Anders Rahbek
Rasmus Søndergaard Pedersen
University of Copenhagen

September 2020

Part III
Estimation of ARCH models

In this part of the notes estimation and asymptotic inference is discussed. Asymptotic distributions of maximum likelihood estimators and likelihood ratio test statistics are derived using classical arguments from asymptotic theory. These are presented such that they can also be used for other types of models including general(ized) ARCH models, see for example the analysis of more general ARCH models in Kristensen and Rahbek (2005, 2009).

In addition, we discuss some non-standard inference issues caused by non-negativity constraints on the parameters of the ARCH model. A key example being that even the simple hypothesis of no ARCH is non-standard as will be demonstrated.

III.1 Estimation of ARCH(1)

Recall that the ARCH(1) model is given by,

$$x_t = \sigma_t(\theta) z_t \tag{III.1}$$
$$\sigma_t^2(\theta) = \sigma^2 + \alpha x_{t-1}^2 \tag{III.2}$$

with z_t $i.i.d.N(0{,}1)$ and x_0 fixed. The parameters of the model are given by

$$\theta = \left(\sigma^2, \alpha\right)',$$

where $\sigma^2 > 0$ and $\alpha \geq 0$. Alternatively,

$$\theta \in \Theta = (0, \infty) \times [0, \infty) = \mathbb{R}_+ \times \mathbb{R}^+.$$

The notation $\sigma_t^2(\theta)$ emphasizes that the conditional variance depends on the parameters in θ. We shall use the notation,

$$\sigma_t^2 = \sigma_t^2(\theta_0) = \sigma_0^2 + \alpha_0 x_{t-1}^2,$$

such that σ_t^2 denotes $\sigma_t^2(\cdot)$ evaluated at the so-called "true parameter value" θ_0. That is, when making probability statements for a particular choice of a parameter value this is emphasized by the subscript "0".

This is useful when discussing estimation where we estimate θ, with $\hat{\theta}$ denoting the estimator. The estimator $\hat{\theta}$ is a (often implicit and complicated) function of the data $(x_t)_{t=1,...,T}$ and when we discuss the properties of $\hat{\theta}$, such as consistency, $\hat{\theta} \xrightarrow{P} \theta_0$, and asymptotic normality of $\sqrt{T}(\hat{\theta} - \theta_0)$, we use θ_0 to denote the parameter value for which the process x_t is generated. For example, we know that if $\alpha_0 < 1$ then $Ex_t^2 < \infty$, while we need $\alpha_0 < 1/\sqrt{3}$ for $Ex_t^4 < \infty$. And, typically, when establishing consistency we need for example $Ex_t^2 < \infty$, while higher order moments such as Ex_t^4 are needed for asymptotic normality of $\hat{\theta}$. Thus different values of θ_0 yields different properties of x_t.

Example III.1.1 *Consider the AR(1) model given by*

$$x_t = \rho x_{t-1} + \varepsilon_t$$

for $t = 1, 2, ..., T$ *and with* $\rho \in \mathbb{R}$, ε_t *i.i.d.*$N(0, \sigma^2)$ *and* x_0 *fixed. The ordinary least squares (OLS) estimator is here identical to the maximum likelihood estimator (MLE) maximizing with* ρ *freely varying in* $\mathbb{R}$,

$$\hat{\rho} = \sum_{t=1}^{T} x_t x_{t-1} / \sum_{t=1}^{T} x_{t-1}^2.$$

That is, $\hat{\rho} = \arg\max_{\rho \in \mathbb{R}} (\ell_T(\theta))$, *where*

$$\ell_T(\theta) = -\tfrac{1}{2} \sum_{t=1}^{T} (\log \sigma^2 + \tfrac{(x_t - \rho x_{t-1})^2}{\sigma^2})$$

We know that if ρ_0 *satisfies* $|\rho_0| < 1$, *and* $\sigma_0^2 > 0$, *then* $y_t = \rho_0 y_{t-1} + \varepsilon_t$ *is weakly mixing, and using the LLN in Lemma I.3.2 we immediately get,*

$$\hat{\rho} \xrightarrow{P} E(x_t x_{t-1}) / E(x_{t-1}^2) = \rho_0 \left(\sigma_0^2 / \left(1 - \rho_0^2\right)\right) / \left(\sigma_0^2 / \left(1 - \rho_0^2\right)\right) = \rho_0.$$

Likewise, using the CLT in Theorem II.4.1,

$$\sqrt{T}(\hat{\rho} - \rho_0) = \frac{\frac{1}{\sqrt{T}} \sum_{t=1}^{T} x_{t-1}\varepsilon_t}{\frac{1}{T} \sum_{t=1}^{T} x_{t-1}^2} \xrightarrow{D} \frac{N\left(0, \sigma_0^4 / \left(1 - \rho_0^2\right)\right)}{\sigma_0^2 / \left(1 - \rho_0^2\right)} = N\left(0, 1 - \rho_0^2\right). \qquad \text{(III.3)}$$

To see this, use that $x_{t-1}\varepsilon_t = x_{t-1}(x_t - \rho_0 x_{t-1})$, *such that* $f(x_t, x_{t-1}) = x_{t-1}(x_t - \rho_0 x_{t-1})$ *in Theorem II.4.1, with* $E(f(x_t, x_{t-1}) | x_{t-1}) = \rho_0 x_{t-1}^2 - \rho_0 x_{t-1}^2 = 0$. *Using conditional expectations, see Lemma I.2.1,*

$$\begin{aligned} Ef^2(x_t, x_{t-1}) &= E\left(f^2(x_t, x_{t-1})\right) = E\left(x_{t-1}^2 \varepsilon_t^2\right) \\ &= E\left(E\left(x_{t-1}^2 \varepsilon_t^2 | x_{t-1}\right)\right) \\ &= E\left(x_{t-1}^2 \sigma_0^2\right) = \sigma_0^4 / \left(1 - \rho_0^2\right). \end{aligned}$$

Note that these results do not apply if $\rho_0 = 1$ for example, where instead Dickey-Fuller type distributions would appear. Thus it is important to emphasize for which value(s) of the parameters the results apply.

Under the assumption of Gaussianity of the innovations z_t, the (Gaussian) log-likelihood function for the ARCH(1) model is given by,

$$\ell_T(\theta) = -\tfrac{1}{2}\sum_{t=1}^{T}\left(\log\sigma_t^2(\theta) + \frac{x_t^2}{\sigma_t^2(\theta)}\right) \tag{III.4}$$

as x_t is conditionally $N(0, \sigma_t^2)$ distributed, see also Example I.3.8. Note that here and most often the constant term $-\frac{T}{2}\log(2\pi)$ is left out as it does not depend on θ and hence plays no role in the discussion of maximization.

The (log-)likelihood function $\ell_T(\theta)$ is maximized over $\theta = (\sigma^2, \alpha) \in \Theta$ with $\sigma^2 > 0$ and $\alpha \geq 0$ from which we get the maximum likelihood (ML) estimator $\hat{\theta}$,

$$\hat{\theta} = \arg\max_{\theta\in\Theta} \ell_T(\theta).$$

Unlike for the AR(1) model, no closed form solution exists for the ML estimator in the ARCH(1) (and general ARCH) model(s) and instead numerical optimization has to be used which is briefly discussed in the next section.

Often the likelihood function in (III.4) is used even if the *i.i.d.* sequence z_t of the ARCH process is not assumed to be Gaussian. In this case, estimation is referred to as *quasi* ML estimation, or QMLE. This is similar to the classic OLS estimator which is the MLE in the case of *i.i.d.* Gaussian observations. Typically theory for the OLS estimator is stated under various and more general assumptions on the "innovations" departing from Gaussianity; in this sense one can consider the OLS estimator as an example of a QML estimator.

In line with existing literature on ARCH and GARCH models, estimation is here (predominantly) discussed under the assumption of a Gaussian likelihood function as in (III.4), while regularity conditions are stated such that some flexibility is allowed for the distribution of the innovations z_t when stating results about consistency and asymptotic normality. That is, we consider QMLE and QLR rather than ML estimators and likelihood ratio (LR) statistics.

At the same time, note that if z_t as in Example I.3.9 is instead assumed to be t_v-distributed (and scaled by $\sqrt{\frac{v-2}{v}}$ such that z_t $i.i.d.t_v(0,1)$), then a t_v–log-likelihood function would be given by $\ell_T^{t_v}(\theta) = \sum_{t=1}^{T}\log f(x_t|x_{t-1})$ with $f(x_t|x_{t-1})$ given in Example I.3.9. That is, in this case, again apart from constants,

$$\ell_T^{t_v}(\theta) = -\tfrac{1}{2}\sum_{t=1}^{T}\left(\log\sigma_t^2(\theta) + (v+1)\log\left(1+\frac{x_t^2}{\sigma_t^2(\theta)(v-2)}\right)\right),$$

with $\sigma_t^2(\theta) = \sigma^2 + \alpha x_{t-1}^2$ as before. Often the degrees of freedom v is considered a parameter to be estimated as well such that $\theta = (\sigma^2, \alpha, v)$. In this case $\ell_T^{t_v}(\theta)$ also has additional two terms, $\log\gamma(v) - \frac{1}{2}\log(v-2)$ as these are no longer constant over θ. Thus if the z_t are believed to be t_v distributed then one may use QML estimation (applying a Gaussian likelihood), or ML estimation (applying the t_v likelihood). Both cases are applied in existing literature.

Finally, as an alternative to (Q)MLE one may also consider moment estimation (M-estimation) based here on OLS from the regression, or estimating, equation,

$$x_t^2 = \sigma^2 + \alpha x_{t-1}^2 + e_t, \tag{III.5}$$

where e_t is the "regression error", $e_t = x_t^2 - E(x_t^2|x_{t-1})$. The ARCH OLS estimator is given by,

$$\hat{\theta}'_{\text{ols}} = \sum x_t^2 w_t' \left(\sum w_t w_t'\right)^{-1}, \quad w_t = \left(1, x_{t-1}^2\right)'$$

and is easy to compute. While it can be shown to be consistent, it is not efficient as it has a larger variance than the MLE, and it is not a good candidate for a final estimator of θ_0. Moreover, the regularity conditions for $\hat{\theta}_{\text{ols}}$ to be consistent and asymptotically Gaussian distributed are much stronger than for the QMLE. Similar to autocorrelation functions of x_t^2, the regularity conditions for $\hat{\theta}_{\text{ols}}$ to be consistent and asymptotically Gaussian are so strong that they are unlikely not to hold in practice. The estimator is sometimes used as an initial starting value for some iterative search algorithm leading to the $\hat{\theta}_{\text{(q)mle}} = \hat{\theta}$ such as the Newton-Raphson discussed in the next.

III.2 Numerical Optimization

In general, with $\theta = (\theta_1, \theta_2, ..., \theta_d)'$ a d–dimensional parameter, (Q)ML estimation is performed by optimization of a log-likelihood function as a function of θ. That is,

$$\hat{\theta} = \arg\max \ell_T(\theta),$$

where optimization is over θ, with $\theta \in \Theta$. For example, in the ARCH(1) model, $d = 2$ with $\theta = (\sigma^2, \alpha)'$ and $\Theta = (0,\infty) \times [0,\infty) = \mathbb{R}_+ \times \mathbb{R}^+$.

There are many algorithms designed for this, such as the classical Newton-Rahpson algorithm. All algorithms share the property that sometimes they converge to a unique maximum $\hat{\theta}$ and sometimes not. This depends on the shape of the log-likelihood function, $\ell_T(\theta)$, and on the type of algorithm. Hence often different algorithms are applied to see which works best in concrete models.

A simple way to try to find $\hat{\theta}$ would be a *grid search* where different values of the parameter $\theta^i = (\theta_1^i, ..., \theta_d^i)'$ are inserted, and $\ell_T(\theta_i)$ compared for these. How to choose the correct grid $(\theta^i)_{i=1,...,M}$ say, is then the problematic part or the essence of various pre-programmed search algorithms. Classic examples include *grid search* with equi-spaced points with M "large". In *alternating grid search,* first $(\theta_2, ..., \theta_d)$ are fixed and only a grid for θ_1, θ_1^i, is searched over, giving $\hat{\theta}_1 = \hat{\theta}_1(\theta_2, ..., \theta_d)$. Next, fixing $(\hat{\theta}_1, \theta_3, \theta_4, ..., \theta_d)$, a grid search is performed over θ_2 and so forth. With *random grid search* instead the grid $\theta^i = (\theta_1^i, ..., \theta_d^i)'$ is chosen in some random way, for example using the uniform distribution.

III.2.1 Newton-Raphson optimization

The Newton-Raphson procedure may be applid in cases where $\ell_T(\theta)$ is two times (continuously) differentiable with respect to θ as the algorithm is based on a second order Taylor expansion of the log-likelihood function.

The first derivative of $\ell_T(\theta)$ with respect to θ – the *score* $s_T(\theta)$ – is given by the $d-$dimensional vector,

$$s_T(\theta) = \partial \ell_T(\theta) / \partial \theta = \begin{pmatrix} \partial \ell_T(\theta) / \partial \theta_1 \\ \vdots \\ \partial \ell_T(\theta) / \partial \theta_d \end{pmatrix}.$$

Likewise, *minus* the second derivative, or the (*observed) information* $i_T(\theta)$, is given by the $(d \times d)$-dimensional symmetric matrix,

$$i_T(\theta) = -\partial^2 \ell_T(\theta) / \partial \theta \partial \theta' = \begin{pmatrix} -\partial^2 \ell_T(\theta) / \partial^2 \theta_1 & \cdots & -\partial^2 \ell_T(\theta) / \partial \theta_1 \partial \theta_d \\ \vdots & & \vdots \\ -\partial^2 \ell_T(\theta) / \partial \theta_d \partial \theta_1 & \cdots & -\partial^2 \ell_T(\theta) / \partial^2 \theta_d \end{pmatrix}.$$

Example III.2.1 *Consider the AR(1) model with σ^2 known such that $\theta = \rho$.*

Then

$$\ell_T(\rho) = -\frac{1}{2}\sum_{t=1}^{T}(x - \rho x_{t-1})^2/\sigma^2,\ s_T(\rho) = \sum_{t=1}^{T}(x_t - \rho x_{t-1})\,x_{t-1}/\sigma^2 \text{ and}$$

$$i_T(\theta) = \sum_{t=1}^{T} x_{t-1}^2/\sigma^2.$$

Example III.2.2 *With the ARCH(1) likelihood function given in (III.4) with* $\theta = (\sigma^2, \alpha)'$,

$$w_t = \left(1, x_{t-1}^2\right)',\ \text{ and } \sigma_t^2(\theta) = \sigma^2 + \alpha x_{t-1}^2 = w_t'\theta.$$

With $\sigma^2, \alpha > 0$ *differentiability is ensured and it follows that the score is given by the* $d = 2$ *dimensional vector,*

$$s_T(\theta) = \tfrac{\partial}{\partial\theta}\ell_T(\theta) = -\tfrac{1}{2}\sum_{t=1}^{T}\tfrac{1}{\sigma_t^2(\theta)}(1 - \tfrac{x_t^2}{\sigma_t^2(\theta)})w_t. \tag{III.6}$$

Likewise the observed information is given by

$$i_T(\theta) = -\tfrac{\partial^2}{\partial\theta\partial\theta'}\ell_T(\theta) = -\tfrac{1}{2}\sum_{t=1}^{T}\tfrac{1}{\sigma_t^4(\theta)}(1 - \tfrac{2x_t^2}{\sigma_t^2(\theta)})w_t w_t'. \tag{III.7}$$

Remark III.2.1 *Note that the constraint in the ARCH(1) model of* $\alpha \geq 0$ *means that the likelihood function is not differentiable at* $\alpha = 0$, *and hence the Newton-Raphson algorithm is not directly applicable unless* $\alpha > 0$ *is assumed. Thus in the ARCH model, either* $\alpha > 0$ *is assumed, or the concept of a directional derivative from the right has to be introduced (see later) in order to define a score and observed information. In particular, the constraint that* $\alpha \geq 0$, *implies one has to use optimization subject to an inequality constraint.*

In the case of a differentiable $\ell_T(\theta)$, then by definition the first order derivative evaluated at $\hat{\theta}$ is zero, that is

$$s_T(\hat{\theta}) = 0.$$

Expanding the score around some point θ^* which is close to $\hat{\theta}$, one finds

$$\begin{aligned} 0 = s_T(\hat{\theta}) &= s_T(\theta^*) + \tfrac{\partial^2}{\partial\theta\partial\theta'}\ell_T(\theta^*)\left(\hat{\theta} - \theta^*\right) + ... \\ &= s_T(\theta^*) - i_T(\theta^*)\left(\hat{\theta} - \theta^*\right) + ... \end{aligned} \tag{III.8}$$

Solving for $\hat{\theta}$, and ignoring the "...." terms,

$$\hat{\theta} \simeq \theta^{*} + i_T\left(\theta^{*}\right)^{-1} s_T\left(\theta^{*}\right),$$

provided that $i_T\left(\theta^{*}\right)$ is non-singular, or invertible. This defines the Newton-Raphson iterations,

$$\hat{\theta}^{n} = \hat{\theta}^{n-1} + i_T(\hat{\theta}^{n-1})^{-1} s_T(\hat{\theta}^{n-1}), \tag{III.9}$$

for $n = 1, 2, ...$ and with initial estimator $\hat{\theta}^{0} = \theta^{*}$ for some choice of θ^{*}.

With $\hat{\theta}$ defined as $\hat{\theta} = \hat{\theta}^{N}$ say for some N, then the algorithm may be stopped provided for example,

$$|\hat{\theta} - \hat{\theta}^{N-1}| = |\hat{\theta}^{N} - \hat{\theta}^{N-1}| < \delta,$$

where δ is some small number. Another criterion for stopping the iterations at $\hat{\theta}^{N}$ could equivalently be in terms of the likelihood values,

$$|\ell_T(\hat{\theta}^{N}) - \ell_T(\hat{\theta}^{N-1})| < \delta.$$

Example III.2.3 *In terms of the AR(1) model in Example III.2.2, we find*

$$\begin{aligned}\hat{\rho}^{n} &= \hat{\rho}^{n-1} + \left(\textstyle\sum_{t=1}^{T} x_{t-1}^{2}\right)^{-1} \left(\textstyle\sum_{t=1}^{T} \left(x_t - \hat{\rho}^{n-1} x_{t-1}\right) x_{t-1}\right) \\ &= \hat{\rho}^{n-1} + \left(\textstyle\sum_{t=1}^{T} x_{t-1}^{2}\right)^{-1} \left(\textstyle\sum_{t=1}^{T} x_t x_{t-1}\right) - \hat{\rho}^{n-1} \\ &= \left(\textstyle\sum_{t=1}^{T} x_{t-1}^{2}\right)^{-1} \left(\textstyle\sum_{t=1}^{T} x_t x_{t-1}\right) = \hat{\rho}_{mle}.\end{aligned}$$

That is, after $N = 1$ iteration, and independently of which value ρ^{*} is set to, the algorithm finds the MLE estimator. This is of course nothing but the OLS estimator, which reflects that the result of convergence after one step applies for quadratic likelihood functions, $\ell_T(\theta)$, where quadratic means quadratic in the parameter θ.

In the ARCH(1) case, and in ARCH(k) models in general – even assuming $\alpha > 0$ and hence not $\alpha \geq 0$ – the Newton-Raphson algorithm does converge towards a unique point $\hat{\theta}_{\text{mle}}$ provided "reasonable" initial values $\theta^{*} = (\sigma_{*}^{2}, \alpha_{*})$ have been used, as $\ell_T(\theta)$ is smooth and has a unique maximum. General results states that under *regularity conditions* on the derivatives of $\ell_T(\theta)$ and when a consistent estimator of θ has been used as initial value θ^{*} then the Newton-Raphson is converging rapidly. Such *regularity conditions* are stated in Appendix B and are very similar to the regularity conditions used below for discussion of the asymptotic distribution of the MLE, including consistency and asymptotic normality.

Remark III.2.2 *As mentioned, in the ARCH(1) model one may for example initiate the algorithm at the consistent estimator, $\theta^{*} = \hat{\theta}_{ols}$,from (III.5).*

III.3 Properties of (Q)ML estimators

We establish here that the ML estimator $\hat{\theta} = (\sigma^2, \hat{\alpha})'$ is consistent and asymptotically Gaussian where $\hat{\theta}$ is maximized over Θ_+ rather than Θ, where Θ_+ excludes $\alpha = 0$, that is,

$$\Theta_+ = \mathbb{R}_+^2 = (0, \infty) \times (0, \infty) \subset \Theta.$$

In particular, $\sigma^2, \alpha > 0$ and the log-likelihood function, $\ell_T(\theta)$ is continuously differentiable for all $\theta \in \Theta_+$.

Thus this section establishes that with[1]

$$\hat{\theta} = \arg\max_{\theta \in \Theta_+} \ell_T(\theta),$$

it holds as $T \to \infty$,

$$\hat{\theta} \xrightarrow{P} \theta_0 \text{ and } \sqrt{T}(\hat{\theta} - \theta_0) \xrightarrow{D} N_2(0, \Sigma).$$

Here the covariance Σ is defined below and θ_0 is some (true) parameter value in Θ_+. The arguments explore classical asymptotic Taylor expansions from likelihood theory, see e.g. Billingsley (1961) and Jensen and Rahbek (2004b).

Some early work on ARCH models is given in Weiss (1986), where it is found that the ARCH process x_t should have finite *fourth order moments,* or $\alpha_0 < \frac{1}{\sqrt{3}}$, for consistency and asymptotic normality of $\hat{\theta}$. The requirement of the existence of fourth order moments is strong and in practice (at least for general ARCH models) not fulfilled often, and here we instead establish milder conditions.

Remark III.3.1 *Note that for ARCH and GARCH models, Jensen and Rahbek (2004a,2004b) have shown that even the condition of stationarity of x_t can be omitted for the theory of QML estimators of the (G)ARCH model.*

III.3.1 The score

As noted, asymptotic theory of likelihood estimators is often based on Taylor expansions where the behaviour of the score function plays a key role. Two examples of score functions are considered next.

[1] Strictly speaking, Theorem III.3.1 below states that that there exists a maximizer of the log-likelihood function on a neighborhood of θ_0 that is consistent and asymptotically normal.

Example III.3.1 *Consider the AR(1) model in Examples III.1.1 and III.2.1 from where we have that the score is given by,*

$$s_T(\rho) = \sum_{t=1}^{T} (x_t - \rho x_{t-1}) x_{t-1}/\sigma^2.$$

Evaluated at $\rho = \rho_0$ and $\sigma^2 = \sigma_0^2$ such that data are generated by,

$$x_t = \rho_0 x_{t-1} + \varepsilon_t,$$

with ε_t i.i.d.$N(0, \sigma_0^2)$,

$$\frac{1}{\sqrt{T}} s_T(\rho_0) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} (x_t - \rho_0 x_{t-1}) x_{t-1}/\sigma_0^2 = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \varepsilon_t x_{t-1}/\sigma_0^2.$$

With $f(x_t, x_{t-1}) = (x_t - \rho_0 x_{t-1}) x_{t-1}/\sigma_0^2$ as in Example III.1.1 with ρ_0 such that $|\rho_0| < 1$,

$$\frac{1}{\sqrt{T}} s_T(\rho_0) \overset{D}{\to} N\left(0, 1/\left(1 - \rho_0^2\right)\right).$$

And from the asymptotic distribution of the OLS, or MLE, estimator of ρ in (III.3) one may observe that the asymptotic distribution of the score seems a cruical part of $\hat{\rho}$'s distribution.

The score for the ARCH(1) is a little more involved so we state the results for this as a lemma. A first observation is that as $\theta = (\sigma^2, \alpha)'$ we need to establish convergence to a two-dimensional distribution, whereas the CLT in Theorem II.4.1 applies only to univariate functions $f(X_t, ..., X_{t-m})$. We then note the following well-known result:

Lemma III.3.1 *Consider $(X_t)_{t=1,2,...,T}$ with $X_t = (X_{1t}, ..., X_{pt}) \in \mathbb{R}^p$. Then*

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} X_t \overset{D}{\to} N_p(0, \Sigma),$$

if, and only if, for any nonzero vector $\lambda = (\lambda_1, ..., \lambda_p)'$,

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \lambda' X_t = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} (\lambda_1 X_{1t} + ... + \lambda_p X_{pt}) \overset{D}{\to} N(0, \lambda' \Sigma \lambda).$$

Lemma III.3.2 *With* $\theta_0 = (\alpha_0, \sigma_0^2)' \in \Theta_+$*, assume that the ARCH(1) process* x_t *in (I.1) satisfies the drift criterion and Theorem I.3.2, and* $Ez_t^4 < \infty$*. Then the score is asymptotically Gaussian,*

$$\tfrac{1}{\sqrt{T}} s_T(\theta_0) = \tfrac{1}{\sqrt{T}} \sum_{t=1}^{T} \tfrac{1}{2\sigma_t^2} \left(1 - \tfrac{x_t^2}{\sigma_t^2}\right) w_t \xrightarrow{D} N_2\left(0, \frac{\kappa}{4}\Sigma\right) \qquad \text{(III.10)}$$

where

$$\Sigma = E\left(\frac{w_t w_t'}{\sigma_t^4}\right) = E\begin{pmatrix} 1/\sigma_t^4 & x_{t-1}^2/\sigma_t^4 \\ x_{t-1}^2/\sigma_t^4 & x_{t-1}^4/\sigma_t^4 \end{pmatrix},$$

$\kappa = E(1 - z_t^2)^2 = Ez_t^4 - 1$ *and* $w_t = \left(1, x_{t-1}^2\right)'$.

Remark III.3.2 *Note that if* z_t *is Gaussian,* $\kappa = 2$ *and the condition for Theorem I.3.2 to hold is that,* $E \log \alpha_0 z_t^2 < 0$*, or* $0 < \alpha_0 \lesssim 3.56$.

Proof: We wish to apply the CLT in Theorem II.4.1 to the score in (III.6). From the two-dimensional score,

$$s_T(\theta_0) = -\sum_{t=1}^{T} \tfrac{1}{2\sigma_t^2} \left(1 - \tfrac{x_t^2}{\sigma_t^2}\right) w_t,$$

we construct the univariate f function by multiplying with a vector $\lambda = (\lambda_1, \lambda_2)$, see Lemma III.3.1 and use that $\lambda' w_t = \lambda_1 + \lambda_2 x_{t-1}^2$,

$$f(x_t, x_{t-1}) = f_t = \tfrac{-1}{2\sigma_t^2}(1 - \tfrac{x_t^2}{\sigma_t^2})\left(\lambda_1 + \lambda_2 x_{t-1}^2\right), \qquad \sigma_t^2 = \sigma_0^2 + \alpha_0 x_{t-1}^2.$$

Then we get

$$E(f_t | x_{t-1}, x_{t-2}) = \frac{-\left(\lambda_1 + \lambda_2 x_{t-1}^2\right)}{2\sigma_t^2} E(1 - z_t^2) = 0. \qquad \text{(III.11)}$$

In (III.11), we need the expectation of $1/\sigma_t^2$ and x_{t-1}^2/σ_t^2 to be finite, which holds by the simple inequalities,

$$\frac{1}{\sigma_t^2} \leq \frac{1}{\sigma_0^2}, \qquad \frac{x_{t-1}^2}{\sigma_t^2} \leq \frac{1}{\alpha_0}, \qquad \text{(III.12)}$$

as $\sigma_0^2, \alpha_0 > 0$. Next, consider the variance,

$$\begin{aligned} E(f_t^2) &= E\left(E\left(\tfrac{1}{4\sigma_t^4}(1 - \tfrac{x_t^2}{\sigma_t^2})^2 \left(\lambda_1 + \lambda_2 x_{t-1}^2\right)^2 | x_{t-1}, x_{t-2}\right)\right) \\ &= E\left(\tfrac{1}{4\sigma_t^4}\left(\lambda_1 + \lambda_2 x_{t-1}^2\right)^2 E(1 - z_t^2)^2\right) \\ &= \frac{E(1 - z_t^2)^2}{4} E\left(\tfrac{1}{\sigma_t^4}\left(\lambda_1 + \lambda_2 x_{t-1}^2\right)^2\right). \end{aligned}$$

Note first that by definition,

$$\frac{E(1-z_t^2)^2}{4} = \kappa/4,$$

with $\kappa = 2$ if z_t is Gaussian. Next,

$$E\left(\tfrac{1}{\sigma_t^4}\left(\lambda_1 + \lambda_2 x_{t-1}^2\right)^2\right) = E\left(\frac{1}{\sigma_t^4}\lambda' w_t w_t' \lambda\right) = \lambda' E\left(\frac{1}{\sigma_t^4} w_t w_t'\right)\lambda = \lambda' \Sigma \lambda,$$

and the result follows. □

Remark: Note that the result requires finite 4^{th} order moments of z_t, but there are no moment requirements for x_t.

III.3.2 Consistency and asymptotic normality for the QMLE

We state here a general result from Jensen and Rahbek (2004b), which can be used to establish consistency and asymptotic normality of general QML estimators of with θ of dimension d, $\theta = (\theta_1, ..., \theta_d)'$. The approach is *local* in the sense that it describes the behaviour of the log-likelihood function $\ell_T(\theta)$ close to the true value θ_0 of the parameter θ. It is not much different from other classic approaches of verifying regularity conditions in asymptotic likelihood theory and has the advantage that it fits within our framework.

To give an idea of the conditions in the theorem, recall the Taylor expansion we applied in connection with the Newton-Raphson iterations in (III.8). Stating this again with θ_0 replacing the initial value θ^*, with $\ell_T(\theta)$ differentiable,

$$0 = s_T(\hat{\theta}) = s_T(\theta_0) - i_T(\theta_0)\left(\hat{\theta} - \theta_0\right) + ... \tag{III.13}$$

we have, again ignoring "...",

$$\sqrt{T}\left(\hat{\theta} - \theta_0\right) \simeq \left(\frac{1}{T} i_T(\theta_0)\right)^{-1} \frac{1}{\sqrt{T}} s_T(\theta_0). \tag{III.14}$$

Regularity condition (A.1) below states that the score $s_T(\theta)$ normalized by $\sqrt{T}$ is indeed asymptotically Gaussian with a $(d \times d)-$dimensional covariance matrix Ω_S when evaluated at the true parameter θ_0.

Likewise, regularity condition (A.2) states that the information matrix $\frac{1}{T} i_T(\theta)$ when evaluated at θ_0 converges in probability to a constant $(d \times$

$d)-$dimensional matrix $\Omega_I > 0$. Hence (III.14) would immediately imply that

$$\sqrt{T}\left(\hat{\theta}-\theta_0\right) \overset{D}{\rightarrow} \Omega_I^{-1} N_d\left(0,\Omega_S\right) \overset{D}{=} N_d\left(0,\Omega_I^{-1}\Omega_S\Omega_I^{-1}\right),$$

which is in fact the asymptotic distribution of the QMLE, see (B.3) below.

In (III.13) we ignored "...". In the case where there is only one parameter to be estimated such that θ is one-dimensional, or $d = 1$, a second order Taylor expansion is given by,

$$0 = s_T(\hat{\theta}) = s_T\left(\theta_0\right) - i_T\left(\theta_0\right)\left(\hat{\theta}-\theta_0\right) + \partial^3 \ell_T\left(\theta^*\right)/\partial\theta^3\left(\hat{\theta}-\theta_0\right)^2/2,$$

where θ^* is some point in between $\hat{\theta}$ and θ_0. Dividing by $\sqrt{T}$ we get,

$$\begin{aligned} 0 = &\underbrace{\left(\frac{1}{\sqrt{T}}s_T\left(\theta_0\right)\right)}_{\text{(A.1)}} - \underbrace{\left(\frac{1}{T}i_T\left(\theta_0\right)\right)}_{\text{(A.2)}}\left(\sqrt{T}\left(\hat{\theta}-\theta_0\right)\right) \\ &+ \underbrace{\left(\frac{1}{T}\partial^3\ell_T\left(\theta^*\right)/\partial\theta^3\right)}_{\text{(A.3)}}\left(\sqrt{T}\left(\hat{\theta}-\theta_0\right)\right)\left(\hat{\theta}-\theta_0\right), \end{aligned}$$

or simply,

$$\sqrt{T}\left(\hat{\theta}-\theta_0\right) = \left[\underbrace{\frac{1}{T}i_T\left(\theta_0\right)}_{\text{(A.2)}} - \underbrace{\frac{1}{T}\partial^3\ell_T\left(\theta^*\right)/\partial\theta^3}_{\text{(A.3)}}\left(\hat{\theta}-\theta_0\right)\right]^{-1}\underbrace{\frac{1}{\sqrt{T}}s_T\left(\theta_0\right)}_{\text{(A.1)}}.$$

The condition (A.3) implies that the term $\frac{1}{T}\partial^3\ell_T\left(\theta^*\right)/\partial\theta^3\left(\hat{\theta}-\theta_0\right) \overset{P}{\rightarrow} 0$ as $\hat{\theta}$ is consistent, see (B.2) below. Condition (A.3) states that the third order derivative

$$\frac{1}{T}\partial^3\ell_T\left(\theta\right)/\partial\theta_i\partial\theta_j\partial\theta_k$$

for $i,j,k = 1,...,d$ in absolute value is bounded by a constant (in probability) for any θ in a neigbourhood of θ_0, that is, for $\theta \in N\left(\theta_0\right)$.

For the ARCH(1) model a neigbourhood of $\theta_0 = (\sigma_0^2,\alpha_0)'$ can be chosen as,

$$N\left(\theta_0\right) = N\left(\sigma_0^2,\alpha_0\right) = [\sigma_L^2,\sigma_U^2]\times[\alpha_L,\alpha_U], \tag{III.15}$$

where

$$0 < \sigma_L^2 < \sigma_0^2 < \sigma_U^2 < \infty \text{ and } 0 < \alpha_L < \alpha_0 < \alpha_U < \infty.$$

Likewise, for $\theta = (\theta_1,...,\theta_d)' \in \Theta$, where Θ is a product of intervals I_i which can be (subintervals of) $\mathbb{R},\mathbb{R}_+$ or $\mathbb{R}^+$ for $i = 1,...,d$, such that $\Theta = I_1\times...\times I_d$, one may choose $N\left(\theta\right)$ as,

$$N\left(\theta_0\right) = [\theta_{1L},\theta_{1U}]\times...\times[\theta_{dL},\theta_{dU}], \tag{III.16}$$

where $\theta_{iL} < \theta_{i0} < \theta_{iU}$ for $i = 1, 2, ..., d$.

Theorem III.3.1 *Consider the log-likelihood function $\ell_T(\theta)$, which is a function of the observations $X_0, X_1..., X_T$ and the parameter $\theta \in \Theta \subseteq \mathbb{R}^d$. Assume that $\ell_T(\theta)$ is three times differentiable in θ with all derivatives continuous. With θ_0 inside the set Θ, assume that:*

(A.1): $\frac{1}{\sqrt{T}} s_T(\theta_0) = \frac{1}{\sqrt{T}} \partial \ell_T(\theta_0)/\partial\theta \xrightarrow{D} N_d(0, \Omega_S),\ \ \Omega_S > 0.$

(A.2): $\frac{1}{T} i_T(\theta_0) = -\frac{1}{T}\partial^2 \ell_T(\theta_0)/\partial\theta\partial\theta' \xrightarrow{P} \Omega_I > 0.$

(A.3): $\max_{h,i,j=1,...,k} \sup_{\theta \in N(\theta_0)} \left| \frac{1}{T} \frac{\partial^3 \ell_T(\theta)}{\partial\theta_h \partial\theta_i \partial\theta_j} \right| \leq c_T,$

where $N(\theta_0)$ is a neighborhood of θ_0, see (III.16), and $0 \leq c_T \xrightarrow{P} c$, $0 < c < \infty$. Then in a fixed open neigborhood $U(\theta_0) \subseteq N(\theta_0)$ of θ_0:

(B.1): As $T \to \infty$, there exists a unique maximum point $\hat{\theta}$ of $\partial \ell_T(\hat{\theta})$, which solves the estimating equation, $\partial \ell_T(\hat{\theta})/\partial\theta = 0$ (with probability tending to one).

(B.2): As $T \to \infty$, $\hat{\theta} \xrightarrow{P} \theta_0$.

(B.3): As $T \to \infty$, $\sqrt{T}(\hat{\theta} - \theta_0) \xrightarrow{D} N_d(0, \Omega_I^{-1}\Omega_S\Omega_I^{-1})$.

Note that in the case of MLE rather than QMLE,

$$\Omega_S = c\Omega_I$$

with c some constant, in which case (B.3) reduces to

$$\sqrt{T}(\hat{\theta} - \theta_0) \xrightarrow{D} N_d(0, c\Omega_I^{-1}),$$

where the constant c depends on whether the likelihood function has been scaled by some constant or not.

If it is *not scaled*, such as in our case with the likelihood function in (III.4), $c = 1$ and the result in (B.3) reduces further to the classic likelihood result that,

$$\sqrt{T}(\hat{\theta} - \theta_0) \xrightarrow{D} N_d(0, \Omega_I^{-1}). \tag{III.17}$$

This clearly demonstrates the importance of the second derivative or the information, see (A.2). Often the Ω_I is consistently estimated by

$$\hat{\Omega}_I = \frac{1}{T} i_T(\hat{\theta}), \tag{III.18}$$

that is, minus the second derivative (divided by T) of the likelihood function evaluated at $\hat{\theta}$.

As an example of a scaled likelihood function, we could choose to maximize the likelihood function in (III.4) multiplied by $c = 2$, that is,

$$\ell_T^{\text{new}}(\theta) = c\ell_T(\theta) = -\sum_{t=1}^{T}\left(\log\sigma_t^2(\theta) + \frac{x_t^2}{\sigma_t^2(\theta)}\right).$$

Then $\Omega_S^{\text{new}} = c^2\Omega_S$, and $\Omega_I^{\text{new}} = c\Omega_I$, such that $\Omega_S^{\text{new}} = c\Omega_I^{\text{new}}$, and (B.3) reduces to $\sqrt{T}(\hat{\theta} - \theta_0) \xrightarrow{D} N_2(0, c\Omega_I^{-1})$. This is clearly somewhat confusing, and to avoid these issues we will predominantly use unscaled likelihood functions.

Remark III.3.3 *Note also that for the QMLE, the identity $\Omega_S = c\Omega_I$ does not hold in general.*

III.3.3 Application of Theorem III.3.1 to the ARCH(1) model

For the ARCH(1) model we have already seen in Lemma III.3.2 that the score is asymptotically normal, that is, we showed that with $\theta_0 \in \Theta_+$,

$$\tfrac{1}{\sqrt{T}}s_T(\theta_0) = \tfrac{1}{\sqrt{T}}\sum_{t=1}^{T}\tfrac{1}{2\sigma_t^2}\left(1 - \tfrac{x_t^2}{\sigma_t^2}\right)w_t \xrightarrow{D} N_2\left(0, \frac{\kappa}{4}\Sigma\right),$$

where

$$\Sigma = E\left(\frac{w_t w_t'}{\sigma_t^4}\right) = E\begin{pmatrix} 1/\sigma_t^4 & x_{t-1}^2/\sigma_t^4 \\ x_{t-1}^2/\sigma_t^4 & x_{t-1}^4/\sigma_t^4 \end{pmatrix}, \tag{III.19}$$

$\kappa = E(1 - z_t^2)^2$ and $w_t = \left(1, x_{t-1}^2\right)'$.

We can now state the following result for the ARCH(1) model:

Theorem III.3.2 *Consider the QMLE $\hat{\theta} = (\hat{\sigma}^2, \hat{\alpha})'$ found by maximizing the Gaussian likelihood function in (III.4) over $\theta \in \Theta_+ = \mathbb{R}_+^2$,*

$$\ell_T(\theta) = -\tfrac{1}{2}\sum_{t=1}^{T}\left(\log\sigma_t^2(\theta) + \frac{x_t^2}{\sigma_t^2(\theta)}\right), \quad \sigma_t^2(\theta) = \sigma^2 + \alpha x_{t-1}^2.$$

Assume for the ARCH(1) process x_t defined in (I.1) that z_t are i.i.d.(0,1), $Ez_t^4 < \infty$ and furthemore that x_t satisfies the drift criterion at $\theta_0 = (\sigma_0^2, \alpha_0)'$, that is $E\log\alpha_0 z_t^2 < 0$ for $\alpha_0 > 0$. Then the conclusions of Theorem III.3.1 hold. In particular, $\hat{\theta} \xrightarrow{P} \theta_0$ and

$$\sqrt{T}(\hat{\theta} - \theta_0) \xrightarrow{D} N_2\left(0, \kappa\Sigma^{-1}\right), \tag{III.20}$$

where $\Sigma = E(\frac{w_t w_t'}{\sigma_t^4})$, $\kappa = E(1 - z_t^2)^2$ and $w_t = \left(1, x_{t-1}^2\right)'$.

Remark III.3.4 *If z_t is Gaussian, such that $\hat{\theta}$ is the MLE, then $\kappa\Sigma^{-1} = \Omega_I^{-1}$, $\kappa = 2$ and moreover the drift criterion is satisfied if $E \log \alpha_0 z_t^2 < 0$ or $\alpha_0 \lesssim 3.56$.*

Proof: We apply Theorem III.3.1 and show that each of the conditions (A.1)–(A.3) hold.

Condition (A.1): By Lemma III.3.2, (A.1) holds, that is,

$$\tfrac{1}{\sqrt{T}} s_T\left(\theta_0\right) = \tfrac{1}{\sqrt{T}} \Sigma_{t=1}^T \tfrac{1}{2\sigma_t^2}(1 - \tfrac{x_t^2}{\sigma_t^2}) w_t \xrightarrow{D} N_2\left(0, \Omega_S\right) \tag{III.21}$$

with $\Omega_S = \frac{\kappa}{4}\Sigma$, $\kappa = E(1 - z_t^2)^2$ and Σ is given in (III.19).

So what is missing is to show the remaining part of the regularity conditions (A.2) and (A.3).

Condition (A.2): By (III.7), and the LLN in Theorem I.7,

$$-\tfrac{1}{T} \tfrac{\partial^2}{\partial\theta\partial\theta'} \ell_T\left(\theta\right)|_{\theta=\theta_0} = -\tfrac{1}{T} \sum_{t=1}^T \tfrac{1}{2\sigma_t^4}(1 - \tfrac{2x_t^2}{\sigma_t^2}) w_t w_t' \xrightarrow{P} \tag{III.22}$$

$$-E((1 - 2z_t^2)\frac{w_t w_t'}{2\sigma_t^4}) = \tfrac{1}{2}\Sigma = \Omega_I$$

as $E((1 - 2z_t^2)\frac{w_t w_t'}{\sigma_t^4}) = E(E((1 - 2z_t^2)\frac{w_t w_t'}{\sigma_t^4}|x_{t-1})) = E\left(1 - 2z_t^2\right) E(\frac{w_t w_t'}{\sigma_t^4})$.

Then as

$$\Omega_I^{-1}\Omega_S\Omega_I^{-1} = \left(\tfrac{1}{2}\Sigma\right)^{-1}\left(\tfrac{\kappa}{4}\Sigma\right)\left(\tfrac{1}{2}\Sigma\right)^{-1} = \kappa\Sigma^{-1}, \tag{III.23}$$

(III.20) holds provided (A.3) holds.

Condition (A.3): Consider $N\left(\theta_0\right)$ as defined in (III.15),

$$N\left(\theta_0\right) = \left[\sigma_L^2, \sigma_U^2\right] \times \left[\alpha_L, \alpha_U\right],$$

where $0 < \sigma_L^2 < \sigma_0^2 < \sigma_U^2$ and $0 < \alpha_L < \alpha_0 < \alpha_U$. The third order derivatives normalized by T are given by:

$$\frac{1}{T}\frac{\partial^3 \ell_T(\theta)}{\partial\alpha^2\partial\theta'} = -\frac{1}{T}\sum_{t=1}^T \left(1 - 3\frac{x_t^2}{\sigma_t^2\left(\theta\right)}\right)\frac{x_{t-1}^4}{\sigma_t^6\left(\theta\right)} w_t$$

$$\frac{1}{T}\frac{\partial^3 \ell_T(\theta)}{\partial(\sigma^2)^2\partial\theta'} = -\frac{1}{T}\sum_{t=1}^T \left(1 - 3\frac{x_t^2}{\sigma_t^2\left(\theta\right)}\right)\frac{1}{\sigma_t^6\left(\theta\right)} w_t$$

Then for any θ in $N(\theta_0)$, e.g.

$$\begin{aligned}
\left|\frac{1}{T}\frac{\partial^3 \ell_T(\theta)}{\partial \alpha^3}\right| &= \left|\frac{1}{T}\sum_{t=1}^{T}\left(1-3\frac{x_t^2}{\sigma_t^2(\theta)}\right)\frac{x_{t-1}^6}{\sigma_t^6(\theta)}\right| \\
&\leq \frac{1}{\alpha_L^3}\frac{1}{T}\sum_{t=1}^{T}\left(1+3\frac{x_t^2}{\sigma_t^2(\theta)}\right) \\
&= \frac{1}{\alpha_L^3}\left(1+\frac{3}{T}\sum_{t=1}^{T}\frac{x_t^2}{\sigma_t^2}\frac{\sigma_t^2}{\sigma_t^2(\theta)}\right) \\
&\leq \frac{1}{\alpha_L^3}\left(1+\frac{3}{T}\sum_{t=1}^{T}z_t^2\left(\frac{\sigma_0^2}{\sigma_L^2}+\frac{\alpha_0}{\alpha_L}\right)\right)
\end{aligned}$$

using,

$$\frac{\sigma_t^2}{\sigma_t^2(\theta)} := \frac{\sigma_0^2+\alpha_0 x_{t-1}^2}{\sigma^2+\alpha x_{t-1}^2} \leq \frac{\sigma_0^2}{\sigma_L^2}+\frac{\alpha_0}{\alpha_L}. \tag{III.24}$$

Identical considerations can be used for the other terms, and (A.3) holds by the LLN applied to averages of z_t^2.

III.3.4 Consistent estimation of the covariance

From Theorem III.3.2 we conclude that, if the z_t's are $N(0,1)$ distributed, then $\hat{\theta} = (\hat{\sigma}^2, \hat{\alpha})'$ is asymptotically Gaussian distributed with covariance,

$$\Omega_I^{-1} = 2\Sigma^{-1}.$$

As mentioned this is simple to provide a consistent estimator for as,

$$\frac{1}{T}i_T(\hat{\theta}) = -\frac{1}{T}\tfrac{\partial^2}{\partial\theta\partial\theta'}\ell_T(\theta)\left.\right|_{\theta=\hat{\theta}} \xrightarrow{P} \Omega_I.$$

Hence for the Gaussian MLE, one can use that

$$\hat{\theta}-\theta_0 = \begin{pmatrix}\hat{\sigma}^2-\sigma_0^2\\ \hat{\alpha}-\alpha_0\end{pmatrix} \simeq N_2\left(0,\frac{1}{T}\left(\frac{1}{T}i_T(\hat{\theta})\right)^{-1}\right) = N_2\left(0,\left(-\tfrac{\partial^2}{\partial\theta\partial\theta'}\ell_T(\theta)\left.\right|_{\theta=\hat{\theta}}\right)^{-1}\right),$$

and report t-ratios using this.

Remark III.3.5 *Most software reports second derivatives of the likelihood function directly.*

III.3.4.1 QML estimator

From Theorem III.3.2 we can also conclude that, if the z_t are not $i.i.d.N\left(0,1\right)$, then $\hat{\theta}=\left(\hat{\sigma}^2,\hat{\alpha}\right)'$ is still asymptotically Gaussian distributed with covariance,

$$\Omega_I^{-1}\Omega_S\Omega_I^{-1}.$$

One may still provide a consistent estimator of the covariance without specifying the distribution of the inovations z_t. For example, Ω_I is consistently estimated by $\frac{1}{T}i_T(\hat{\theta})$ as

$$\frac{1}{T}i_T(\hat{\theta})=-\frac{1}{T}\tfrac{\partial^2}{\partial\theta\partial\theta'}\ell_T\left(\theta\right)|_{\theta=\hat{\theta}}\overset{P}{\to}\Omega_I.$$

For a consistent estimator of Ω_S, that is the variance of the score, write the likelihood function in (III.4) as

$$\ell_T\left(\theta\right)=\sum_{t=1}^{T}l_t\left(\theta\right),\qquad l_t\left(\theta\right)=\tfrac{1}{2}\left(\log\sigma_t^2\left(\theta\right)+\frac{x_t^2}{\sigma_t^2\left(\theta\right)}\right).$$

Then the score equals,

$$\frac{1}{\sqrt{T}}s_T\left(\theta\right)=\frac{1}{\sqrt{T}}\tfrac{\partial}{\partial\theta}\ell_T\left(\theta\right)=\frac{1}{\sqrt{T}}\sum_{t=1}^{T}s_t\left(\theta\right),\qquad\text{where } s_t\left(\theta\right)=\tfrac{\partial}{\partial\theta}l_t\left(\theta\right),$$

and Ω_S is consistently estimated by (this is often referred to as the "outer-product"),

$$\hat{\Omega}_S=\frac{1}{T}\sum_{t=1}^{T}s_t(\hat{\theta})s_t(\hat{\theta})'.$$

Collecting terms, with $\hat{\theta}$ the (Gaussian) QMLE,

$$\begin{aligned}
&\hat{\theta}-\theta_0=\begin{pmatrix}\hat{\sigma}^2-\sigma_0^2\\ \hat{\alpha}-\alpha_0\end{pmatrix}\simeq N_2\left(0,\frac{1}{T}\hat{\Omega}_I^{-1}\hat{\Omega}_S\hat{\Omega}_I^{-1}\right)\\
&=N_2\left(0,\frac{1}{T}\left(\frac{1}{T}i_T\left(\hat{\theta}\right)\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^{T}s_t(\hat{\theta})s_t(\hat{\theta})'\right)\left(\frac{1}{T}i_T\left(\hat{\theta}\right)\right)^{-1}\right)\\
&=N_2\left(0,\left(-\tfrac{\partial^2}{\partial\theta\partial\theta'}\ell_T\left(\theta\right)|_{\theta=\hat{\theta}}\right)^{-1}\sum_{t=1}^{T}s_t(\hat{\theta})s_t(\hat{\theta})'\left(-\tfrac{\partial^2}{\partial\theta\partial\theta'}\ell_T\left(\theta\right)|_{\theta=\hat{\theta}}\right)^{-1}\right)\\
&=N_2\left(0,\left(\sum_{t=1}^{T}\tfrac{\partial^2}{\partial\theta\partial\theta'}l_t\left(\theta\right)|_{\theta=\hat{\theta}}\right)^{-1}\left(\sum_{t=1}^{T}\tfrac{\partial}{\partial\theta}l_t\left(\theta\right)\tfrac{\partial}{\partial\theta'}l_t\left(\theta\right)|_{\theta=\hat{\theta}}\right)\left(\sum_{t=1}^{T}\tfrac{\partial^2}{\partial\theta\partial\theta'}l_t\left(\theta\right)|_{\theta=\hat{\theta}}\right)^{-1}\right)
\end{aligned}$$

where the first derivatives like the second can be obtained from standard software during optimization.

III.4 LR | The likelihood ratio test

A cruical part of "inference" is testing hypotheses on parameters of the model. For example, testing in the ARCH model that α has some value, say, $H : \alpha = \alpha_0$, $\alpha_0 > 0$, and more generally testing the simple hypothesis,

$$H : \theta = \theta_0,$$

with $\theta_0 \in \Theta_+$.

Remark III.4.1 *Note that $\alpha_0 = 0$ is excluded here and is discussed later.*

The likelihood ratio test statistic is by definition given by,

$$LR= -2\left(\ell\left(\theta_0\right) - \ell(\hat{\theta})\right)$$

and is under the regularity conditions (A.1)-(A.3) in Theorem III.3.1 asymptotically χ^2_d where d is the number of parameters in θ. This follows by expanding the log-likelihood function similar to the already applied expansions. Thus,

$$\begin{aligned} LR &= -2\left(\ell\left(\theta_0\right) - \ell(\hat{\theta})\right) \qquad \text{(III.25)} \\ &= 2\left(s_T(\theta_0)\left(\hat{\theta} - \theta_0\right) - \frac{1}{2}\left(\hat{\theta} - \theta_0\right)' i_T\left(\theta_0\right)\left(\hat{\theta} - \theta_0\right) + ...\right) \end{aligned}$$

which combined with

$$s_T(\hat{\theta}) = s_T\left(\theta_0\right) - i_T\left(\theta_0\right)\left(\hat{\theta} - \theta_0\right) + ...$$

gives,

$$LR = \left(\hat{\theta} - \theta_0\right)' i_T\left(\theta_0\right)\left(\hat{\theta} - \theta_0\right) + ...$$

The theory above implies that the terms indicated by "..." can be ignored and we write "$o_p\left(1\right)$" henceforth for terms that tends to zero in probablity. That is, under the conditions of Theorem III.3.1,

$$LR = \left(\hat{\theta} - \theta_0\right)' i_T\left(\theta_0\right)\left(\hat{\theta} - \theta_0\right) + o_p\left(1\right)$$

Applying the results on the asymptotic distribution of $\hat{\theta}$, one finds,

$$\begin{aligned} LR &= \sqrt{T}\left(\hat{\theta} - \theta_0\right)' T^{-1} i_T\left(\theta_0\right)\sqrt{T}\left(\hat{\theta} - \theta_0\right) + o_p\left(1\right) \\ &\overset{D}{\to} \mathcal{LR} \overset{D}{=} \gamma'\Omega_I\gamma, \end{aligned}$$

where γ is $N_d\left(0, \Omega_I^{-1}\right)$ distributed. And, with $\gamma^* = \Omega_I^{1/2}\gamma \overset{D}{=} N_d(0, I_d)$, one has for the likelihood ratio statistic, $LR \overset{D}{\to} \mathcal{LR}$, where

$$\mathcal{LR} \overset{D}{=} \gamma' \Omega_I \gamma \overset{D}{=} \gamma^{*\prime}\gamma^* \overset{D}{=} \chi_d^2.$$

Remark III.4.2 *If one is interested in only testing a hypothesis on part of θ, the above results can be modified accordingly. In general with a hypothesis restricting k of the d parameters in θ to be known,*

$$LR \overset{D}{\to} \chi_k^2.$$

For example, testing $H : \alpha = \alpha_0$ in the ARCH(1) model, gives

$$LR = -2\left(\ell(\tilde{\theta}) - \ell(\hat{\theta})\right) \overset{D}{\to} \chi_1^2,$$

where $\tilde{\theta}$ is the MLE where $\alpha = \alpha_0$ is fixed and only σ^2 is estimated by MLE.

III.4.1 QLR asymptotic distribution

For the case of QMLE,

$$\sqrt{T}\left(\hat{\theta} - \theta_0\right) \overset{D}{\to} \gamma_{\text{QMLE}} = N_d\left(0, \Omega_I^{-1}\Omega_S\Omega_I^{-1}\right)$$

and hence

$$QLR \overset{D}{\to} \mathcal{LR}_Q = \gamma_{\text{QMLE}}^{*\prime}\gamma_{\text{QMLE}}^*$$

where

$$\gamma_{\text{QMLE}}^* = N_d\left(0, \Omega_I^{-1/2}\Omega_S\Omega_I^{-1/2}\right).$$

In general, $\mathcal{LR}_Q$ is not χ_d^2 which, needless to say, have implications if this is ignored. Often however,

$$\mathcal{LR}_Q = s\chi_d^2,$$

where s is some scaling factor.

For example, for the ARCH(1) case, where from (III.23),

$$\Omega_I^{-1}\Omega_S\Omega_I^{-1} = \left(\tfrac{1}{2}\Sigma\right)^{-1}\left(\tfrac{\kappa}{4}\Sigma\right)\left(\tfrac{1}{2}\Sigma\right)^{-1} = \kappa\Sigma^{-1},$$

we have

$$\gamma_{\text{QMLE}} = N_d\left(0, \kappa\Sigma^{-1}\right), \text{ and } \Omega_I = \tfrac{1}{2}\Sigma.$$

Hence in this case,

$$QLR \overset{D}{\to} N_d\left(0, \kappa\Sigma^{-1}\right)\tfrac{1}{2}\Sigma N_d\left(0, \kappa\Sigma^{-1}\right) \overset{D}{=} c\chi_2^2,$$

where $c = \kappa/2$ with $\kappa = E(1 - z_t^2)^2$. This means that the scaling factor, c (where $c = 1$ if z_t are Gaussian) in general appears in the limiting distribution of the quasi-LR statistic.

Remark III.4.3 *Note that c may be estimated consistently by $\hat{c} = \sum_{t=1}^{T}(1 - \hat{z}_t^2)^2/2$, where $\hat{z}_t = x_t/\sqrt{\sigma_t^2(\hat{\theta})}$ are the standardized residuals. Hence $QLR/\hat{c} \xrightarrow{D} \chi_2^2$.*

III.5 Allowing $\alpha_0 = 0$ | Asymptotics at the boundary

The theory discussed so far has been based on a second (or third) order Taylor expansion of the log-likelihood function $\ell_T(\theta)$ and it was assumed that $\alpha_0 > 0$ (in addition to $\sigma_0^2 > 0$) in the derivations in order to ensure differentiability, and hence validity of the Taylor expansion. Thus Theorem III.3.2 establishes consistency and asymptotic normality of $\hat{\theta}$ for $\theta_0 = (\sigma_0^2, \alpha_0) \in \Theta_+ = \mathbb{R}_+^2$.

If $\ell_T(\theta)$ is instead maximized over Θ, where

$$\Theta = \mathbb{R}_+ \times \mathbb{R}^+ = (0, \infty) \times [0, \infty),$$

the log-likelihood function is not differentiable at $\theta_0 = (\sigma_0^2, 0)' \in \Theta$. That is, for $\alpha_0 = 0$, then – even with $\sigma_0^2 > 0$ – by definition, $\theta_0 = (\sigma_0^2, 0)'$ is not an interior point of Θ. It is a so-called boundary point. However, obviously asymptotic theory is needed for the case where $\alpha_0 = 0$, as for example one may want to test the hypohesis of no ARCH,

$$H : \alpha = 0.$$

To discuss this, observe first that (with σ^2 fixed) by definition the log-likelihood function $\ell_T(\theta)$ is differentiable from the right at $\alpha_0 = 0$. Stated differently, the directional derivative of $\ell_T(\theta)$ is well-defined at $\theta_0 = (\sigma_0^2, 0) \in \Theta = \mathbb{R}_+ \times \mathbb{R}^+$. This is exploited in Andrews (1999), see also Silvapulle and Sen (2005), as the usual expansion of $\ell_T(\theta)$ holds in terms of directional derivatives rather than classical derivatives.

In fact, similar to the expansion used for the LR statistic in (III.25), one has in terms of directional derivatives for θ_0 which includes the boundary of $\Theta = \mathbb{R}_+ \times \mathbb{R}^+$,

$$\ell_T(\theta) - \ell_T(\theta_0) = s_T(\theta_0)'(\theta - \theta_0) - \frac{1}{2}(\theta - \theta_0)' i_T(\theta_0)(\theta - \theta_0) + ... \quad \text{(III.26)}$$

where (the directional) $s_T(\theta_0)$ and $i_T(\theta_0)$ satisfy the regularity conditions (A.1)–(A.3) of Theorem III.3.1. As to the remainder term(s), it follows by Theorem 3 in Andrews (1999) that under regularity conditions that these can be ignored and the expansion used to derive the asymptotic distribution of $\hat{\theta}$ independently of whether θ_0 is an interior point or not.

III.5.1 Regularity conditions for the QMLE when $\alpha_0 \geq 0$

Conditions under which the expansion in (III.26) can be used for deriving the limiting distribution of $\hat{\theta}$ are stated for the compact parameter space $\Theta_C \subset \Theta$. This is common in much of the literature on asymptotic theory in nonlinear models, such as the ARCH(1) model, where often the parameter set Θ as here is replaced by a compact subset Θ_C, $\Theta_C \subset \Theta$. For the ARCH model,

$$\Theta_C = \left[\sigma_L^2, \sigma_U^2\right] \times [0, \alpha_U] \subset \Theta. \tag{III.27}$$

where $0 < \sigma_L^2 < \sigma_U^2 < \infty$, $0 < \alpha_U < \infty$. Moreover, σ_U^2 and α_U are here aribrarily large (and σ_L^2 arbitrarily close to zero). This way[2] the set Θ_C is defined such that maximizing the likelihood function over Θ or Θ_C effectively makes little, or no, difference. Moreover, the fact that Θ_C is compact means that existence of maxima (or minima) for continuous functions by definition are guaranteed, and also that the upper and lower bounds for the parameters may be exploited when deriving the theory.

Assumption III.5.1 *With $\theta_0 \in \Theta_C$ in (III.27), assume that:*

(C.1) $\hat{\theta}_C = \arg\max_{\theta \in \Theta_c} \ell_T(\theta) \xrightarrow{P} \theta_0$.
(C.2) Reguarity conditions (A.1)-(A.3) in Theorem III.3.1 hold

Here condition (C.1) is a high-level condition which requires that $\hat{\theta}_C$ is consistent. For the ARCH(1) model with $\hat{\theta}$ maximizing $\ell_T(\theta)$ over $\Theta_+ = \mathbb{R}_+^2$ this was established as part of Theorem III.3.2 using differentiability of $\ell_T(\theta)$. We briefly discuss below how to modify the arguments to establish consistency of $\hat{\theta}_C$ when maximizing over Θ_C.

As to condition (C.2) this was established to hold for the ARCH(1) model for $\sigma_0^2, \alpha_0 > 0$, that is, for θ_0 in the interior of Θ_C. Moreover, it was assumed that x_t was geometrically ergodic,

$$E \log\left(\alpha_0 z_t^2\right) < 0,$$

and $E z_t^4$. We show next that when $\alpha_0 = 0$, (A.1)-(A.3) hold under the mild condition that $E z_t^4 < \infty$.

[2] Note that in the following derivations, it is implicitly assumed that $\sigma_L^2 < \sigma_0^2 < \sigma_U^2$ and $\alpha_0 < \alpha_U$.

III.5.1.1 Condition (C.2) for $\alpha_0 = 0$.

For the directional derivative case where $\alpha_0 = 0$, observe that in this case the ARCH process becomes *i.i.d.*, that is

$$x_t = \sigma_0 z_t.$$

The score at $\theta_0 = (\sigma_0^2, 0)'$ is given in (III.21) which when $\alpha_0 = 0$ reduces to,

$$\frac{1}{\sqrt{T}} s_T(\theta_0) = \tfrac{1}{\sqrt{T}} \Sigma_{t=1}^T \tfrac{1}{2\sigma_t^2}(1 - \tfrac{x_t^2}{\sigma_t^2}) w_t = \tfrac{1}{\sqrt{T}} \Sigma_{t=1}^T \tfrac{1}{2\sigma_0^2}(1 - z_t^2)\left(1, \sigma_0^2 z_{t-1}^2\right)'.$$

Thus the CLT applies to $s_T(\theta_0)$ provided that $Ez_t^4 < \infty$,

$$\frac{1}{\sqrt{T}} s_T(\theta_0) \xrightarrow{D} N(0, \Omega_S), \quad \Omega_S = \frac{\kappa}{4}\Sigma_0,$$

where, using $\kappa = E(1 - z_t^2)^2 = Ez_t^4 - 1$,

$$\Sigma_0 = \begin{pmatrix} 1/\sigma_0^4 & 1/\sigma_0^2 \\ 1/\sigma_0^2 & 1 + \kappa \end{pmatrix}.$$

Likewise for the information, see (III.22), at $\theta_0 = (\sigma_0^2, 0)'$ and with $w_t = \left(1, \sigma_0^2 z_{t-1}^2\right)'$,

$$\tfrac{1}{T} i_T(\theta) = -\tfrac{1}{T} \sum_{t=1}^T \tfrac{1}{2\sigma_0^4}(1 - 2z_t^2) w_t w_t' \xrightarrow{P} \tfrac{1}{2}\Sigma_0 = \Omega_I,$$

and similarly for the third order derivatives.

III.5.2 Asymptotics of the QMLE when $\alpha_0 \geq 0$.

To present the results, define initially the score normalized by the information,

$$Z_T = i_T(\theta_0)^{-1} \sqrt{T} s_T(\theta_0), \tag{III.28}$$

such that by simple manipulations, the likelihood expansion in (III.26) can be written as,

$$\begin{aligned} &2\left(\ell_T(\theta) - \ell_T(\theta_0)\right) && \text{(III.29)} \\ &= -\left(\sqrt{T}(\theta - \theta_0) - Z_T\right)' \left(\frac{1}{T} i_T(\theta_0)\right) \left(\sqrt{T}(\theta - \theta_0) - Z_T\right) + Z_T' \left(\frac{1}{T} i_T(\theta_0)\right) Z_T + \ldots \end{aligned}$$

It thus follows as in Andrews (1999, Theorem 3) that maximizing $\ell_T(\theta)$ over $\theta \in \Theta_C$ is (asymptotically) equivalent to *minimizing* the quadratic form,

$$Q_T\left(\sqrt{T}(\theta-\theta_0)\right) = \left(\sqrt{T}(\theta-\theta_0) - Z_T\right)' \left(\frac{1}{T} i_T(\theta_0)\right) \left(\sqrt{T}(\theta-\theta_0) - Z_T\right).$$

Next observe that, with

$$Q_T(v) = (v - Z_T)' \left(\frac{1}{T} i_T(\theta_0)\right)(v - Z_T),$$

and Θ_C defined in (III.27), by definition

$$Q_T\left(\sqrt{T}\left(\hat{\theta}_C - \theta_0\right)\right) = \inf_{\theta\in\Theta_C} Q_T\left(\sqrt{T}(\theta-\theta_0)\right) = \inf_{\lambda_T\in\Lambda_T} Q_T(\lambda_T)$$

where

$$\Lambda_T = \sqrt{T}(\Theta_C - \theta_0) = \sqrt{T}[\sigma_L^2 - \sigma_0^2, \sigma_U^2 - \sigma_0^2] \times \sqrt{T}[-\alpha_0, \alpha_U - \alpha_0]. \quad \text{(III.30)}$$

Theorem 3 in Andrews (1999) states that under Assumption III.5.1,

$$\sqrt{T}\left(\hat{\theta}_C - \theta_0\right) \xrightarrow{D} \lambda,$$

where λ is the "limit of $\inf_{\lambda_T\in\Lambda_T} Q_T(\lambda_T)$".

As to the "limit of $\inf_{\lambda_T\in\Lambda_T} Q_T(\lambda_T)$" observe first that by the regularity condition (C.2), $\frac{1}{T} i_T(\theta_0) \xrightarrow{P} \Omega_I$ and

$$Z_T = \left(\frac{1}{T} i_T(\theta_0)\right)^{-1} \frac{1}{\sqrt{T}} s_T(\theta_0) \xrightarrow{D} Z \overset{D}{=} \Omega_I^{-1} N_2(0, \Omega_S).$$

Next, with Λ_T defined in (III.30) note that $\sqrt{T}\alpha_0 \to \infty$ if $\alpha_0 > 0$ while $\sqrt{T}\alpha_0 \to 0$ if $\alpha_0 = 0$.

Thus with θ_0 an inner point of Θ, such that $\alpha_0 > 0$

$$\Lambda_T \to \Lambda = \mathbb{R} \times \mathbb{R}.$$

On the other hand, if $\alpha_0 = 0$, then

$$\Lambda_T \to \Lambda = \mathbb{R} \times \mathbb{R}^+.$$

Collecting terms the limit(s) of $\inf_{\lambda_T\in\sqrt{T}(\Theta-\theta_0)} Q_T(\lambda)$ it follows that

$$\lambda = \arg\inf_{v\in\Lambda} (v - Z)' \Omega_I (v - Z),$$

where either $\Lambda = \mathbb{R} \times \mathbb{R}$ or $\Lambda = \mathbb{R} \times \mathbb{R}^+$ depending on whether $\alpha_0 > 0$ or $\alpha_0 = 0$.

We summarize the results in the following theorem:

Theorem III.5.1 *Assume for the ARCH(1) process x_t defined in (I.1) that the sequence z_t i.i.d.(0,1) with $Ez_t^4 < \infty$ and that x_t satisfies the drift criterion, that is $E \log \alpha_0 z_t^2 < 0$ for $\alpha_0 > 0$. With,*

$$\Theta_C = [\sigma_L^2, \sigma_U^2] \times [0, \alpha_U] \subset \Theta$$

it follows for $\theta_0 \in \Theta_C$ that with $\hat{\theta}_C = \arg\max_{\theta \in \Theta_C} \ell_T(\theta)$,

$$\sqrt{T}\left(\hat{\theta}_C - \theta_0\right) \overset{D}{\to} \lambda = \arg\inf_{v \in \Lambda} (v - Z)' \Omega_I (v - Z),$$

where $Z \overset{D}{=} N_2\left(0, \Omega_I^{-1}\Omega_S\Omega_I^{-1}\right)$, and where $\Lambda = \mathbb{R} \times \mathbb{R}$ if $\alpha_0 > 0$, while $\Lambda = \mathbb{R} \times \mathbb{R}^+$ if $\alpha_0 = 0$.

Initially, note that if $\Lambda = \mathbb{R} \times \mathbb{R}$, or θ_0 is an interior point of Θ_C, then $\lambda = Z$ such that

$$\sqrt{T}\left(\hat{\theta}_C - \theta_0\right) \overset{D}{\to} Z \overset{D}{=} N_2\left(0, \Omega_I^{-1}\Omega_S\Omega_I^{-1}\right),$$

and the asymptotic distribution is (as expected) as before for $\hat{\theta} = \arg\max_{\theta \in \Theta_+} \ell_T(\theta)$.

However, if $\alpha_0 = 0$, then $\Lambda = \mathbb{R} \times \mathbb{R}^+$ and the distribution is non-standard as it is non-Gaussian.

To give an idea of the non-standard distribution, consider the MLE in the ARCH model with σ^2 fixed such that $\theta = \alpha$. In this case and with $\alpha_0 = 0$, then $\Lambda = \mathbb{R}^+$, and $Z = N(0, \omega)$ where $\omega = \kappa/(1 + \kappa)$. One finds immediately,

$$\lambda = \arg\inf_{v \in \mathbb{R}^+} (v - Z)' \Omega_I (v - Z) = \begin{cases} Z & \text{if } Z > 0 \\ 0 & \text{if } Z \leq 0 \end{cases} = \max(0, Z). \quad \text{(III.31)}$$

This is an example of a "half-normal" distribution. In general the non-standard distributions are unknown, and are commonly, as in Theorem III.5.1, stated implicitly in terms of the limiting infimum over the relevant set (or, cones) Λ which depends on θ_0.

III.5.3 Testing the hypothesis $\alpha = 0$

The likelihood ratio statistic of the hypothesis of $\alpha = 0$ in the ARCH(1) model is by definition given by,

$$LR = -2\left(\ell_T(\tilde{\theta}) - \ell(\hat{\theta}_C)\right),$$

where $\tilde{\theta}$ is the estimator under the hypothesis, that is with $\Theta_C^0 = [\sigma_L^2, \sigma_U^2] \times \{0\}$,

$$\tilde{\theta} = \arg\max_{\theta \in \Theta_C^0} \ell_T(\theta).$$

By the considerations above, it follows that

$$LR \xrightarrow{D} \mathcal{LR} = \max(0, U)^2,$$

where U is $N(0,1)$ distributed.

Remark III.5.1 *Note that while $U^2 = \chi_1^2$, $\mathcal{LR}$ is not χ^2 distributed as with probability 1/2 it takes the value zero. Therefore $\mathcal{LR}$ is sometimes written as $\frac{1}{2}\chi_1^2$.*

To better understand the result consider again the ARCH(1) model with σ^2 fixed. In this case $\theta = \alpha$, $\tilde{\theta} = 0$, and as for the QMLE expansion, with $U = Z\sqrt{\Omega_I} = N(0,1)$, see also (III.31),

$$\begin{aligned} LR &= -2(\ell_T(0) - \ell(\hat{\alpha}_C)) = Z_T^2 \left(\frac{1}{T} i_T(0)\right) - \left(\sqrt{T}\hat{\alpha}_C - Z_T\right)^2 \left(\frac{1}{T} i_T(0)\right) \\ &\xrightarrow{D} U^2 - (\max(0,U) - U)^2 \\ &= \begin{cases} U^2 & \text{if } U > 0 \\ 0 & \text{if } U \leq 0 \end{cases} = \frac{1}{2}\chi_1^2 \end{aligned}$$

III.5.4 The regularity condition (C.1) : Consistency

The regularity condition (C.1) of consistency of $\hat{\theta}$ in Assumption III.5.1 is a non-trivial condition which needs to be verified. Kristensen and Rahbek (2005, 2009) establishes consistency of the MLE in a wide range of ARCH models, including the ARCH(k) and asymmetric ARCH(k) models. Jeantheau (1998) discusses consistency in multivariate ARCH models, and Berkes, Horváth and Kokoszka (2003) and Francq and Zakoïan (2019) provide general theory on GARCH models.

For the ARCH(1) model, Kristensen and Rahbek (2005) establishes that with the Gaussian log-likelihood,

$$\ell_T(\theta) = -\frac{1}{2} \sum_{t=1}^{T} \left(\log \sigma_t^2(\theta) + x_t^2/\sigma_t^2(\theta)\right),$$

the MLE on (the non-compact Θ),

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \ell_T(\theta)$$

satisfies (C.1), that is $\hat{\theta} \xrightarrow{P} \theta_0$. This is derived under the following conditions: (i) the $i.i.d.(0,1)$ sequence z_t satisfies $Ez_t^2 < \infty$, and (ii) geometric ergodicity, or weakly mixing x_t.

The arguments in Kristensen and Rahbek (2005) hold as well for $\hat{\theta}_C = \arg\max_{\theta \in \Theta_C} \ell_T(\theta)$, and here a brief summary is given of the main key arguments. For more details, see Jeantheau (1998) for a theory on compact parameter sets in general ARCH models, and Kristensen and Rahbek (2005, proof of Theorem 1) for non-compact sets.

The key idea is, similar to most classical analyses, to study the limit $\ell^*(\theta)$ of $\ell_T^*(\theta) = (-2/T)\,\ell_T(\theta)$ and show that the limit has a unique minimum at θ_0.

Consider first point-wise convergence of $\ell_T^*(\theta)$ for a $\theta \in \Theta_C$, that is,

$$\ell_T^*(\theta) = \frac{1}{T}\sum_{t=1}^{T}\left(\log\sigma_t^2(\theta) + x_t^2/\sigma_t^2(\theta)\right) \xrightarrow{P} \ell^*(\theta).$$

By standard application of the LLN, and using $x_t = \sigma_t z_t$, it follows that

$$\ell(\theta) = E\left(\log\sigma_t^2(\theta) + x_t^2/\sigma_t^2(\theta)\right) = E\left(\log\sigma_t^2(\theta) + \sigma_t^2/\sigma_t^2(\theta)\right)$$

Next, note that, using the inequality, $-\log x \geq 1-x$, and $\ell(\theta_0) = E(\log\sigma_t^2(\theta)) + 1$,

$$\ell(\theta) - \ell(\theta_0) = E\left(-\log\left(\sigma_t^2/\sigma_t^2(\theta)\right) + \sigma_t^2/\sigma_t^2(\theta)\right) - 1 \geq 0,$$

with equality if and only if, (with probatility one)

$$\begin{aligned}&\sigma_t^2(\theta) = \sigma_t^2, \text{ or}\\ &\sigma^2 - \sigma_0^2 + (\alpha - \alpha_0)\,x_{t-1}^2 = 0\,.\end{aligned}$$

Now if $\alpha = \alpha_0$, then clearly, $\sigma_0^2 = \sigma^2$. If $\alpha \neq \alpha_0$, the condition reduces to $x_{t-1}^2 = c$ (with probability one) for some constant c. However, it holds by the theory of the drift criterion (the proof of weakly mixing) that the stationary distribution of x_t has a well-defined density with respect to the Lebesgue measure, such that in particular, $P(x_t^2 = c) = 0$ for c a constant. Hence, $\sigma_t^2(\theta) = \sigma_t^2$ implies $\theta = \theta_0$ such that $\ell(\theta)$ attains its minimum in θ_0.

The just given argument was point-wise, that is for one $\theta \in \Theta_C$. By definition,

$$\hat{\theta}_C = \arg\max_{\theta \in \Theta_C} \ell_T(\theta) = \arg\min_{\theta \in \Theta_C} \ell_T^*(\theta),$$

and what is needed is therefore,

$$\arg\min_{\theta \in \Theta_C} \ell_T^*(\theta) \xrightarrow{P} \arg\min_{\theta \in \Theta_C} \ell^*(\theta),$$

which holds provided *uniform* convergence of $\ell_T^*(\theta)$ holds over Θ_C. This again holds provided,

$$E \sup_{\theta\in\Theta_C} \left|\log \sigma_t^2(\theta) + x_t^2/\sigma_t^2(\theta)\right| < \infty,$$

by the *uniform* LLN in Jensen, Lange and Rahbek (2011, Lemma 3), see also Kristensen and Rahbek (2005).

III.6 (Asymmetric) ARCH(k) and GARCH(1,1)

We shall not go through the details of the derivations for the (asymmetric) ARCH(k) and GARCH(1,1) models but instead state results known from the literature. These have been derived using techniques similar to the just described for the ARCH(1) model.

III.6.1 ARCH(k) and Asymmetric ARCH(k)

The linear ARCH(k) model is given by the equations, for $t = k, k+1, ..., T$,

$$\begin{aligned} x_t &= \sigma_t(\theta) z_t \\ \sigma_t^2(\theta) &= \sigma^2 + \alpha_1 x_{t-1}^2 + ... + \alpha_k x_{t-k}^2 \end{aligned} \tag{III.32}$$

with $x_0, ..., x_{k-1}$ fixed and z_t *i.i.d.*(0,1). The parameters to be estimated are given by $\theta = (\sigma^2, \alpha_1, ..., \alpha_k)'$ with $\sigma^2 > 0$ and $\alpha_i \geq 0$ for all $i = 1, ..., k$.

The *asymmetric* (or GJR) ARCH(k) model is for $t = k, k+1, ..., T$, given by

$$\begin{aligned} x_t &= \sigma_t(\theta) z_t \\ \sigma_t^2(\theta) &= \sigma^2 + \alpha_{1n} 1(x_{t-1} < 0) x_{t-1}^2 + \alpha_{1p} 1(x_{t-1} \geq 0) x_{t-1}^2 + ... \\ &\quad + \alpha_{kn} 1(x_{t-k} < 0) x_{t-k}^2 + \alpha_{kp} 1(x_{t-k} \geq 0) x_{t-k}^2 \end{aligned} \tag{III.33}$$

with $x_0, ..., x_{k-1}$ fixed and z_t *i.i.d.*(0,1). The parameters to be estimated are given by $\theta = (\sigma^2, \alpha_{1n}, \alpha_{1p}..., \alpha_{kn}, \alpha_{kp})'$ with $\sigma^2 > 0$ and $\alpha_{in}, \alpha_{ip} \geq 0$ for all $i = 1, ..., k$.

In both cases, the QMLE $\hat{\theta}$ is found by maximizing the Gaussian log-likelihood function,

$$\ell_T(\theta) = \sum_{t=k}^{T} l_t(\theta), \qquad l_t(\theta) = -\tfrac{1}{2}(\log \sigma_t^2(\theta) + \frac{x_t^2}{\sigma_t^2(\theta)}), \tag{III.34}$$

with $\sigma_t^2(\theta)$ defined in (III.32) for the ARCH(k) and in (III.33) for the asymmetric ARCH(k).

Kristensen and Rahbek (2005, 2009) provide results for various variants of the ARCH models including the two mentioned here. It follows that provided z_t is $i.i.d.$(0,1) with $Ez_t^4 < \infty$, and such that x_t is weakly mixing and satisfies a drift criterion, then $\hat{\theta}$ is consistent. Moreover,

$$\sqrt{T}\left(\hat{\theta}-\theta_0\right) \xrightarrow{D} N_d\left(0, \Omega_I^{-1}\Omega_S\Omega_I^{-1}\right), \tag{III.35}$$

where

$$\frac{1}{T}i_T(\theta_0) = -\frac{1}{T}\tfrac{\partial^2}{\partial\theta\partial\theta'}\ell_T(\theta)\,|_{\theta=\theta_0} \xrightarrow{P} \Omega_I.$$
$$\tfrac{1}{\sqrt{T}}s_T(\theta_0) = \frac{1}{\sqrt{T}}\tfrac{\partial}{\partial\theta\partial}\ell_T(\theta)\,|_{\theta=\theta_0} \xrightarrow{D} N_d(0,\Omega_S).$$

Note that like in the ARCH(1) case, if z_t are $i.i.d.N(0,1)$ such that $\hat{\theta}$ is the MLE, then

$$\Omega_I^{-1}\Omega_S\Omega_I^{-1} = \Omega_I^{-1} = 2\Sigma^{-1},$$

with

$$\Sigma = E\left(\frac{w_t w_t'}{\sigma_t^4}\right),$$

where in the ARCH(k) and asymmetric ARCH(k) cases respectively,

$$w_t = \left(1, x_{t-1}^2, ..., x_{t-k}^2\right)' \quad \text{and} \tag{III.36}$$
$$w_t = \left(1, 1\left(x_{t-1}<0\right)x_{t-1}^2, 1\left(x_{t-1}\geq 0\right)x_{t-1}^2, ..., 1\left(x_{t-k}\geq 0\right)x_{t-k}^2\right)'.$$

III.6.2 Consistent estimation of the covariance

The result in (III.35) implies that one can do χ^2 inference using likelihood ratio tests for hypotheses on the parameters in θ, see below.

To compute simple standard $t-$ratios on individual parameters say, one can also provide consistent estimators of the covariance matrix as discussed before for the ARCH(1).

Thus if the z_t are assumed to be $i.i.d.N(0,1)$ then $\hat{\theta}$ is the MLE, and

$$\sqrt{T}\left(\hat{\theta}-\theta_0\right) \xrightarrow{D} N_d\left(0, \Omega_I^{-1}\right),$$

where we as noted can estimate Ω_I consistently by,

$$\frac{1}{T}i_T(\hat{\theta}) = -\frac{1}{T}\tfrac{\partial^2}{\partial\theta\partial\theta'}\ell_T(\theta)\,|_{\theta=\hat{\theta}}.$$

If the z_t are $i.i.d.(0,1)$ but not Gaussian, then $\hat{\theta}$ is the QMLE, and

$$\sqrt{T}\left(\hat{\theta}-\theta_0\right)\xrightarrow{D}N_2\left(0,\Omega_I^{-1}\Omega_S\Omega_I^{-1}\right).$$

Thus in this case we also need a consistent estimator of Ω_S. As before for the ARCH(1) model,

$$\begin{aligned}\frac{1}{\sqrt{T}}s_T\left(\theta\right)=\frac{1}{\sqrt{T}}\tfrac{\partial}{\partial\theta}\ell_T\left(\theta\right)&=\frac{1}{\sqrt{T}}\sum_{t=1}^{T}s_t\left(\theta\right),\\ &=\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\tfrac{1}{2\sigma_t^2(\theta)}(1-\tfrac{x_t^2}{\sigma_t^2(\theta)})w_t,\end{aligned}$$

where w_t are defined in (III.36) and Ω_S is consistently estimated by

$$\hat{\Omega}_S=\frac{1}{T}\sum_{t=1}^{T}s_t(\hat{\theta})s_t(\hat{\theta})'.$$

III.6.3 On weakly mixing

A key assumption for the asymptotic normality of the $\hat{\theta}$ in the ARCH(k) and asymmetric ARCH(k) was the one of weakly mixing or that the drift criterion applies. Thus while we need the drift criterion to apply, we do not need any moments for the ARCH process x_t. Hence it would be natural to find a minimal condition for this, in line with the ARCH(1) where the condition is $E\log\left(\alpha_0 z_t^2\right)<0$ or $\alpha_0<3.56$. While it can be done, this is not so simple for the ARCH(k) case, so we will restrict attention to the case where $Ex_t^2<\infty$ in which case the results are simple to state.

Tedious calculations show that similar to the discussion in Part II about the ARCH(2) process, the companion form of the ARCH(k) process,

$$X_t=(x_t,...,x_{t-k+1})'$$

is weakly mixing with drift function $\delta\left(X\right)=1+|X|^2$ and hence finite second order moments, $E|X_t|^2=EX_t'X_t=E\left(x_t^2+...+x_{t-k+1}^2\right)<\infty$ if,

$$\alpha_1+\alpha_2+...+\alpha_k<1. \tag{III.37}$$

Likewise for the asymmetric ARCH(k),

$$X_t=(x_t,...,x_{t-k+1})'$$

is weakly mixing with finite second order moments, $E|X_t|^2=EX_t'X_t=E\left(x_t^2+...+x_{t-k+1}^2\right)<\infty$ and hence $Ex_t^2<\infty$ if,

$$\max\left(\alpha_{1n},\alpha_{1p}\right)+...+\max\left(\alpha_{kn},\alpha_{kp}\right)<1, \tag{III.38}$$

see e.g. Kristensen and Rahbek (2009) for details.

III.6.4 GARCH(1,1)

Consider the GARCH(1,1) model, where for $t = 1, 2, 3, ..., T$

$$\begin{aligned} x_t &= \sigma_t(\theta) z_t \\ \sigma_t^2(\theta) &= \sigma^2 + \alpha x_{t-1}^2 + \beta \sigma_{t-1}^2(\theta), \end{aligned}$$

z_t *i.i.d.*N(0,1), x_0 and $\sigma_0^2(\theta)$ are fixed and the parameters to be estimated are given by $\theta = (\sigma^2, \alpha, \beta)$, where $\sigma^2 > 0$ and $\alpha, \beta \geq 0$.

Fixing, or conditioning on, the *observed* initial value x_0 and the *unobserved* initial value $\sigma_0^2(\theta)$, the Gaussian log-likelihood function is as for the ARCH case given by,

$$\ell_T(\theta) = \sum_{t=1}^{T} l_t(\theta) = -\tfrac{1}{2} \sum_{t=1}^{T} (\log \sigma_t^2(\theta) + \frac{x_t^2}{\sigma_t^2(\theta)}). \qquad \text{(III.39)}$$

In practice, often the unobserved $\sigma_0^2(\theta)$ is set equal to the sample variance of x_t, that is

$$\sigma_0^2(\theta) = \frac{1}{T} \sum_{t=1}^{T} x_t^2.$$

The asymptotic distribution of $\hat{\theta} = \left(\hat{\sigma}^2, \hat{\alpha}, \hat{\beta}\right)$ is derived in for example Berkes et.al (2003) using similar arguments as for the ARCH(1) model, see also Jensen and Rahbek (2004b). The main conclusion is that provided *(i)* $Ez_t^4 < \infty$, *(ii)* the process $(x_t, \sigma_t^2)'$ is weakly mixing, and that *(iii)* $\alpha_0, \beta_0 > 0$ hold, then asymptotic normality holds,

$$\sqrt{T}\left(\hat{\theta} - \theta\right) \xrightarrow{D} N_3\left(0, \Omega_I^{-1} \Omega_S \Omega_I^{-1}\right),$$

with $\Omega_S = \Omega_I$ if the z_t are *i.i.d.*N$(0, 1)$.

III.6.4.1 Consistent estimation of the covariance

With the GARCH(1,1) likelihood function given in (III.39), then as before we can estimate Ω_I consistently by,

$$\frac{1}{T} i_T\left(\hat{\theta}\right) = -\frac{1}{T} \tfrac{\partial^2}{\partial\theta\partial\theta'} \ell_T(\theta)\,|_{\theta=\hat{\theta}}.$$

and Ω_S consistently by

$$\hat{\Omega}_S = \frac{1}{T} \sum_{t=1}^{T} s_t(\hat{\theta}) s_t(\hat{\theta})', \text{ where } s_t = \tfrac{\partial}{\partial\theta} l_t(\theta).$$

Note in this respect that the derivatives, while still simple, are more complicated than in the ARCH case. For example,

$$\begin{aligned}\tfrac{\partial}{\partial\theta}l_t(\theta) &= \tfrac{\partial}{\partial\theta}\left(-\frac{1}{2}(\log\sigma_t^2(\theta)+\frac{x_t^2}{\sigma_t^2(\theta)})\right)\\ &= \frac{1}{2}\left(\frac{x_t^2}{\sigma_t^2(\theta)}-1\right)\frac{\tfrac{\partial}{\partial\theta}\sigma_t^2(\theta)}{\sigma_t^2(\theta)}.\end{aligned}$$

With $\theta = (\sigma^2, \alpha, \beta)$ we find,

$$\begin{aligned}\tfrac{\partial}{\partial\sigma^2}\sigma_t^2(\theta) &= 1+\beta\tfrac{\partial}{\partial\sigma^2}\sigma_{t-1}^2(\theta)\\ \tfrac{\partial}{\partial\alpha}\sigma_t^2(\theta) &= x_{t-1}^2+\beta\tfrac{\partial}{\partial\alpha}\sigma_{t-1}^2(\theta)\\ \tfrac{\partial}{\partial\beta}\sigma_t^2(\theta) &= \sigma_{t-1}^2(\theta)+\beta\tfrac{\partial}{\partial\sigma^2}\sigma_{t-1}^2(\theta).\end{aligned}$$

Thus we see that all derivatives are given by recursions[3].

III.6.4.2 On weakly mixing

Recall from Part II, that if

$$E\log\left(\beta_0+\alpha_0 z_t^2\right)<0,$$

then the GARCH(1,1) process is weakly mixing. This implies $\beta_0 < 1$ and that $\alpha_0 + \beta_0 = 1$ are included. And thus the regularity conditions discussed do not conflict with $\hat{\alpha} + \hat{\beta} = 1$ (approximately) as found in many GARCH analyses.

Remark III.6.1 *Note that if the hypothesis $\alpha = 0$ is considered for the GARCH(1,1) model, then β is not identified since in this case*

$$\sigma_t^2 = \omega + \beta\sigma_{t-1}^2 = \omega/(1-\beta),$$

if σ_0^2 is initiated at $\sigma_0^2 = \omega/(1-\beta)$, $\beta < 1$. This means that testing for "no GARCH" effects in GARCH models is not a simple task.

[3] With $\sigma_0^2(\theta)$ fixed, then $\tfrac{\partial}{\partial\theta}\sigma_0^2(\theta) = 0$, and the recursions give directly,

$$\begin{aligned}\tfrac{\partial}{\partial\sigma^2}\sigma_t^2(\theta) &= 1+\beta\tfrac{\partial}{\partial\sigma^2}\sigma_{t-1}^2(\theta) = \sum_{i=0}^{t-1}\beta^i\\ \tfrac{\partial}{\partial\alpha}\sigma_t^2(\theta) &= x_{t-1}^2+\beta\tfrac{\partial}{\partial\alpha}\sigma_{t-1}^2(\theta) = \sum_{i=0}^{t-1}\beta^i x_{t-1-i}^2\\ \tfrac{\partial}{\partial\beta}\sigma_t^2(\theta) &= \sigma_{t-1}^2(\theta)+\beta\tfrac{\partial}{\partial\sigma^2}\sigma_{t-1}^2(\theta) = \sum_{i=0}^{t-1}\beta^i\sigma_{t-1-i}^2(\theta)\end{aligned}$$

Appendix

A Proof of Theorem III.3.1

Some notation: Define the normalized version of the log-likelihood function,

$$\tilde{\ell}_T(\theta) = -\frac{2}{T}\ell_T(\theta), \tag{III.40}$$

such that

$$\begin{aligned} D\tilde{\ell}_T(\theta) &\equiv \partial\tilde{\ell}_T(\theta)/\partial\theta = -\frac{2}{T}\partial\ell_T(\theta)/\partial\theta, \\ D^2\tilde{\ell}_T(\theta) &\equiv \partial^2\tilde{\ell}_T(\theta)/\partial\theta\partial\theta' = -\frac{2}{T}\partial^2\ell_T(\theta)/\partial\theta\partial\theta'. \end{aligned}$$

Set furthermore,

$$\tilde{\Omega}_I \equiv 2\Omega_I \quad \text{and} \quad \tilde{\Omega}_S = 4\Omega_S.$$

An initial result: Note first that by (A.3), it follows that for any vectors $v_1, v_2 \in \mathbb{R}^k$, and any $\theta \in N(\theta_0)$,

$$\left| v_1' \left(D^2\tilde{\ell}_T(\theta) - D^2\tilde{\ell}_T(\theta_0) \right) v_2 \right| \leq \|v_1\| \, \|v_2\| \, \|\theta - \theta_0\| \, \tilde{c}_T, \tag{III.41}$$

where $\tilde{c}_T = 2k^{3/2}c_T$.

To see this, note that the l.h.s of (III.41) is $|f(1) - f(0)| = |\partial f(\lambda^*)/\partial\lambda|$ for some $0 \leq \lambda^* \leq 1$, where, $f(\lambda) = v_1' \left[D^2\tilde{\ell}_T\left(\theta_0 - \lambda\left(\theta - \theta_0\right)\right) \right] v_2$, $0 \leq \lambda \leq 1$. By Taylors formula and (A.3),

$$\begin{aligned} &|\partial f(\lambda^*)/\partial\lambda| = \left| \textstyle\sum_{i,j,l=1}^k v_{1,i}v_{2,j}(\theta_l - \theta_{0,l})\partial^3\tilde{\ell}_T(\theta_0 - \lambda^*(\theta - \theta_0)/\partial\theta_i\partial\theta_j\partial\theta_l \right| \\ &\leq 2c_T \textstyle\sum_{i=1}^k |v_{1,i}| \sum_{j=1}^k |v_{2,j}| \sum_{l=1}^k |\theta_l - \theta_{0,l}| \leq \tilde{c}_T\|v_1\|\|v_2\|\|\theta - \theta_0\|. \end{aligned}$$

Existence and uniqueness of $\hat{\theta}$:

Next, by definition the continuous function $\tilde{\ell}_T(\theta)$ attains its minimum in any compact neighborhood $K(\theta_0, r) = \{\theta | \ \|\theta - \theta_0\| \leq r\} \subseteq N(\theta_0)$ of θ_0. We proceed by showing that with a probability tending to one as $T \to \infty$, $\tilde{\ell}_T(\theta)$ cannot obtain its minumum on the boundary of $K\left(\theta_0, r\right)$ and that $\tilde{\ell}_T(\theta)$ is convex in the interior of $K\left(\theta_0, r\right)$, $intK\left(\theta_0, r\right)$.

With $v_\theta = (\theta - \theta_0)$, and θ^* on the line from θ to θ_0, Taylors formula gives,

$$\begin{aligned} &\tilde{\ell}_T(\theta) - \tilde{\ell}_T(\theta_0) = D\tilde{\ell}_T(\theta_0)v_\theta + \tfrac{1}{2}v_\theta' D^2\tilde{\ell}_T(\theta^*)v_\theta = \qquad \text{(III.42)} \\ &D\tilde{\ell}_T(\theta_0)v_\theta + \tfrac{1}{2}v_\theta' \left[\tilde{\Omega}_I + (D^2\tilde{\ell}_T(\theta_0) - \tilde{\Omega}_I) + (D^2\tilde{\ell}_T(\theta^*) - D^2\tilde{\ell}_T(\theta_0)) \right] v_\theta. \end{aligned}$$

Denote by ρ_T and ρ, $\rho > 0$, the smallest eigenvalues of $\left[D^2\tilde{\ell}_T(\theta_0) - \tilde{\Omega}_I\right]$ and $\tilde{\Omega}_I$ respectively. Note that $\rho_T \xrightarrow{P} 0$ by (A.2) and the fact that the smallest eigenvalue of a $k \times k$ symmetric matrix M, $\inf_{\{v \in \mathbb{R}^k | \|v\|=1\}} v'Mv$ is continuous in M.

Then (A.1) and (A.3), with $\tilde{c} = 2k^{3/2}c$, and the uniform upper bound in (III.41) imply that

$$\inf_{\theta:||v_\theta||=r}[\tilde{\ell}_T(\theta)-\tilde{\ell}_T(\theta_0)] \geq -\|D\tilde{\ell}_T(\theta_0)\|r+\tfrac{1}{2}\left[\rho+\rho_T-\tilde{c}_T r\right]r^2 \xrightarrow{P} \tfrac{1}{2}\left[\rho-\tilde{c}r\right]r^2 \equiv \eta.$$

Hence, if $r < \rho/\tilde{c}$ then $\inf_{\theta:||v_\theta||=r}[\tilde{\ell}_T(\theta) - \tilde{\ell}_T(\theta_0)] \geq \eta > 0$ with probability tending to one. As $\tilde{\ell}_T(\theta)|_{\theta=\theta_0} - \tilde{\ell}_T(\theta_0) = 0$, this implies that the probability that $\tilde{\ell}_T(\theta)$ attains its minimum on the boundary of $K(\theta_0, r)$ tends to zero.

Next, for $\theta \in K(\theta_0, r)$ and $v \in \mathbb{R}^k$, rewriting $v'D^2\tilde{\ell}_T(\theta)v$ as in (III.42),

$$\begin{aligned} v'D^2\tilde{\ell}_T(\theta)v &= v'\left[\tilde{\Omega}_I + (D^2\tilde{\ell}_T(\theta_0) - \tilde{\Omega}_I) + (D^2\tilde{\ell}_T(\theta) - D^2\tilde{\ell}_T(\theta_0))\right] \\ &\geq \|v\|^2(\rho + \rho_T - r\tilde{c}_T) \xrightarrow{P} \|v\|^2(\rho - r\tilde{c}). \end{aligned}$$

Hence, if $r < \rho/\tilde{c}$ the probability that $\tilde{\ell}_T(\theta)$ is strongly convex in the interior of $K(\theta_0, r)$ tends to 1, and therefore it has at most one stationary point. This establishes (B.1): If $r < \rho/\tilde{c}$ and $K(\theta_0, r) \subseteq N(\theta_0)$, there is with a probability tending to 1 exactly one solution $\hat{\theta}$ to the likelihood equation in the interior $U(\theta_0) = intK(\theta_0, r)$. It is the unique minimum point of $\tilde{\ell}_T(\theta)$ in $U(\theta_0)$ and, as it is a stationary point, it solves $D\tilde{\ell}_T(\theta) = 0$, and hence also $D\ell_T(\theta) = 0$.

Establishing consistency:

By the same argument, for any δ, $0 < \delta < r$ there is with a probability tending to 1 a solution to the likelihood equation in $K(\theta_0, \delta)$. As $\hat{\theta}$ is the unique solution to the likelihood equation in $K(\theta_0, r)$, it must therefore be in $K(\theta_0, \delta)$ with a probability tending to 1. Hence we have proved that θ_T is consistent. That is, for any $0 < \delta < r$, the probability that $\hat{\theta}$ is a unique solution to $D\tilde{\ell}_T(\hat{\theta}) = 0$ in $K(\theta_0, r)$ and $\|\hat{\theta} - \theta_0\| \leq \delta$ tends to one, which establishes the needed.

Asymptotic normality:

That $\hat{\theta}$ is asymptotically Gaussian follows from (A.1) and by Taylors formula for the functions $\partial\tilde{\ell}_T(\theta)/\partial\theta_j, j = 1, \ldots, k$:

$$\sqrt{T}D\tilde{\ell}_T(\theta_0) = (\tilde{\Omega}_I + A_T(\hat{\theta}))\sqrt{T}(\hat{\theta} - \theta_0). \tag{III.43}$$

Here the elements in the matrix $A_T(\hat{\theta})$ are of the form $v_1'(D^2\tilde{\ell}_T(\theta_T^*) - 2\Omega_I)v_2$ with v_1, v_2 unit vectors in $\mathbb{R}^k$ and θ_T^* a point on the line from θ_0 to $\hat{\theta}$. Note that θ_T^* depends on the first vector v_1. Next, by (III.41),

$$|v_1'(D^2\tilde{\ell}_T(\theta_T^*) - \Omega_I)v_2| \leq |v_1'(D^2\tilde{\ell}_T(\theta_0) - \Omega_I)v_2| + \|v_1\|\|v_2\|\|\theta_T^* - \theta_0\|\tilde{c}_T.$$

Since $\theta_T^* \xrightarrow{P} \theta_0$ and $\tilde{c}_T \xrightarrow{P} \tilde{c} < \infty$ it follows from (A.2) that the right hand side tends in probability to 0. Hence $A_T(\theta_T) \xrightarrow{P} 0$ and using (A.1), (III.43) gives

$$\sqrt{T}(\hat{\theta} - \theta_0) \xrightarrow{D} N\left(0, \tilde{\Omega}_I^{-1}\tilde{\Omega}_S\tilde{\Omega}_I^{-1}\right) = N_d\left(0, \Omega_I^{-1}\Omega_S\Omega_I^{-1}\right)$$

showing the needed. □

B Convergence of Newton-Raphson algorithm

The result on convergence of the Newton-Raphson algorithm is here formulated in terms of any differentiable likelihood function $\ell_T(\theta)$ for some time series model for observations $X_0, X_1, ..., X_T$ and where,

$$\hat{\theta} = \arg\max_{\theta\in\Theta} \ell_T(\theta) \ .$$

Clearly $\ell_T(\theta)$ is a function of θ, and so are derivatives of $\ell_T(\theta)$ as for example the score and information, $s_T(\theta)$ and $i_T(\theta)$.

Recall initially the concept of a neigborhood around θ_0, $N(\theta_0)$. With the score defined in Example III.2.2 for the ARCH(1) model,

$$s_T(\theta) = -\tfrac{1}{2}\sum_{t=1}^{T} \tfrac{1}{\sigma_t^2(\theta)}(1 - \tfrac{x_t^2}{\sigma_t^2(\theta)})w_t, \quad \text{with } \sigma_t^2(\theta) = \sigma^2 + \alpha x_{t-1}^2,$$

with $w_t = \left(1, x_{t-1}^2\right)'$, we see that if we let σ^2 and α take different values in some intervals, $[\sigma_L^2, \sigma_U^2]$ and $[\alpha_L, \alpha_U]$ say, then $s_T(\theta)$ will likewise vary in value.

With $\theta = (\theta_1, ..., \theta_d)'$ also recall the notation,

$$\frac{\partial^3\ell_T(\theta)}{\partial\theta_h\partial\theta_i\partial\theta_j}, \quad \text{for } h, i, j = 1, 2, ..., d$$

for third order derivatives. For example $\partial^3\ell_T(\theta)/\partial\theta_1\partial\theta_2\partial\theta_2$ in the ARCH(1) case means

$$\frac{\partial^3\ell_T(\theta)}{\partial\sigma^2\partial\alpha\partial\alpha}.$$

We are now in position to state the convergence result:

Theorem B.1 *Consider the log-likelihood function $\ell_T(\theta)$, which is a function of the observations $X_1, ..., X_T$ and the parameter $\theta \in \Theta$, where Θ is a subset of $\mathbb{R}^k$. Assume that $\ell_T(\theta)$ is three times differentiable in θ with all derivatives continuous. Assume that θ_0 is inside Θ, and that we have for the Newton-Raphson iterations in $\hat{\theta}_n$ that:*

(i) *Initial Estimator:* $\theta^* \xrightarrow{P} \theta_0$, *and* $\sqrt{T}\,(\theta^* - \theta_0) \xrightarrow{D} N\,(0, \Sigma_*)$

(ii) *Information:* $\frac{1}{T} i_T\,(\theta_0) \xrightarrow{P} \Omega_I > 0.$

(iii) *Third Derivatives:* $\max_{h,i,j=1,\dots,d} \sup_{\theta \in N(\theta_0)} \left| \frac{1}{T} \frac{\partial^3 \ell_T(\theta)}{\partial\theta_h \partial\theta_i \partial\theta_j} \right| \leq c_T \xrightarrow{P} c,$

(iv) *QML Estimator* $\sqrt{T}\left(\hat{\theta} - \theta_0\right) \xrightarrow{D} N_k\,(0, \Sigma)$

with $c > 0$. Then for $n = 1, 2, ...$ and for some small δ, $\delta > 0$,

$$T^{n-\delta} |\hat{\theta} - \hat{\theta}_n| \xrightarrow{P} 0, \tag{III.44}$$

as $T \to \infty$.

The result in (III.44) means that for each iteration n, we get closer to $\hat{\theta}$. That is, $|\hat{\theta} - \hat{\theta}_n|$ tends to zero (in probability) as T tends to infinity. Moreover, the speed by which $|\hat{\theta} - \hat{\theta}_n|$ tends to zero is increasing in the iterations n so if it converges, it converges rapidly.

Example B.1 *In terms of the AR(1) model in Example III.2.2, we know from Example III.2.1 that $i_T\,(\theta) = \sum_{t=1}^{T} x_{t-1}^2 / \sigma^2$, and hence if $|\rho_0| < 1$,*

$$\frac{1}{T} i_T\,(\theta_0) = \frac{1}{T} \sum_{t=1}^{T} x_{t-1}^2 / \sigma^2 \xrightarrow{P} E x_{t-1}^2 / \sigma^2 = 1/\left(1 - \rho_0^2\right) > 0,$$

such that (ii) holds in Theorem B.1. Also $\frac{\partial^3 \ell_T(\theta)}{\partial \rho^3} = 0$ such that (iii) trivially holds. This is again reflecting that in case the likelihood function is quadratic optimization is straightforward.

References

Andrews, D., 1999, Estimation When a Parameter Is on a Boundary: Theory and Applications, Econometrica.

Basawa, I.V., Feigin, P.D. and Heyde, C.C.,1976, Asymptotic Properties of Maximum Likelihood Estimators for Stochastic Processes. Sankya Series A 38, 259-270.

Berkes, I., L. Horváth and P. Kokoszka, 2003, GARCH processes: Structure and Estimation, Bernoulli, 201–227.

Billingsley., P., 1961, Statistical Inference for Markov Processes, University of Chicago Press

Francq, C. and J.-M. Zakoïan, 2019, GARCH Models: Structure, Statistical Inference and Financial Applications, Wiley.

Jeantheu, T., 1998, Strong Consistency of Estimators for Multivariate ARCH Models, Econometric Theory.

Jensen, S.T. and A. Rahbek, 2004a, Non-stationary and No Moments Asymptotics for the ARCH Model, Econometrica.

Jensen, S.T. and A. Rahbek, 2004b, Asymptotic Normality for Non-Stationary, Explosive GARCH, Econometric Theory, 20:6:1203-1226.

Kristensen, D. and A. Rahbek, 2005, Asymptotics of the QMLE for a Class of ARCH(q) Models, Econometric Theory.

Kristensen, D. and A. Rahbek, 2009, Asymptotics of the QMLE for Non-Linear ARCH Models, Journal of Time Series Econometrics.

Lange, T., S.T.Jensen and A. Rahbek, 2011, Estimation and Asymptotic Inference in the AR-ARCH Model, Econometric Reviews.

Ling, S., 2006, Self-Weighted and Local Quasi-Maximum Likelihood Estimators for ARMA-GARCH/IGARCH Models, Journal of Econometrics.

Silvapulle, M.J. and P. Sen, 2005, Constrained Statistical Inference: Inequality, Order, and Shape Restrictions, Wiley.

Weiss, A., 1986, Asymptotic Theory for ARCH models, Econometric Theory.