1}{4}(\bar{g}_{\gamma\gamma} - \bar{g}_{\alpha\gamma}\bar{g}^{T}_{\alpha\gamma})I_2(f) \equiv \tfrac{1}{4}\,\bar{g}^{*}_{\gamma\gamma}I_2(f).$$

We therefore estimate $J_{\beta\beta}$ and $J^{*}_{\gamma\gamma}$ by

$$\hat{J}_{T\beta\beta}(\bar{\theta}_T,\hat{f}) = \{\hat{\bar{g}}_{\beta\beta 1}\hat{I}_{T1}(\bar{\theta}_T,\hat{f}) + \hat{\bar{g}}_{\beta\beta 2}\hat{I}_{T2}(\bar{\theta}_T,\hat{f})\}M_T(0).$$

$$\hat{J}^{*}_{T\gamma\gamma}(\bar{\theta}_T,\hat{f}) = \tfrac{1}{4}\hat{I}_{T2}(\bar{\theta}_T,\hat{f})\hat{\bar{g}}^{*}_{\gamma\gamma},$$

We now establish the main result of the paper:

Theorem 6: Assume that A1-A6, B1 and condition C^{*} holds. Furthermore, let $\bar{\theta}_T$ be a discretised $\sqrt{T}$ consistent estimator of θ. Let

$$\hat{\theta}_T = \bar{\theta}_T + T^{-1/2}\hat{J}_{T\theta\theta}(\bar{\theta}_T,\hat{f})^{-1}\hat{S}_{T\theta}(\bar{\theta}_T),$$

Then

1) $\hat{J}_{T\theta\theta}(\bar{\theta}_T,\hat{f}) = J_{\theta\theta}(\theta_0,f) + o_p(1),$

2) $\sqrt{T}(\theta_T-\theta_0) = J_{\theta\theta}(\theta_0,f)^{-1}S_{T\theta}(\theta_0) + o_p(1).$

Therefore,

$$\sqrt{T}(\hat{\theta}_T-\theta_0) \Rightarrow N(0,J_{\theta\theta}(\theta_0,f)^{-1}),$$

for all densities f that satisfy our conditions. Furthermore, $\hat{J}_{T\theta\theta}(\bar{\theta}_T, \hat{f})^{-1}$ is a consistent estimator of the asymptotic variance of the adaptive estimator which can be used to form confidence intervals or carry out hypothesis tests.

7. CONCLUSIONS

We have shown how to construct estimates of the identifiable parameters in an ARCH model when the error density is of unknown shape. We have shown that our estimates are adaptive; they have the same asymptotic distribution as the MLE based on the true density. The only substantive restriction we require on the error density is that it be symmetric about zero.

We expect that a number of extensions of these results are possible. Firstly, the assumption of symmetry could be relaxed as in Bickel [2] and Kreiss [22], although it is not known whether γ is adaptively estimable in this case. Secondly, an extension to the GARCH model of Bollerslev [5] should be straightforward following the results of Lumsdaine [27]. It should also be possible to allow the bandwidth parameter b(T) to be data dependent justifying standard cross validation methods for bandwidth choice. Finally, the conditions on our regressors could no doubt be relaxed to allow for trending regressors as well as to include lagged dependent variables.

Bickel's orthogonality condition also holds for the Exponential ARCH model defined in (2.1), provided r(.) is a symmetric function. In

this case we should be able to obtain adaptive estimates of the identifiable parameters, although it remains to provide initial estimates of γ that are $\sqrt{T}$ consistent for any error density. Furthermore, establishing stationarity of the process for given r(.) is not a trivial problem.

We employed a number of techniques to establish the asymptotic theory of our estimator: sample splitting, trimming, and discretisation. In practice, it may be necessary to trim out the score function estimates, but it is generally agreed – see Hsieh and Manski [20] and Bickel [2] – that sample splitting is unnecessary and undesirable as far as implementing the procedure is concerned.

Our results appear to contradict the simulation evidence of Engle and Gonzalez-Rivera [13] who found a substantial information loss when going from the MLE to the semiparametric estimator, but in fact our analysis predicts this should happen in their parameterization. In their parameterization, the ARCH/GARCH variance parameters all contain information about scale. Since knowledge of the error density conveys valuable information about overall scale, one does indeed suffer an information loss when estimating these parameters – see Steigerwald [41]. However, the relative effects that are captured by our parameterization are adaptively estimable.

APPENDIX I

Consider the Exponential ARCH model

$$\log[\sigma_t^2] = \alpha + \sum_{j=1}^{P} \gamma_j r(\varepsilon_{t-j}\sigma_{t-j}),$$

where r(.) is a known function. This specification differs from that in Nelson [30] in that the innovations are

$$\varepsilon_{t-j}\sigma_{t-j} = y_{t-j} - \beta' x_{t-j},$$

which do not depend on α or γ. The scores in this model are

$$\dot{\ell}_{T\beta} = -\sum_{t=1}^{T} \{\sigma_t(\theta)^{-1} x_t \psi_1(\varepsilon_t(\theta)) + \tilde{W}_t(\theta)\psi_2(\varepsilon_t(\theta))\} = \dot{\ell}_{T\beta 1} + \dot{\ell}_{T\beta 2},$$

where

$$\tilde{W}_t(\theta) = -\tfrac{1}{2}\sum_{j=1}^{P} \gamma_j x_{t-j} r'(y_{t-j} - \beta^T x_{t-j}),$$

and

$$\dot{\ell}_{T\gamma} = -\tfrac{1}{2}\sum_{t=1}^{T} \psi_2(\varepsilon_t(\theta)) r_t \; ; \; \dot{\ell}_{T\alpha} = -\tfrac{1}{2}\sum_{t=1}^{T} \psi_2(\varepsilon_t(\theta)),$$

where $r_t=(r_{1t}, r_{2t}, .., r_{Pt})^T$, and

$$r_{jt}(\theta) = r(y_{t-j} - \beta^T x_{t-j}), \; j=1,2,..,P.$$

Provided $E[r'(\varepsilon_t\sigma_t)]=0$, the scores for β are orthogonal to those for α and γ, and to any symmetric function of ε_t. In this case, the efficient score function for γ in the presence of unknown α is

$$\dot{\ell}^{*}_{T\gamma} = -\tfrac{1}{2}\sum_{t=1}^{T}\psi_2(\varepsilon_t(\theta))(r_t - E[r_t]),$$

which is orthogonal to any symmetric function of ε_t. In this case, α is unidentifiable when f is unknown, while both β and γ are adaptively estimable provided $\sqrt{T}$ consistent initial estimates of γ can be found.

APPENDIX II

Proof of Theorem 1.

Our treatment is very similar to Swensen [42,43] and Steigerwald [40]. The main difference arises in verifying the quadratic mean differentiability of the log likelihood ratio. We discuss this point below.

We now verify conditions 1)-6) given in proposition 1. Recall that

$$Z_{T,t} = T^{-1/2}\{ \pi_{1t}\psi_1(\varepsilon_t) + \pi_{2t}\psi_2(\varepsilon_t) \},$$

where π_{jt} depend linearly on the fixed regressors $\{x_{t-j},\ j=1,2,..,P\}$ and on the bounded random variables σ_t^{-1}, $\tilde{v}_t$, and $\{\tilde{w}_{t-j},\ j=1,2,..,P\}$ which are all measurable with respect to $\mathfrak{F}_{t-1}$.

Condition 6 is satisfied by A1. Condition 2 holds by virtue of assumptions B1, A2 and A4. For example,

$$T^{-1}\sum_{t=1}^{T} h_\beta x_t x_t^T h_\beta E[\sigma_t^{-2}\psi_1(\varepsilon_t)^2] \le E[\sigma_t^{-2}] I_1(f) T^{-1}\sum_{t=1}^{T} h_\beta x_t x_t^T h_\beta,$$

which is bounded by assumption B1.

Conditions 3), 4) and 5) can be verified exactly as in Swensen [43]. Without loss of generality, π_{jt} can be treated as deterministic constants obeying Lemma 1.2 below, since the bounded random variables on which they depend can be factored out. Then 3) follows, since in particular:

$$T^{-1/2} \underset{1 \le t \le T}{\text{Max}} \; |\pi_{jt}\psi_2(\varepsilon_t)| \xrightarrow{P} 0,$$

because

$$\Pr[T^{-1/2}\underset{1 \le t \le T}{\text{Max}} \; |\pi_{jt}\psi_2(\varepsilon_t)| > \delta] \le \delta^{-2}T^{-1}\sum_{t=1}^{T} \pi_{jt}^2 E[\psi_2(\varepsilon_t)^2 1(|\pi_{jt}\psi_2(\varepsilon_t)| > \delta\sqrt{T}),$$

by Dvoretzky's inequality, see Hall and Heyde [19], Lemma 2.5. This latter quantity tends to zero with T. Condition 4) follows by a) and b) below, where

a) $T^{-1}\sum_{t=1}^{T} \pi_{jt}^2 \psi_j(\varepsilon_t)^2 \xrightarrow{P} V_j > 0.$

This follows because $T^{-1/2}\sum_{t=1}^{T} \pi_{jt}\psi_j(\varepsilon_t)$ is asymptotically normal – see Swensen [43]. Asymptotic normality is itself a consequence of the following negligibility condition

$$[\sum_{t=1}^{T} \pi_{jt}^2]^{-1} \underset{1 \le t \le T}{\text{Max}} \; \pi_{jt}^2 \to 0,$$

which is satisfied by Lemma 1.2 below. The constants V_j are readily calculated.

b) $T^{-1}\sum_{t=1}^{T} \pi_{1t}\pi_{2t}\psi_1(\varepsilon_t)\psi_2(\varepsilon_t) \xrightarrow{P} 0.$

This follows since

$$\Pr[\ T^{-1}\sum_{t=1}^{T}\pi_{1t}\pi_{2t}\psi_1(\varepsilon_t)\psi_2(\varepsilon_t)\ >\ \delta] \le T^{-2}\sum_{t=1}^{T}\pi_{1t}^2\pi_{2t}^2 E[\psi_1(\varepsilon_t)^2\psi_2(\varepsilon_t)^2],$$

by Markov's inequality. Since

$$E[\psi_1(\varepsilon_t)^2\psi_2(\varepsilon_t)^2] \le E[\psi_1(\varepsilon_t)^4]^{1/2}E[\psi_2(\varepsilon_t)^4]^{1/2} < \infty,$$

by Cauchy-Schwarz and A4(3).

Condition 5) can be verified exactly as in Swensen [43].

We now verify condition 1). Firstly, we need some background on quadratic mean differentiability. Let ε be a random variable defined on the probability space $(\Omega,\mathfrak{F},P)$, with Lebesgue density f. Then define the stochastic process $\zeta(\varepsilon;\tau)$ on $(\Omega,\mathfrak{F},P)$, where $\tau=(\delta,\eta)$, and

$$\zeta(\varepsilon;\tau) = [f((\varepsilon+\delta)/(1+\eta)^{1/2})/(f(\varepsilon)(1+\eta)^{1/2})]^{1/2}.$$

We verify the following lemma below:

Lemma 1.1: Assume that A1 and A2 hold. Then $\zeta(\varepsilon;\tau)$ is jointly quadratic mean differentiable at any (δ,η), where $\eta>-1$. In other words, there exists a vector process $d\zeta(\varepsilon;\tau)=(d\zeta_1,d\zeta_2)^T$ such that for any $u=(m,s)^T\to 0$

$$\underset{m,s\Rightarrow 0}{Lim}\ E\{\ (m^2+s^2)^{-1/2}[\zeta(\varepsilon;\delta+m,\eta+s) - \zeta(\varepsilon;\delta,\eta) - u^T d\zeta(\varepsilon;\delta,\eta)]\}^2 \to 0,$$

independently of the path $(m,s)\to 0$.

This concept was introduced in Le Cam [25]. Joint quadratic mean differentiability implies marginal q.m.d, and we define $d\zeta_1(\varepsilon;\tau)$ and $d\zeta_2(\varepsilon;\tau)$ as satisfying

$$\lim_{m\Rightarrow 0} E\{\, m^{-1}[\zeta(\varepsilon;\delta+m,\eta)-\zeta(\varepsilon;\delta,\eta)] - d\zeta_1(\varepsilon;\delta,\eta)\}^2 \Rightarrow 0,$$
$$\lim_{s\Rightarrow 0} E\{\, s^{-1}[\zeta(\varepsilon;\delta,\eta+s)-\zeta(\varepsilon;\delta,\eta)] - d\zeta_2(\varepsilon;\delta,\eta)\}^2 \Rightarrow 0,$$

where

$$d\zeta_1(\varepsilon;\tau) = \zeta(\varepsilon;\tau)\frac{f'}{f}\left(\frac{\varepsilon+\delta}{(1+\eta)^{1/2}}\right)(1+\eta)^{-1/2} ;$$
$$d\zeta_2(\varepsilon;\tau) = -\zeta(\varepsilon;\tau)\frac{1}{2}\left\{ \left(\frac{\varepsilon+\delta}{(1+\eta)^{1/2}}\right)\frac{f'}{f}\left(\frac{\varepsilon+\delta}{(1+\eta)^{1/2}}\right) + 1\right\}(1+\eta)^{-1} ;$$

It remains to show quadratic mean differentiability for the regression model with ARCH errors. We have to show that

$$\sum_{t=1}^{T} E(X_{T,t}-Z_{T,t})^2 \Rightarrow 0.$$

We establish (E1) and (E2) below which together imply this:

$$\text{(E1)} \quad \sum_{t=1}^{T} E(X_{T,t}-Z^*_{T,t})^2 \Rightarrow 0 \quad ; \quad \text{(E2)} \quad \sum_{t=1}^{T} E(Z_{T,t}-Z^*_{T,t})^2 \Rightarrow 0$$

where

$$Z^*_{T,t} = -\tfrac{1}{2}\,\xi^T_{T,t}\psi(\varepsilon_t), \quad \xi_{T,t} = (\delta_{T,t}, \tfrac{1}{2}\eta_{T,t})^T.$$

We first show (E2).

$$E\sum_{t=1}^{T}(Z_{T,t}-Z^{*}_{T,t})^2 \le qI_2(f)\{\ E[\sigma_t^{-4}][T^{-2}\sum_{t=1}^{T}\sum_{j}\sum_{k}(c_{j0}c_{k0}(h_\beta^T x_{t-j})^2(h_\beta^T x_{t-k})^2)\ ]$$
$$+\ 4T^{-1}\sum_{t=1}^{T}\sum_{j}(c_{j0}-c_{jT})^2(x_{t-j}^T h_\beta)^2 E[\sigma_t^{-4}\varepsilon_{t-j}^2\sigma_{t-j}^2]$$
$$+\ E[\sigma_t^{-4}]\ T^{-2}\sum_{t=1}^{T}\sum_{j}\sum_{k\neq j}(c_{j0}-c_{jT})(c_{k0}-c_{kT})(h_\beta^T x_{t-j})^2(h_\beta^T x_{t-k})^2$$
$$+\ 2E[\sigma_t^{-4}]\ T^{-2}\sum_{t=1}^{T}\sum_{j}\sum_{k}(c_{j0}-c_{jT})c_{k0}(h_\beta^T x_{t-j})^2(h_\beta^T x_{t-k})^2\ \}$$
$$+\ O(T^{-1}),$$

where q is a constant reflecting the number of times we expanded out a sum of squares, while the $O(T^{-1})$ remainder term we did not specify arises from the approximations

$$e^{\alpha_T-\alpha_0} - 1 = \alpha_T - \alpha_0 + O(T^{-1})\ ;$$
$$c_{jT} - c_{j0} = e^{\alpha_0}(\gamma_{jT} - \gamma_{j0}) + \gamma_{j0}(\alpha_T - \alpha_0) + O(T^{-1}).$$

Therefore,

$$E\sum_{t=1}^{T}(Z_{T,t}-Z^{*}_{T,t})^2 \Rightarrow 0,$$

because for example

$$T^{-2}\sum_{t=1}^{T}\sum_{j}\sum_{k}(c_{j0}c_{k0}(h_\beta^T x_{t-j})^2(h_\beta^T x_{t-k})^2) \Rightarrow 0$$

by Lemma 1.2.

We now show (E1). Since ε_t is independent of both $\delta_{T,t}$ and $\eta_{T,t}$, we

have to show that $L \equiv \sum_{t=1}^{T} E[L_{T,t}] \Rightarrow 0$, where

$$L_{T,t} = \int [\zeta(\varepsilon;\delta_{T,t},\eta_{T,t})] - 1 - \tfrac{1}{2}\xi_{T,t}^{T}\psi(\varepsilon)]^2 f(\varepsilon)d\varepsilon.$$

We use the following lemma:

Lemma 1.2: There is a constant c and a deterministic sequence k(T)⇒0 such that

$$\underset{1\le t\le T}{\text{Max}} (\eta_{T,t}^2 + \delta_{T,t}^2) \le k(T) \; ; \sum_{t=1}^{T} (\eta_{T,t}^2 + \delta_{T,t}^2) < c < \infty.$$

Define the following family of neighborhoods of zero:

$$B(k) = \{ (\delta,\eta): (\eta^2 + \delta^2) < k \}.$$

Then,

$$L \le \sum_{t=1}^{T} (\delta_{T,t}^2+\eta_{T,t}^2) * \underset{B(k)}{\text{Sup}} \{(\delta^2+\eta^2)^{-1} \int [\zeta(\varepsilon;\delta,\eta) - 1 - \tfrac{1}{2}\xi^{T}\psi(\varepsilon)]^2 f(\varepsilon)d\varepsilon \}.$$

Since $k(T)\Rightarrow 0$, $L\Rightarrow 0$ as required. ■

Proof of Theorem 2

Since $\{g_{tj}(\theta_0)\}_{t=1}^{T}$ is a bounded stationary ergodic process, we have

$$T^{-1}\sum_{t=1}^{T} g_{tj}(\theta_0) \overset{p}{\Rightarrow} \bar{g}_j(\theta_0)$$

by the ergodic theorem – see for example Hall and Heyde [19] p281.It is easy to verify that in each case there is a positive constant K and a neighborhood N_{θ_0} of θ_0 on which

$$(E3) \quad T^{-1}\sum_{t=1}^{T} |g_{tj}(\theta_1) - g_{tj}(\theta_2)| \leq K \|\theta_1 - \theta_2\|,$$

for large T. For example, consider $g_{tj}(\theta)=\sigma_t^{-2}(\theta)$. In this case

$$[\sigma_t^{-2}(\theta_T) - \sigma_t^{-2}(\theta_0)] = \sigma_t^{-2}(\theta_T)\eta_{T,t},$$

for any sequence θ_T such that $\sqrt{T}(\theta_T-\theta_0)$ stays bounded. Therefore,

$$|T^{-1}\sum_{t=1}^{T} [\sigma_t^{-2}(\theta_T) - \sigma_t^{-2}(\theta_0)]| \leq \|\theta_T - \theta_0\| \; |T^{-1}\sum_{t=1}^{T} a_t|,$$

where $\{a_t\}$ is a deterministic sequence derived from $\{\eta_{T,t}\}$, and $|T^{-1}\sum_{t=1}^{T} a_t|$ is bounded.

It follows that

$$T^{-1}\sum_{t=1}^{T} (g_{tj}(\bar{\theta}_T) - g_{tj}(\theta_0)) \xrightarrow{P} 0,$$

where $\bar{\theta}_T$ is a discrete $\sqrt{T}$ consistent estimator of θ_0, and $\bar{J}_{T\theta\theta}(\bar{\theta}_T, f)$ is consistent. ■

Proof of Theorem 3: Firstly, we write

$$S_{T\theta}(\theta_T) - S_{T\theta}(\theta_0) + J_{\theta\theta}(\theta_0,f)\sqrt{T}(\theta_T-\theta_0) = Q_{T\theta}(\theta_T) - Q_{T\theta}(\theta_0) + Y_{T\theta}(\theta_T),$$

where

$$Q_{T\theta}(\theta) = T^{-1/2}\sum_{t=1}^{T}\{\Delta_{t\theta}(\theta) - E_{\theta_0}[\Delta_{t\theta}(\theta)|\mathfrak{F}_{t-1}]\},$$

is a standardized sum of martingale differences with respect to $\mathfrak{F}_{t-1}$, and

$$Y_{T\theta}(\theta_T) = T^{-1/2}\sum_{t=1}^{T}\{E_{\theta_0}[\Delta_{t\theta}(\theta_T)|\mathfrak{F}_{t-1}] + J_{\theta\theta}(\theta_0,f)\sqrt{T}(\theta_T-\theta_0),$$

where $E_{\theta_0}[\Delta_{t\theta}(\theta_0)|\mathfrak{F}_{t-1}]=0$. We show

$$\text{(E4) } Q_{T\theta}(\theta_T) - Q_{T\theta}(\theta_0) = o_p(1) \text{ ; (E5) } Y_{T\theta}(\theta_T) = o_p(1),$$

where $Q_{T\theta}=(Q_{T\beta}^T, Q_{T\alpha}, Q_{T\gamma}^T)^T$, and $Q_{T\beta}=Q_{T\beta 1}+Q_{T\beta 2}$.

Proof of E4

We establish that $Q_{T\beta}(\theta_T)-Q_{T\beta}(\theta_0)$, $Q_{T\alpha}(\theta_T)-Q_{T\alpha}(\theta_0)$, and $Q_{T\gamma}(\theta_T)-Q_{T\gamma}(\theta_0)$ are $o_p(1)$.

By the triangle inequality

$$E\{||Q_{T\beta}(\theta_T)-Q_{T\beta}(\theta_0)||^2\} \leq$$

$$2E\{||Q_{T\beta_1}(\theta_T)-Q_{T\beta_1}(\theta_0)||^2\} + 2E\{||Q_{T\beta_2}(\theta_T)-Q_{T\beta_2}(\theta_0)||^2\}.$$

Let

$$\vartheta_{T,t} = \sigma_t(\theta_T)^{-1}x_t\psi_1(\varepsilon_t(\theta_T)) - E_{\theta_0}[\sigma_t(\theta_T)^{-1}x_t\psi_1(\varepsilon_t(\theta_T))|\mathfrak{F}_{t-1}] - \sigma_t^{-1}x_t\psi_1(\varepsilon_t),$$

Then

$$E\{||Q_{T\beta_1}(\theta_T)-Q_{T\beta_1}(\theta_0)||^2 \le T^{-1}\sum_{t=1}^{T} E[\vartheta_{T,t}\vartheta_{T,t}^T],$$

since $\forall T$ $\{\vartheta_{T,t}\}$ is a martingale difference sequence with respect to $\mathfrak{F}_{t-1}$. Furthermore,

$$T^{-1}\sum_{t=1}^{T} E[||\vartheta_{T,t}\vartheta_{T,t}^T||] \le 2||A|| + 2||B||,$$

where

$$A = T^{-1}\sum_{t=1}^{T} x_t x_t^T E_{\theta_0}\{\sigma_t(\theta_T)^{-1}\psi_1(\varepsilon_t(\theta_T)) - \sigma_t^{-1}\psi_1(\varepsilon_t)\}^2,$$

and

$$B = T^{-1}\sum_{t=1}^{T} x_t x_t^T E_{\theta_0}\{\sigma_t(\theta_T)^{-1}E[\psi_1(\varepsilon_t(\theta_T))|\mathfrak{F}_{t-1}]\}^2.$$

We can bound $||A||$ by $2||A_1||+2||A_2||$, where

$$A_1 = T^{-1}\sum_{t=1}^{T} x_t x_t^T[E_{\theta_0}\{\sigma_t(\theta_T)^{-2}[\psi_1(\varepsilon_t(\theta_T)) - \psi_1(\varepsilon_t)]^2\} ;$$

$$A_2 = T^{-1}\sum_{t=1}^{T} x_t x_t^T E_{\theta_0}\{[\sigma_t(\theta_T)^{-1}-\sigma_t^{-1}]\psi_1(\varepsilon_t)\}^2].$$

Since $\sigma_t(\theta_T)^{-2}$ is eventually bounded from above by some $\delta<\infty$, we have for large T

$$(E6)\quad \sigma_t(\theta_T)^{-2}[\psi_1(\varepsilon_t(\theta_T)) - \psi_1(\varepsilon_t)]^2 \le \delta[\psi_1(\varepsilon_t(\theta_T)) - \psi_1(\varepsilon_t)]^2$$

Furthermore, $\psi_1(\varepsilon_t)$ is independent of σ_t^{-1} and $\sigma_t(\theta_T)^{-1}$, because these latter quantities depend only on the past. Therefore,

$$(E7)\quad E_{\theta_0}\{(\sigma_t(\theta_T)^{-1} - \sigma_t^{-1})\psi_1(\varepsilon_t)\}^2 = I_1(f)E_{\theta_0}\{(\sigma_t(\theta_T)^{-1} - \sigma_t^{-1})^2\}.$$

Together, (E6) and (E7) imply that we have to estimate the norms of the following matrices

$$A_1' = T^{-1}\sum_{t=1}^{T} x_t x_t^T E\{\psi_1(\varepsilon_t(\theta_T)) - \psi_1(\varepsilon_t)\}^2 \;;\; A_2' = I_1(f)T^{-1}\sum_{t=1}^{T} x_t x_t^T E\{\sigma_t(\theta_T)^{-1} - \sigma_t^{-1}\}^2.$$

Since

$$(E8)\quad \operatorname*{Max}_{1\le t\le T} E_{\theta_0}[\sigma_t(\theta_T)^{-1} - \sigma_t^{-1}]^2] \to 0,$$

by Lemma 1.2, and

$$(E9)\quad E\left[\int\{\psi_1((\varepsilon+\delta_{T,t})/(1+\eta_{T,t})^{1/2}) - \psi_1(\varepsilon)\}^2 f(\varepsilon)d\varepsilon\right] \le$$

$$\operatorname*{Sup}_{B(k)} \int\{\psi_1((\varepsilon+\delta)/(1+\eta)^{1/2}) - \psi_1(\varepsilon)\}^2 f(\varepsilon)d\varepsilon \to 0,$$

then $\|A_1'\|$, $\|A_2'\| \Rightarrow 0$.

Similarly, $B \Rightarrow 0$ because

$$T^{-1}\sum_{T=1}^{T} E_{\theta_0}\{E[\psi_1(\varepsilon_t(\theta_T))|\mathfrak{F}_{t-1}]\}^2 \leq \sup_{B(k)} \{\int\{\psi_1((\varepsilon+\delta)/(1+\eta)^{1/2})f(\varepsilon)d\varepsilon\}^2 \Rightarrow 0$$

by A3(4).

Since $\tilde{w}_s(\theta_T)$, $\tilde{v}_s(\theta_T)$, and $\sigma_s(\theta_T)^{-1}$ are eventually bounded, the same reasoning can be applied to show that

$$E\{\|Q_{T\beta 2}(\theta_T) - Q_{T\beta 2}(\theta_0)\|\}^2 \Rightarrow 0,$$
$$E\{\|Q_{T\alpha}(\theta_T) - Q_{T\alpha}(\theta_0)\|\}^2 \Rightarrow 0,$$
$$E\{\|Q_{T\gamma}(\theta_T) - Q_{T\gamma}(\theta_0)\|\}^2 \Rightarrow 0 .$$

Proof of E5

We first examine the terms due to β. We have to show that

$$T^{-1/2}\sum_{t=1}^{T} E_{\theta_0}[\Delta_{t\beta j}(\theta_T)|\mathfrak{F}_{t-1}] + J_{\beta\beta j}\sqrt{T}(\beta_T - \beta) = o_p(1), \quad j=1,2,$$

This amounts to showing that

$$T^{-1/2}\sum_{t=1}^{T} E[\Delta_{t\beta j}(\theta_T)] = J_{\beta\beta j}h_\beta + o(1), \quad j=1,2.$$

We examine Y_{T1}, the argument for Y_{T2} is the same, and is omitted. We have

$$E_{\theta_0}[\Delta_{t\beta 1}(\theta_T)|\mathfrak{F}_{t-1}] = \sigma_t(\theta_T)^{-1}x_t\int \psi_1((\varepsilon+\delta_{T,t})/(1+\eta_{T,t})^{1/2})f(\varepsilon)d\varepsilon.$$

Therefore,

$$|T^{-1/2}\sum_{t=1}^{T} E[(\sigma_t(\theta_T)^{-1}x_t\int\psi_1((\varepsilon+\delta_{T,t})/(1+\eta_{T,t})^{1/2})f(\varepsilon)d\varepsilon - T^{-1/2}\sigma_t^{-2}x_t x_t^T h_\beta I_1(f)]|$$

$$\leq |T^{-1/2}\sum_{t=1}^{T} E[(\sigma_t(\theta_T)^{-1}x_t\int\psi_1((\varepsilon+\delta_{T,t})/(1+\eta_{T,t})^{1/2})f(\varepsilon)d\varepsilon - T^{-1/2}(\sigma_t(\theta_T)^{-2}x_t x_t^T h_\beta I_1(f)]|$$
$$+ |T^{-1}\sum_{t=1}^{T}(\sigma_t^{-2}(\theta_T)-\sigma_t^{-2})x_t x_t^T h_\beta I_1(f)]|.$$

Using the argument given in Theorem 2,

$$||T^{-1}\sum_{t=1}^{T}(\sigma_t(\theta_T)^{-2}-\sigma_t^{-2})x_t x_t^T|| \xrightarrow{P} 0.$$

By the eventual boundedness of $\sigma_t(\theta_T)^{-1}$ the first term is less than

$$|T^{-1/2}\sum_{t=1}^{T} x_t E[\int\psi_1((\varepsilon+\delta_{T,t})/(1+\eta_{T,t})^{1/2})f(\varepsilon)d\varepsilon - T^{-1/2}\sigma_t^{-1}x_t x_t^T h_\beta I_1(f)]|,$$

$$\leq |T^{-1/2}\sum_{t=1}^{T}[\delta_{T,t}x_t \operatorname*{Sup}_{B(k)} \delta^{-1}[\int\psi_1((\varepsilon+\delta)/(1+\eta)^{1/2})f(\varepsilon)d\varepsilon - T^{-1/2}\sigma_t^{-1}x_t x_t^T h_\beta I_1(f)]|,$$
$$\leq |T^{-1}\sum_{t=1}^{T}[x_t x_t^T h_\beta \sigma_t^{-1} \operatorname*{Sup}_{B(k)} |\delta^{-1}[\int\psi_1((\varepsilon+\delta)/(1+\eta)^{1/2})f(\varepsilon)d\varepsilon + I_1(f)]| \Rightarrow 0$$

by A3(2). ■

Proof of Lemmas

1.1) Swensen [42] Lemmas 3 and 4 p.56-57 establish that assumptions A1a) and A2 are sufficient to guarantee that the related process

$$\tilde{\zeta}(\varepsilon;\tau) = [f((\varepsilon+\mu)/\sigma)/\sigma f(\varepsilon)]^{1/2}$$

is jointly differentiable in quadratic mean $\forall\mu,\sigma$ for $\sigma>0$. We change variables from σ to $(1+\eta)^{1/2}$. Since $(1+x)^{1/2}$ is continuous in x the result follows. ■

1.2) Assumption B1 implies that:

$$1)\ \max_{1\le t\le T} x_t^T\Big(\sum_{t=1}^{T} x_t x_t^T\Big)^{-1}x_t \to 0 \quad ;\ 2)\ T^{-1}\max_{1\le t\le T} x_t^T x_t \to 0,$$

see Wu [46], Lemma 3. Expand out $\delta_{T,t}^2+\eta_{T,t}^2$ using the triangle inequality. Since σ_t^{-1}, $\sigma_t^{-2}\varepsilon_{t-j}\sigma_{t-j}$ and $\sigma_t^{-2}\varepsilon_{t-j}^2\sigma_{t-j}^2$ are bounded, they can be factored out of the expression. Then for example,

$$T^{-2}\max_{1\le t\le T}(h_\beta^T x_{t-j})^4 \le [T^{-1}\max_{1\le t\le T}(h_\beta^T x_{t-j})^2]^2 \to 0,$$

$$T^{-2}\sum_{t=1}^{T}(h_\beta^T x_{t-j})^4 \le T^{-1}\max_{1\le t\le T}(h_\beta^T x_{t-j})^2\ T^{-1}\sum_{t=1}^{T}(h_\beta^T x_{t-j})^2 \to 0.$$

The result follows. ■

Proof of Theorem 5:

All calculations below are carried out under the measure P_{T,θ_T}, by contiguity the convergence to zero in probability also holds under P_{T,θ_0}. Now consider

$$E\|\hat{S}_{T\beta}(\theta_T)-S_{T\beta}(\theta_T)\|^2 \leq$$

$$2E\|\hat{S}_{T\beta_1}(\theta_T)-S_{T\beta_1}(\theta_T)\|^2 + 2E\|\hat{S}_{T\beta_2}(\theta_T)-S_{T\beta_2}(\theta_T)\|^2.$$

By construction $\hat{\psi}_{T,t}(x;\theta_T)$ is antisymmetric about zero, i.e.

$$\hat{\psi}_{T,t}(-x;\theta_T) = -\hat{\psi}_{T,t}(x;\theta_T),$$

for all x. Therefore, as in Kreiss [21] and Bickel [2] we obtain

$$E\|\hat{S}_{T\beta_1}(\theta_T)-S_{T\beta_1}(\theta_T)\|^2 = T^{-1}\sum_{t=1}^{T} x_t x_t^T E[\sigma_t(\theta_T)^{-2}\int\{\hat{\psi}_{T,t}(x;\theta_T)-\psi_1(x)\}^2 f(x)dx],$$

and because $\sigma_t(\theta_T)^{-2}$ is bounded for large T, we can apply directly the results of Kreiss [21]. We have

$$\underset{1\leq t\leq T}{\text{Max}}\ E\int\{\hat{\psi}_{T,t}(x;\theta_T) - \psi_1(x)\}^2 f(x)dx \to 0.$$

Therefore,

$$E\|\hat{S}_{T\beta 1}(\theta_T) - S_{T\beta 1}(\theta_T)\|^2 \Rightarrow 0.$$

To establish the same for $E\|\hat{S}_{T\beta 2}(\theta_T)-S_{T\beta 2}(\theta_T)\|^2$ we must exploit the functional form of $\tilde{W}_t$. The estimated scale score – $\hat{\psi}_{T,t}(x;\theta_T)x+1$ – is symmetric in both x and $\varepsilon_{t-j}(\theta_T)$ for any j, while $\tilde{W}_t$ is antisymmetric in $\varepsilon_{t-1}(\theta_T)$, i.e.

$$\tilde{W}_t(-\varepsilon_{t-1}(\theta_T)) = -\tilde{W}_t(\varepsilon_{t-1}(\theta_T)).$$

Therefore, the cross products drop out

$$E[\tilde{W}_t\tilde{W}_s^T\{\varepsilon_t\varepsilon_s\{\hat{\psi}_{T,t}(\varepsilon_t;\theta_T) - \psi_1(\varepsilon_t)\}\{\hat{\psi}_{T,s}(\varepsilon_s;\theta_T) - \psi_1(\varepsilon_s)\}] = 0,$$

and

$$E\|\hat{S}_{T\beta 2}(\theta_T)-S_{T\beta 2}(\theta_T)\|^2 = T^{-1}\sum_{t=1}^{T} E[\tilde{W}_t(\theta_T)\tilde{W}_t(\theta_T)^T]\int\{\hat{\psi}_{T,t}(x;\theta_T)-\psi_1(x)\}^2x^2f(x)dx$$

By a minor modification of Kreiss' [21] arguments we can establish that

$$\text{(E10)}\quad \max_{1\le t\le T} E_{\theta_T}[\int\{\hat{\psi}_{T,t}(x;\theta_T) - \psi_1(x)\}^2x^2f(x)dx] \Rightarrow 0,$$

via a sequence of standard arguments collected below in Lemmas 5.1-5.5. Therefore, since

$$T^{-1}\sum_{t=1}^{T} E[\tilde{W}_t(\theta_T)\tilde{W}_t(\theta_T)^T] < M,$$

for some $M<\infty$, the result follows.

Notice that Kreiss's argument does not require any sample splitting. However, when we examine the estimated scores for γ, we are unable to exploit symmetry properties and the argument becomes considerably more involved. We adopt a form of sample splitting in order to provide a simple proof. We split the sample into two sub–samples

$$I_1 = \{ t: t=1,2,..,T_1\} \; ; \; I_2 = \{ t : t=T_1+1,..,T\},$$

where

$$T_1(T) \Rightarrow \infty \; ; \; T_1/T \Rightarrow 0 \text{ as } T \Rightarrow \infty.$$

The first sub–sample is used to estimate the score function $\psi_1(x)$, while the remaining observations are used to construct the estimator.

In this case

$$\hat{S}_{T\gamma}(\theta_T) - S_{T\gamma}(\theta_T) = T^{-1/2}\sum_{t\in I_2} \varepsilon_t(\theta_T)\,(\hat{\psi}_{T,t}(\varepsilon_t(\theta_T)) - \psi_1(\varepsilon_t(\theta_T)))\,(\tilde{v}_t(\theta_T) - \hat{\bar{v}}),$$

where $\hat{\bar{v}}(\theta_T)=T^{-1}\sum_{t\in I_2} \tilde{v}_t(\theta_T)$. For economy of notation, we drop the θ_T argument. We have

$$\hat{S}_{T\gamma} - S_{T\gamma} = T^{-1/2}\sum_{t\in I_2}\{\varepsilon_t(\hat{\psi}_{T,t}(\varepsilon_t) - \psi_1(\varepsilon_t))(\tilde{v}_t - \bar{v})\}$$

$$+ T^{1/2}(\bar{v} - \hat{\bar{v}})T^{-1}\sum_{t\in I_2}\varepsilon_t(\hat{\psi}_{T,t}(\varepsilon_t) - \psi_1(\varepsilon_t)),$$

$$= I + II.$$

But $\tilde{v}_t - \bar{v}$ is zero mean and independent of $\varepsilon_t(\hat{\psi}_{T,t}(\varepsilon_t) - \psi_1(\varepsilon_t))$, and hence

$$E[||T^{-1/2}\sum_{t\in I_2}\varepsilon_t(\hat{\psi}_{T,t}(\varepsilon_t) - \psi_1(\varepsilon_t))(\tilde{v}_t - \bar{v})||^2] =$$

$$T^{-1}\sum_{t\in I_2}E\{\varepsilon_t(\hat{\psi}_{T,t}(\varepsilon_t) - \psi_1(\varepsilon_t))(\tilde{v}_t - \bar{v})\}^2.$$

Since $\tilde{v}_t$ is bounded, if (5.1) holds then $I=o_p(1)$ as required. The second term II is $o_p(1)$ because

E11) $T^{1/2}(\bar{v} - \hat{\bar{v}}) = O_p(1)$,

E12) $T^{-1}\sum_{t\in I_2}\varepsilon_t(\hat{\psi}_{T,t}(\varepsilon_t) - \psi_1(\varepsilon_t)) = o_p(1)$,

by (E10).

We now establish the fundamental property (E10). Let

$$f_b(x) = \int \phi(x - y;b)f(y)dy \; ; \; \psi_b(x) = \frac{f_b'}{f_b}(x).$$

By repeated addition and subtraction, we get that

$$\int x^2\{\hat{\psi}_{T,t}(x;\theta_T) - \psi_1(x)\}^2 f(x)dx$$

$$\leq 3\ \{\ \int x^2\{\hat{\psi}_{T,t}(x;\theta_T) - \hat{\psi}_{T,t}(x;\theta_T)[\frac{f_b}{f}(x)]^{1/2}\}^2 f(x)dx$$

$$+ \int x^2\{\hat{\psi}_{T,t}(x;\theta_T)[\frac{f_b}{f}(x)]^{1/2} - \psi_b(x)[\frac{f_b}{f}(x)]^{1/2}\}^2 f(x)dx$$

$$+\int x^2\{\ [\frac{f_b}{f}(x)]^{1/2}[\frac{f_b'}{f_b}(x)] - [\frac{f'}{f}(x)]\}^2 f(x)dx\ \}.$$

$$= 3\int x^2\hat{\psi}_{T,t}(x;\theta_T)^2[f_b(x)^{1/2} - f(x)^{1/2}]^2dx$$

$$+ 3\int x^2[\hat{\psi}_{T,t}(x;\theta_T) - \psi_b(x)]^2 f_b(x)dx$$

$$+ 3\int x^2[\ \frac{f_b'}{f_b^{1/2}}(x) - \frac{f'}{f^{1/2}}(x)\ ]^2dx.$$

The second term is essentially a 'variance' term, while the remaining terms are 'biases'.

The following lemmas are, apart from the factor x^2, identical to Lemmas 6.5 and 6.6 proved in Kreiss [21]

Lemma 5.1 For each $x\in\mathbb{R}$, *there are constants* κ_0 *and* κ_1 *such that*

$$E_{\theta_T}[x^2 f_b^{-1}(x)\{\hat{f}_{b,t}(x;\theta_T) - f_b(x)\}^2] \leq \frac{1}{b(T)T}\{\kappa_0 x^2 + \kappa_1 x^4 T^{-1})\}.$$

Lemma 5.2 For each $x\in\mathbb{R}$, there are constants K_0 and K_1 such that

$$E_{\theta_T}[f_b^{-1}(x)x^2\{\hat{f}'_{b,t}(x;\theta_T) - f'_b(x)\}^2] \le \frac{1}{b(T)^3T}\{K_0x^2 + K_1x^4T^{-1}\}.$$

Lemma 5.3: As $T\Rightarrow\infty$

$$\int x^2[\frac{f'_b}{f_b^{1/2}}(x) - \frac{f'}{f^{1/2}}(x)]^2dx = o(1).$$

This holds because

$$\int \psi_b(x)^2f_b(x)x^2dx < \int \psi_1(x)^2f(x)x^2dx < \infty,$$

since $I_2<\infty$. Therefore, we can apply dominated convergence. ■

Lemma 5.4: Provided $Tb(T)^3c(T)^{-2}e(T)^{-2}\Rightarrow 0$,

$$\underset{1\le t\le T}{\text{Max}}\ E_{\theta_T}\int x^2[\hat{\psi}_{T,t}(x;\theta_T) - \psi_b(x)]^2f_b(x)dx = o(1).$$

This is the same as Lemma 6.8 in Kreiss [21] apart from the additional factor of x^2 and some constants. We have

$$[\hat{\psi}_{T,t}(x;\theta_T) - \psi_b(x)]^2f_b(x) \le$$

$$2[\frac{\hat{f}'_{b,t}}{\hat{f}_{b,t}}(x) - \frac{\hat{f}'_{b,t}}{f_b}(x)]^2f_b(x) + 2[\frac{\hat{f}'_{b,t}}{f_b}(x) - \frac{f'_b}{f_b}(x)]^2f_b(x).$$

Since

$$[\frac{\hat{f}'_{b,t}}{\hat{f}_{b,t}}(x) - \frac{\hat{f}'_{b,t}}{f_b}(x)]^2 f_b(x) = [\frac{\hat{f}'_{b,t}}{\hat{f}_{b,t}}(x)]^2 [\hat{f}_{b,t}(x) - f_b(x)]^2 f_b(x)^{-1}$$

$$\leq c(T)^2 [\hat{f}_{b,t}(x) - f_b(x)]^2 f_b(x)^{-1},$$

we can apply Lemmas 5.1 and 5.2 above, after truncating the integral according to the sets

$$A_{T,t}=\{x;\hat{f}_{b,t}(x;\theta_T)\geq d_T\} \ ; \ B_{T,t}=\{x;|x|<e_T\} \ ; \ C_{T,t}=\{x;\hat{\psi}_{T,t}(x;\theta_T)\leq c_T\}.$$

The proof is exactly the same as Kreiss [21].■

Lemma 5.5: Provided $b(T)c(T)\Rightarrow 0$,

$$\underset{1\leq t\leq T}{\text{Max}} E_{\theta_T}\{\int\{x^2\hat{\psi}^2_{T,t}(x;\theta_T)[\sqrt{f_b}(x) - \sqrt{f}(x)]^2 dx\} = o(1)$$

Since $\hat{\psi}^2_{T,t}\leq c(T)^2$, this random variable is bounded by

$$c(T)^2 \int\{x^2[\sqrt{f_b}(x) - \sqrt{f}(x)]^2 dx\} = O(b(T)^2 c(T)^2),$$

by Bickel [2] Lemma 6.3. It is $o(1)$ provided $b(T)c(T)\Rightarrow 0$. ■

REFERENCES

1. Abramson, I.S. On Bandwidth Variation in Kernel Estimates – A Square Root Law. *The Annals of Statistics* 10 (1980): 1217-1223.

2. Bickel, P.J. On Adaptive Estimation. *Annals of Statistics* 10 (1982): 647-671.

3. Bickel, P.J., and Y. Ritov. Achieving Information Bounds in non and Semiparametric Models. *Annals of Statistics* 18 (1990): 925-938.

4. Bickel, P.J., C.A.J. Klaassen, Y. Ritov, and J.A. Wellner. *Efficient and Adaptive Inference in Semiparametric Models* (1991): Forthcoming Monograph: Johns Hopkins University Press.

5. Bollerslev, T. Generalized Autoregressive Conditional Heteroscedasticity. *Journal of Econometrics* 31 (1986): 307-327.

6. Bollerslev, T., and J.M. Wooldridge. Quasi-Maximum Likelihood Estimation of Dynamic Models with Time Varying Covariances. (1991) Forthcoming in *Econometric Reviews.*

7. Bollerslev, T., R.Chou, and K.Kroner. ARCH Modelling in Finance: A Review of the Theory and Empirical Evidence. *Journal of Econometrics* 52 (1992): 5-59.

8. Bougerol, P. and N. Picard. Stationarity of GARCH Processes and some Non-Negative Time Series. *Journal of Econometrics* 52 (1992)

9 Cramer, H. *Mathematical Methods in Statistics* Princeton University Press 1946.

10. Engle, R.F. Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of UK Inflation. *Econometrica* 50 (1982): 987-1008.

11. Engle, R.F. Estimates of the Variance of U.S. Inflation Based upon the ARCH Model. *Journal of Money, Credit, and Banking* 15 (1983): 286-301.

12. Engle, R.F., and T. Bollerslev. Modelling the Persistence of Conditional Variances. *Econometric Reviews* 5 (1986): 1-50.

13. Engle, R.F., and G. Gonzalez-Rivera. Semiparametric ARCH Models. *Journal of Business and Economic Statistics* 9 (1991): 345-360.

14. Fabian, V., and J. Hannan. On Estimation and Adaptive Estimation for Locally Asymptotically Normal Families. *Zeitschrift für Wahrsheinlichskeitstheorie und verwandte Gebiete.* 59 (1982): 459-478.

15. Hajek, J. Asymptotically Most Powerful Rank Order Tests. *Annals of Mathematical Statistics* 33 (1962): 1124-1147.

16. Hajek, J. Local Asymptotic Minimax and Admissibility in Estimation. *Proceedings of the Sixth Berkeley Symposium in Mathematical Statistics* 1972

17. Gallant, A.R., D. Hsieh and G.Tauchen. On Fitting a Recalcitrant Series: The Dollar/Pound Exchange Rate, 1974-1983 in *Nonparametric and Semiparametric Methods in Econometrics and Statistics* eds W.Barnett, J.Powell, and G.Tauchen. Cambridge University Press 1991.

18. Geweke, J. Comment on: Modelling the Persistence of Conditional Variances. *Econometric Reviews* v.5 (1986): 57-61.

19. Hall, P., and C.C. Heyde. *Martingale Limit Theory* Academic Press New York 1980.

20. Hsieh, D.A., and C.F. Manski. Monte Carlo Evidence on Adaptive Maximum Likelihood Estimation of a Regression. *The Annals of Statistics* 15 (1987): 541-551.

21. Kreiss, J.P. On Adaptive Estimation in Stationary ARMA Processes. *Annals of Statistics* 15 (1987): 112-133.

22. Kreiss, J.P. On Adaptive Estimation in Autoregressive Models when there are Nuisance Functions. *Statistics and Decisions* 5 (1987): p59-76.

23. Le Cam, L. Local Asymptotically Normal Families of Distributions. *University of California Publications in Statistics* 3 (1960): 267-284.

24. Le Cam, L. Likelihood Functions for Large Numbers of Independent Observations. in *Festschrift for J.Neyman* ed. F.N. David, Wiley 1966.

25. Le Cam, L. On the Assumptions used to Prove Asymptotic Normality of Maximum Likelihood Estimates. *The Annals of Mathematical Statistics* 41 (1970): 802-828.

26. Lind, B., and G. Roussas. Cramer-Type Conditions and Quadratic Mean Differentiability. *Annals of the Institute of Statistical Mathematics* 29 (1977): 189-201.

27. Lumsdaine, R.L. Asymptotic Properties of the Quasi-Maximum Likelihood Estimator in GARCH(1,1) and IGARCH(1,1) Models. Manuscript, Harvard University (1990).

28. Mandelbrot, B. The Variation of Certain Speculative Prices. *Journal of Business* 36 (1963): 394-419.

29. Milhoj, A. The Moment Structure of ARCH Processes. *Scandinavian Journal of Statistics* 12 (1985):281-292.

30. Nelson, D.B. Conditional Heteroscedasticity in Asset Returns: A New Approach. *Econometrica* 59 (1991): 347-370.

31. Nelson, D.B. Stationarity and Persistence in the GARCH(1,1) Model. *Econometric Theory* 6 (1990): 318-334.

32. Nelson, D.B., and C.Q. Cao. A note on the Inequality Constraints in the Univariate GARCH Model. Forthcoming in *Journal of Business and Economic Statistics* (1991).

33. Nemec, A.F. and Linnell. Conditionally Heteroscedastic Autoregression. University of Washington, Department of Statistics, Technical Report #43 (1984).

34. Phillips, P.C.B. Optimal Inference in Cointegrated Systems. *Econometrica* 59 (1991): 283-306.

35. Roussas, G.G. *Contiguity of Probability Measures: some applications in statistics.* Cambridge University Press 1972.

36. Roussas, G.G. Asymptotic Distribution of the Log-Likelihood Function for Stochastic Processes. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* 43 (1979): 31-46.

37. Sampson, M. A Stationarity Condition for the GARCH(1,1) Process. Department of Economics, Concordia University (1988).

38. Shephard, N. A Local Scale Model: an Unobserved Component Alternative to Integrated GARCH Processes. LSE Discussion Paper EM/90/220 (1990).

39. Spanos, A. A Parametric Approach to Dynamic Heteroscedasticity: The Students' t and Related Linear Models. Unpublished Manuscript (1990).

40. Steigerwald, D. Adaptive Estimation in Time Series Regression Models. Working Paper, Department of Economics, UC Santa Barbara (1990).

41. Steigerwald, D. Efficient Estimation of Models with Conditional Heteroscedasticity. Paper presented at ESWM Louisana 1991.

42. Swensen, A.R. Asymptotic Inference for a Class of Stochastic Processes. Ph.D Thesis, Department of Statistics, UC Berkeley (1980).

43. Swensen, A.R. The Asymptotic Distribution of the Likelihood Ratio for Autoregressive Time Series with a Regression Trend. *Journal of Multivariate Analysis* 16 (1985): 54-70.

44. Weiss, A. Asymptotic Theory for ARCH Mcdels: Estimation and Testing. *Econometric Theory* 2 (1986): 107-131.

45. Whistler, D. Semiparametric ARCH Estimation of Intra-Daily Exchange Rate Volatility. Manuscript London School of Economics (1988).

46. Wu, C. Asymptotic Theory of Nonlinear Least Squares Estimation. *The Annals of Statistics* 9 (1981): 501-513.

ENDNOTES

1. I would like to thank Paul Ruud, Tom Rothenberg, Doug Steigerwald, and Peter Bickel for helpful discussions.

2. Another example of this phenomenon is the linear regression model with intercept when the errors are allowed to be asymmetric about zero. At first blush, the scores for the slope parameters are not orthogonal to the tangent space for f. Bickel [2] shows that one must first project the slope scores orthogonally to the scores for the intercept. The identifiable parameters – the slopes – are adaptively estimable in this case.

3. Lumsdaine [27] also establishes the positive definiteness of $J_{\theta\theta}(\theta_0)$. She dispenses with assumption A5(3) at the cost of strengthening A4(2).