Anders Rahbek
Rasmus Søndergaard Petersen
University of Copenhagen

September 2022

Part I
ARCH(1) | An introduction to the theory of non-linear time series.

I.1 Introduction

In order to discuss econometric modeling of (conditional) time-varying volatility, or variance, this part introduces key ideas and theoretical probabilistic properties of the Autoregressive Conditional Heteroscedastic (ARCH) process of order one, ARCH(1).

The ARCH(1) is the simplest example of the rich class of ARCH models, and their generalized versions, the so-called GARCH models, which are dominating in empirical studies of (conditional) time-varying volatility in financial time series data. While the ARCH(1) process at a first glance appears simple, the probability analysis, as well as the statistical, or econometric analysis, of the ARCH(1) model, require general concepts from recent theory of non-linear models in statistics and econometrics.

The probability theory needed is introduced here such that it can also be applied in the context of the later studies and applications of the vast class of general non-linear ARCH volatility models and hence not alone for the ARCH(1).

To present the idea of modelling time-varying volatility, recall that in the option pricing literature a key assumption of the Black-Scholes model is that spot prices follow a geometric Brownian motion with constant drift and variance. That is, with $t = 1, 2, 3, ..$ denoting time and S_t the price of the underlying asset, log returns $r_t = \log(S_t) - \log(S_{t-1})$ are given by the equation,

$$r_t = \mu + \varepsilon_t,$$

where μ is the drift parameter and ε_t is identically and independently distributed (*i.i.d.*) following a Gaussian distribution with mean zero and variance

σ^2. Here σ (or, sometimes confusingly, σ^2) is commonly referred to as the 'volatility' which is, as is well-known, essential in option pricing and elsewhere. In this simple set-up of the classic Black-Scholes model there are two assumptions which are known from empirical analyses not to hold: *(i)* the (conditional) variance σ^2 is constant over time, and, *(ii)* log returns r_t are Gaussian distributed.

The class of ARCH processes, are processes for which the conditional volatility is allowed to be stochastic and time dependent - with periods of high conditional volatility and periods with low conditional volatility. Moreover, the tails of an ARCH-type process are more thick ('fat tails') than those of the Gaussian distribution. Thus the ARCH process seems a natural starting point for theoretical matching of empirically observed features of the data. Another typical feature of daily and weekly observed financial data series such as returns, is that the data in levels appear to be uncorrelated but not independent as the absolute value or the squared data series often appear empirically to be correlated. Again this is also reflected in the theoretical properties of the ARCH process.

Introductions to modelling issues of (G)ARCH models appear several places, see for example Taylor (2005) and Franses and van Dijk (2000, ch.3-4). A comparison of ARCH and other types of stochastic volatility models can be found in Shephard (1996). General overviews of the rich ARCH modelling literature are also found in e.g. Bollerslev, Chou and Kroner (1992), Bera and Higgins (1992) and Pagan (1996). Finally, the textbook by Francq and Zakoïan (2019) provides a rigorous mathematical statistical introduction to the theory of (G)ARCH models.

I.1.1 The ARCH(1) process

Consider initially the classic ARCH(1) model for (log returns) x_t, which for $t = 1, 2, \ldots$ can be represented as

$$x_t = \sigma_t z_t \tag{I.1}$$
$$\sigma_t^2 = \sigma^2 + \alpha x_{t-1}^2 \tag{I.2}$$

with initial value x_0 and where the z_t's are *i.i.d.* N(0,1). This way the distribution of x_t conditional on past information up to time $t-1$ as given by the variables $(x_{t-1}, ..., x_0)$ depends only on x_{t-1}. This is commonly referred to as the Markov property.

Moreover, the equations (I.1)-(I.2) imply that x_t conditionally on $(x_{t-1}, ..., x_0)$ is Gaussian distributed with conditional variance σ_t^2. However, *importantly, this does not imply that x_t are Gaussian distributed unconditionally.* Instead

the marginal, or unconditional distribution of x_t is non-Gaussian, and – under regularity conditions discussed below – has in particular a more "fat tailed" distribution and can take "larger values" than expected if it was a Gaussian distribution.

The level parameter σ^2 is strictly positive, $\sigma^2 > 0$ while $\alpha \geq 0$. Specifically if $\alpha = 0$ then x_t is simply an *i.i.d.* N(0,σ^2) sequence (conditionally and unconditionally).

It should be emphasized that in particular σ_t^2 is non-constant and stochastic.

The probabilistic behavior of the ARCH process x_t was until recently not fully described as a full discussion of for example the concept *stationarity* and *existence of moments* of x_t demands rather technical analysis. As indicated, the theory will be presented in such a way that other types of ARCH processes can be handled.

Specifically, using non-linear time series theory, we demonstrate below that while the ARCH sequence x_t is *uncorrelated,* it is *dependent.* Moreover, we will derive a simple restriction on the parameters for which x_t is well-behaved process in the sense that it is *stationary* and *asymptotically independent (weakly mixing)* and hence satisfying the *law of large numbers* – which again is important for a discussion of estimation and inference in ARCH and other econometric volatility models.

The concepts *stationarity, asymptotic independence* and *the law of large numbers* will be given precise meaning and definitions in the following sections.

Note that the specification of σ_t^2 is referred to as the linear ARCH(1) model as it is linear in x_{t-1}^2 and depends only on one lag of x_t. The linearity assumption is empirically questionable as for example responses to changes in x_{t-1}^2 often appear nonlinear and asymmetric. In the above mentioned surveys of ARCH modelling much more general functional forms of σ_t^2 are introduced . We shall return to some of these and the ideas behind them later but introduce the theory in terms of the linear ARCH process, simply referred to as the ARCH process henceforth unless otherwise stated. In addition to the surveys, classic references for ARCH models are Engle (1982) and Nelson (1990).

I.2 Conditional moments

Here some key conditional and unconditional moments of the ARCH process are derived. Conditional expectations play a key role in these considerations and will therefore be discussed briefly. Throughout this section we will work

under the assumption that all moments, conditional as well as unconditional, are well-defined such that the calculations are valid. Note that if x_t is *i.i.d.* Gaussian, then all moments of x_t are by definition finite, that is,

$$E|x_t|^k < \infty,$$

for any $k \geq 0$. As will be demonstrated this is *not* the case for non-Gaussian distributions such as for example the ARCH process. In the next section we do establish under which assumptions on the two parameters σ^2 and α the different moments are well-defined and finite.

First we need to give meaning to for example the conditional expectation $E(x_t|x_{t-1})$ and be able to work with conditional expectations in general.

I.2.1 Conditional expectations

Consider in general two random, or equivalently stochastic, variables X and Y, with X having finite expectation $E|X| < \infty$. We shall work with real valued stochastic variables, $X \in \mathbb{R}$, and vector valued, $X \in \mathbb{R}^p$ for some $p \geq 1$ and likewise for Y. We shall furthermore work under the assumption that X has density $f(x)$, Y has density $f(y)$ and the conditional distribution of X given Y has density $f(x|y)$.

In terms of the density, recall that the expectation of X, $X \in \mathbb{R}$ or $X \in \mathbb{R}^p$, can be computed as

$$E(X) = \int_{\mathbb{R}^p} xf(x)\,dx. \tag{I.3}$$

Likewise, we define the conditional expectation of X given $Y = y$, denoted $E(X|Y = y)$, by

$$E(X|Y = y) = \int_{\mathbb{R}^p} xf(x|y)\,dx, \tag{I.4}$$

which, by definition, depends on the value y.

Example I.2.1 *If X is $N(\mu_x, \sigma_x^2)$ distributed, then the density of X is given by,*

$$f(x) = \frac{1}{\sqrt{2\pi\sigma_x^2}} \exp\left(-\frac{1}{2\sigma_x^2}(x-\mu_x)^2\right),$$

and from probability analysis, $E(X) = \mu_x = \int_{\mathbb{R}} xf(x)\,dx$.

Example I.2.2 *Consider $(X,Y)'$ which is bivariate $N_2(\mu,\Omega)$ distributed, with mean $\mu = E(X,Y)'$ and variance matrix $\Omega = Var(X,Y)'$ given by,*

$$\mu = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \ \Omega = \begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{yx} & \sigma_y^2 \end{pmatrix}.$$

In particular, X is $N(\mu_x, \sigma_x^2)$ and Y is $N\left(\mu_y, \sigma_y^2\right)$ distributed, and $Cov(X,Y) = \sigma_{xy}$.

Recall the result that the conditional expectation of X given $Y = y$, $E(X|Y=y)$, is given by

$$E(X|Y=y) = \mu_x + \omega(y - \mu_y) = \mu_{x|y}, \quad \text{where } \omega = \sigma_{xy}/\sigma_y^2. \tag{I.5}$$

In fact, the conditional distribution of X given $Y = y$ is $N\left(\mu_{x|y}, \sigma_{x|y}^2\right)$ distributed, with

$$\sigma_{x|y}^2 = \sigma_{xx}^2 - \omega\sigma_{yx},$$

and hence the conditional density is given by

$$\begin{aligned} f(x|y) &= \frac{1}{\sqrt{2\pi\sigma_{x|y}^2}} \exp\left(-\frac{1}{2\sigma_{x|y}^2}\left(x - \mu_{x|y}\right)^2\right) \\ &= \frac{1}{\sqrt{2\pi\sigma_{x|y}^2}} \exp\left(-\frac{1}{2\sigma_{x|y}^2}\left(x - \mu_x - \omega\left(y - \mu_y\right)\right)^2\right). \end{aligned}$$

Thus as illustrated in Example I.2.2, by defining the conditional expectation $E(X|Y=y)$ as in (I.4), for each value of y we get a different value. In general, we wish to use the conditional expectation for all possible values of Y, that is to treat the conditional expectation as a random variable which we call $E(X|Y)$. With $g(y) = E(X|Y=y)$ defined in (I.4), this is accomplished by simply setting,

$$E(X|Y) = g(Y). \tag{I.6}$$

This is a function of Y and therefore a random variable and defines $E(X|Y)$.

Example I.2.3 *In terms of Example I.2.2, then*

$$E(X|Y) = \mu_x + \omega(Y - \mu_y).$$

This is a random variable (it is a function of Y), and moreover it has the same expectation as X since

$$E(E(X|Y)) = E(\mu_x + \omega(Y - \mu_y)) = \mu_x + \omega(E(Y) - \mu_y) = \mu_x = E(X). \tag{I.7}$$

The fact that $E(E(X|Y)) = E(X)$ in the above example is a general feature of the conditional expectation, often referred to as the law of iterated

expectations. This is a much used implication of the defintion of conditional expectations. Before listing some rules much useful for calulations with conditional expectations we note that if the joint distribution of X and Y has density $f(x,y)$, the conditional density of X given Y is given by

$$f(x|y) = f(x,y)/f(y). \tag{I.8}$$

This corresponds to the well-known result that

$$P(X \in A|Y \in B) = P(X \in A, Y \in B)/P(Y \in B).$$

Lemma I.2.1 *Consider the random variables X, Y and Z with joint density $f(x,y,z)$ and finite expectation. For the conditional expectation $E(X|Y)$ it holds that $E|E(X|Y)| < \infty$ and the law of iterated expectations apply,*

$$E(X) = E(E(X|Y)). \tag{I.9}$$

If X and Y are independent,

$$E(X|Y) = E(X). \tag{I.10}$$

Moreover,

$$E(X|Y) = E(E(X|Y,Z)|Y) \qquad \text{and} \tag{I.11}$$
$$E(X|X) = X \tag{I.12}$$

Generally, with g and h functions such that $g(Y)$ and $h(Y)$ take values in $\mathbb{R}$,

$$E(g(Y) + h(Y)X|Y) = g(Y) + h(Y)E(X|Y). \tag{I.13}$$

Example I.2.4 *Consider the AR process given by,*

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

for $t = 1, 2, ..., T$, initial value x_0 and where ε_t is i.i.d.N(0,σ^2). Then

$$E(x_t|x_{t-1}) = E(\rho x_{t-1} + \varepsilon_t|x_{t-1}) = \rho E(x_{t-1}|x_{t-1}) + E(\varepsilon_t) = \rho x_{t-1},$$

where we have used (I.13), (I.12) and (I.10).

Example I.2.5 *Consider the ARCH process x_t in (I.1), then using first (I.13) and next (I.10) we find that the conditional expectation of x_t is zero as,*

$$\begin{aligned} E\left(x_t|x_{t-1}\right) &= E\left(\sqrt{\left[\sigma^2+\alpha x_{t-1}^2\right]}z_t|x_{t-1}\right) \\ &= \sqrt{\left[\sigma^2+\alpha x_{t-1}^2\right]}E\left(z_t|x_{t-1}\right) \\ &= \sqrt{\left[\sigma^2+\alpha x_{t-1}^2\right]}E\left(z_t\right) = 0. \end{aligned}$$

Hence the ARCH(1) process has mean zero as $E\left(x_t\right) = E\left(E\left(x_t|x_{t-1}\right)\right) = 0$ by (I.9).

Also note for later use when discussing for example the ARCH-implied Value-at-Risk, that the conditional distribution of X given a set $A = \{x : h\left(x\right) > a\}$, with $P\left(X \in A\right) > 0$ and $h : \mathbb{R} \to \mathbb{R}$, has the density

$$f\left(x|A\right) = \frac{f\left(x\right)}{P\left(X \in A\right)} \quad \text{for } h\left(x\right) > a,$$

such that,

$$E\left(X|X \in A\right) = \int x f\left(x|A\right) dx.$$

Example I.2.6 *With x_t $N\left(0,\sigma^2\right)$, distributed, the density of the distribution of x_t conditional on $x_t > 0$ is (using $P\left(x_t > 0\right) = \frac{1}{2}$) given by*

$$\begin{aligned} f\left(x_t|x_t > 0\right) &= \frac{f\left(x_t\right)}{P\left(x_t > 0\right)}1\left(x_t > 0\right) = \frac{\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}x_t^2\right)}{\frac{1}{2}}1\left(x_t > 0\right) \\ &= \sqrt{\frac{2}{\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}x_t^2\right)1\left(x_t > 0\right). \end{aligned}$$

Next, with z_t $N\left(0,1\right)$ distributed,

$$\begin{aligned} E\left(x_t|x_t > 0\right) &= \int_{-\infty}^{\infty} x\sqrt{\frac{2}{\sigma^2\pi}}\exp\left(-\frac{1}{2\sigma^2}x^2\right)1\left(x > 0\right)dx \\ &= 2\int_0^{\infty} x\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}x^2\right)dx = \int_{-\infty}^{\infty}|x|\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}x^2\right)dx \\ &= E\left|\sigma z_t\right| = \sigma\sqrt{\frac{2}{\pi}}, \end{aligned}$$

using well-known properties of the Gaussian distribution: (i) symmetry, and (ii) $E|z_t| = \sqrt{2/\pi}$.

I.2.2 The ARCH(1) process continued

We discuss here some first properties of the ARCH(1) process x_t defined in (I.1). It should be emphasized that all calculations in the next are done under the assumption of relevant finite order moments of the ARCH process which we will establish in the next section.

It is useful to discuss conditioning not only only x_{t-1} as in Example I.2.5 but also on all past variables $(x_{t-1}, x_{t-2}..., x_0)$. This is often referred to as 'conditioning on all past information' in the literature.

In terms of the definition of conditional expectations, we give meaning to $E(x_t|x_{t-1}, ..., x_0)$ by setting $X = x_t$ and $Y = (x_t, x_{t-1}, ..., x_0)$, and write it as

$$E(x_t|\mathcal{F}_{t-1}),$$

with

$$\mathcal{F}_{t-1} = (x_{t-1}, x_{t-2} \ldots, x_0).$$

In terms of this notation we can as in the example find the first moment of the ARCH process x_t,

$$E(x_t) = E(E(x_t|\mathcal{F}_{t-1})) = E(\sigma_t E(z_t)) = 0.$$

That the calculations are identical, reflects that by the defintion of the ARCH(1) process, the distribution of x_t conditional on $(x_{t-1}, x_{t-2}, ..., x_0)$ depends only on x_{t-1}.

Consider next the correlation between x_t and x_{t-1}, say

$$E(x_{t-1}x_t) = E(E(x_t x_{t-1}|\mathcal{F}_{t-1})) = E(x_{t-1}E(x_t|\mathcal{F}_{t-1})) = 0$$

and likewise,

$$E(x_{t-k}x_t) = E\left(E\left(x_{t-k}x_t|\mathcal{F}_{t-k}\right)\right) = 0 \text{ for any } k \geq 1$$

Hence the ARCH(1) process is a mean zero and uncorrelated process.

Note that this – unlike for Gaussian variables – *does not imply* that x_t and x_{t-1} are independent as σ_t^2, and hence x_t, depends on x_{t-1}^2.

Now turn to the second order moment, or as $Ex_t = 0$, the variance:

$$V(x_t) = E(x_t^2) = E\left(E\left(x_t^2|\mathcal{F}_{t-1}\right)\right) = E(\sigma_t^2 E(z_t^2)) = E(\sigma_t^2) = \sigma^2 + \alpha E\left(x_{t-1}^2\right) \tag{I.14}$$

Hence, if $E(x_t^2)$ is constant and finite (or, $\alpha < 1$), the variance of x_t is given by,

$$V(x_t) = E\left(x_t^2\right) = \frac{\sigma^2}{1-\alpha}. \tag{I.15}$$

Next, consider the "tail" behavior of x_t by considering 3. and 4. order moments. As odd moments of the $N(0,1)$ distribution are zero it follows as above for the first order moment that,

$$E(x_t^{2k+1}) = 0$$

i.e. all odd moments are zero. For the 4th order moment it follows that,

$$E(x_t^4) = 3(\tfrac{1-\alpha^2}{1-3\alpha^2})E(x_t^2)^2 > 3E(x_t^2)^2 \tag{I.16}$$

if $\alpha < 1/\sqrt{3}$. As $E(x_t^4)/E(x_t^2)^2 = 3$ for the Gaussian case, we conclude that the ARCH(1) process has indeed fatter tails or so-called *excess kurtosis* since $1-\alpha^2 > 1-3\alpha^2$, or $\alpha < 1/\sqrt{3}$.

I.3 Stationarity and mixing

Above we computed moments $E\left(x_t^k\right)$ for the ARCH(1) process assuming that they were the same for all time points, that is $E\left(x_t^k\right) = E\left(x_{t+n}^k\right)$ for all n, and that they were finite. To show under which conditions on the parameters this is the case we introduce theory from probability analysis. We first discuss stationarity.

I.3.1 Stationarity

A time series or, equivalently, a stochastic process, $(X_t)_{t=0,1,2,...}$ is a sequence of stochastic variables with each X_t taking values in $\mathbb{R}$ or $\mathbb{R}^p$. A key concept in time series analysis is that of stationarity of a stochastic process which can be formally defined as follows:

Definition I.3.1 *The process $(X_t)_{t=0,1,2,...}$ is said to be stationary, or simply X_t is said to be stationary, if for all $h, h \geq 0$, and t the joint distribution of $(X_t, \ldots, X_{t+h})$ does not depend on $t, t \geq 0$.*

Note that by definition for a stationary process with well-defined second order moments, the expectation $E(X_t)$ and variance $V(X_t)$ are constant, while the covariance between X_t and X_{t+h}, $Cov\left(X_t, X_{t+h}\right)$ depends only on h, and not on t.

Example I.3.1 *With x_t i.i.d. $N(0,\sigma^2)$ for all $t \geq 0$, then for $h \geq 0$,*

$$(x_t, \ldots, x_{t+h}) \text{ is } N_{h+1}\left(0, \Omega_h\right),$$

with $\Omega_h = \sigma^2 I_{h+1}$ *where* I_{h+1} *is the* $(h+1)$*-dimensional identity matrix. We can also write*

$$\Omega_h = diag\left(\sigma^2, ..., \sigma^2\right).$$

This distribution does not depend on t *and naturally the i.i.d. sequence is stationary.*

This was a very simple example of a Gaussian process, where X_t is said to be Gaussian if $(X_t, ..., X_{t+h})$ is Gaussian distributed for all t and h. As the Gaussian distribution is characterized alone by the first two moments, it holds that X_t is stationary if, and only if, $E(X_t)$ is constant and $Cov(X_t, X_{t+h}) = v(h)$ that is, the covariance is a function of h and hence independent of t. Thus for Gaussian processes it is enough to consider the first two moments when discussing stationarity.

Example I.3.2 *The univariate Gaussian moving average process* x_t *of order* 1, $MA(1)$, *is given by*

$$x_t = \varepsilon_t + \theta\varepsilon_{t-1},$$

with ε_t *i.i.d.* $N(0, \sigma^2)$. *In particular* x_t *is a stationary process with* $Ex_t = 0$, $V(x_t) = (1+\theta^2)\sigma^2$, $Cov(x_t, x_{t+1}) = \theta\sigma^2$ *and*

$$Cov(x_t, x_{t+h}) = 0 \quad for\ h > 1.$$

Hence the MA(1) process is stationary.

In the next example we use the result:

Lemma I.3.1 *With* $\phi \in \mathbb{R}$ *and* $\phi \neq 1$, *then*

$$1 + \phi + \phi^2 + ... + \phi^n = \sum_{i=0}^{n} \phi^i = \left(1 - \phi^{n+1}\right) / (1 - \phi).$$

If moreover $|\phi| < 1$, $\phi^n \to 0$ *as* $n \to \infty$, *and*

$$\sum_{i=0}^{\infty} \phi^i = 1/(1-\phi). \tag{I.17}$$

Example I.3.3 *Consider the AR process given by,*

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

for $t = 1, 2, ..., T$, initial value x_0 and where ε_t i.i.d.N(0,σ^2). Simple recursion shows that

$$x_t = \sum_{i=0}^{t-1} \rho^i \varepsilon_{t-i} + \rho^t x_0. \tag{I.18}$$

Hence in particular,

$$E(x_t) = \rho^t x_0$$

which depends on t and the AR process is thus not stationary. However, we can make it stationary by choosing x_0 as below and restricting ρ such that $|\rho| < 1$. In this case x_t has the stationary distribution,

$$x_t^* = \sum_{i=0}^{\infty} \rho^i \varepsilon_{t-i}.$$

To see this, we can give the initial value x_0 the distribution of x_0^*, that is $x_0 = \sum_{i=0}^{\infty} \rho^i \varepsilon_{0-i}$. Then simple insertion in (I.18) indeed gives,

$$x_t = \sum_{i=0}^{t-1} \rho^i \varepsilon_{t-i} + \rho^t \sum_{i=0}^{\infty} \rho^i \varepsilon_{0-i} = \sum_{i=0}^{t-1} \rho^i \varepsilon_{t-i} + \sum_{i=t}^{\infty} \rho^i \varepsilon_{t-i} = x_t^*.$$

And as (x_t^*) is Gaussian with $E(x_t^*) = 0$, using (I.17) with $\phi = \rho^2$,

$$Var(x_t^*) = \sum_{i=0}^{\infty} \rho^{2i} \sigma^2 = \sigma^2 / \left(1 - \rho^2\right).$$

and using that ε_t are i.i.d. such that $Cov(\varepsilon_t \varepsilon_{t+h}) = 0$ for $h \geq 1$, we also find the formula for the covariance of a stationary AR(1) process,

$$\begin{aligned} Cov\left(x_t^*, x_{t+h}^*\right) = E\left(x_t^* x_{t+h}^*\right) &= E\left(\sum_{i=0}^{\infty} \rho^i \varepsilon_{t-i} \sum_{i=0}^{\infty} \rho^i \varepsilon_{t+h-i}\right) \\ &= E\left(\sum_{i=0}^{\infty} \rho^i \varepsilon_{t-i} \sum_{i=h}^{\infty} \rho^i \varepsilon_{t+h-i}\right) \\ &= E\left(\sum_{i=0}^{\infty} \rho^i \varepsilon_{t-i} \sum_{j=0}^{\infty} \rho^i \varepsilon_{t-i} \rho^h\right) = \rho^h V(x_t^*), \end{aligned}$$

we conclude that x_t^* is stationary.

That is, if $|\rho| < 1$ the initial value x_0 can be given an initial distribution such that $x_t = x_t^*$ is stationary.

Note that all computations done here requires in particular $\sum_{i=0}^{\infty} \rho^i \varepsilon_{t-i}$ to be well-defined. The process is an example of linear processes known from

time series analysis and is well-defined for $|\rho| < 1$ and ε_t i.i.d.. The theory below will show that it is in fact not needed to introduce the inifinite sums and the explicit stationary solution x_t^* to discuss stationarity and dependence structure of the AR process.

More generally, we can apply the concept of stationarity to the familiar autoregressive (AR) and moving average (MA) processes by considerations as above. However, we cannot do so for the ARCH processes and different techniques will be applied. These techniques can of course also be applied to AR and MA processes as noted in the previous example.

I.3.2 Weakly mixing and asymptotic independence

The definition of stationarity adresses the joint distribution of the variables $X_{t+1}, \ldots, X_{t+h}$ for all h and t, but states nothing about dependence over time. To measure correlation over time for stationary processes usually the covariance function,

$$v(h) \equiv Cov(X_t, X_{t+h}), \tag{I.19}$$

is considered. As already noted for a stationary process this does not depend on t. For a univariate stationary process the well-known autocorrelation function is defined by,

$$\rho(h) = Corr\left(X_t, X_{t+h}\right) = \frac{Cov(X_t, X_{t+h})}{\sqrt{V\left(X_t\right) V\left(X_{t+h}\right)}} = \frac{v(h)}{v(0)}, \tag{I.20}$$

where the last equality holds by stationarity.

The functions $\rho\left(h\right)$ and $v\left(h\right)$ for various h describe the correlatedness over periods of time. Weakly mixing, or asymptotic indepencence, which is defined next, states that the dependence between X_t and X_{t+h} vanishes as h increases and hence replaces independence. More precisely:

Definition I.3.2 *A stationary process $\left(X_t\right)_{t=0,1,\ldots}$ is said to be weakly mixing, if for all t, h and sets A, B,*

$$\begin{aligned} &P((X_0, \ldots, X_t) \in A,\ (X_h, \ldots,\ X_{t+h}) \in B) \\ &\rightarrow P((X_0, \ldots, X_t) \in A) P((X_0, \ldots,\ X_t) \in B) \end{aligned}$$

as $h \rightarrow \infty$.

Intuitively this means that events removed from one another in time are independent. The next result relates correlation with weakly mixing:

Lemma I.3.2 *If the stationary process* $(X_t)_{t=0,1,2,..}$ *is weakly mixing and has finite variance, then the covariance* $v(h) = Cov(X_t, X_{t+h})$ *tends to zero as* $h \to \infty$.

If X_t *is a stationary Gaussian process and* $v(h) \to 0$, $h \to \infty$ *then* X_t *is weakly mixing.*

Example I.3.4 *It follows by Example I.3.2 that the MA(1) process is weakly mixing and likewise, if* $|\rho| < 1$, *the AR(1) process is weakly mixing by Example I.3.3.*

A key implication of mixing is that a law of large numbers (LLN) apply:

Theorem I.3.1 *Assume that with* $X_t \in \mathbb{R}^p$, $(X_t)_{t=0,1,2,3,...}$ *is a weakly mixing process, and assume that the function* $g : \mathbb{R}^{p(m+1)} \to \mathbb{R}$, $m \geq 0$, *satisfies* $E|g(X_t, X_{t+1}, ..., X_{t+m})| < \infty$. *Then as* $T \to \infty$,

$$\frac{1}{T}\sum_{t=1}^{T} g(X_t, X_{t+1}, ..., X_{t+m}) \xrightarrow{P} E(g(X_t, X_{t+1}, ..., X_{t+m})). \tag{I.21}$$

This version of the LLN applies a quite general formulation in terms of the function g.

Example I.3.5 *The formulation means that for example if* X_t *is univariate and mixing with finite second order moments, then*

$$\frac{1}{T}\sum_{t=1}^{T} X_t^2 \xrightarrow{P} E(X_t^2) \quad \text{and} \quad \frac{1}{T}\sum_{t=1}^{T} X_t X_{t+1} \xrightarrow{P} E(X_t X_{t+1}),$$

by applying Theorem I.3.1 with

$$g(X_t) = X_t^2 \text{ and } g(X_t, X_{t+1}) = X_t X_{t+1}.$$

Example I.3.6 *If* $EX_t = 0$, *such that* $Var(X_t) = E(X_t^2)$ *and* $Cov(X_t, X_{t+h}) = E(X_t X_{t+h})$ *it follows that the empirical autocorrelation function, see (I.20), as defined by,*

$$\widehat{\rho}(h) \equiv \frac{\frac{1}{T}\sum_{t=1}^{T-h} X_t X_{t+h}}{\sqrt{\frac{1}{T}\sum_{t=1}^{T} X_t^2 \frac{1}{T}\sum_{t=1}^{T-h} X_{t+h}^2}}, \tag{I.22}$$

will converge in probability to (the theoretical) $\rho(h)$ *by Theorem I.3.1 motivating that most software for time series allows one to compute these directly.*

I.3.3 Moments and stationarity using the drift criterion

Before turning to the concept of a drift function and the drift criterion, we note the key implication that if $(X_t)_{t=0,1,2,...}$ satifies the drift criterion, the initial value, X_0, can be given a distribution such that indeed X_t is stationary. This resembles the considerations we made for the AR(1) process x_t in Example I.3.3 where we could explicitly choose an initial distribution of x_0 such that x_t became stationary. Moreover, with X_0 given this initial distribution, the stationary process is also *weakly mixing* such that the law of large numbers can be applied. And in addition, the drift criterion, implies that X_t has finite moments as we will see.

Hence establishing the drift criterion is very helpful in many ways. The presentation here is based on Meyn and Tweedie (1993), Tjøstheim (1990) and Hansen and Rahbek (1998), see alsoCarrasco and Chen (2002) for general ARCH and related stochastic volatility processes.

I.3.3.1 Assumptions

Now a common key feature of the the AR(1) process in Example I.2.4 and the ARCH(1) process in (I.1) is that with X_t denoting either of the two, then the conditional distribution of,

$$X_t \text{ given } (X_{t-1}, X_{t-2}, ..., X_0)$$

depends only on X_{t-1}. More precisely, in the AR(1) case x_t conditionally on x_{t-1}, is $N(\rho x_{t-1}, \sigma^2)$ distributed, while for the ARCH(1) x_t conditionally on x_{t-1}, is $N(0, \sigma_t^2)$ distributed with $\sigma_t^2 = \sigma^2 + \alpha x_{t-1}^2$. In both cases, the conditional distribution has a Gaussian density which has some attractable features. We make the following assumption:

Assumption I.3.1 *Assume that for $(X_t)_{t=0,1,2,...}$ with $X_t \in \mathbb{R}^p$ it holds that:*

(i) *the conditional distribution of X_t given $(X_{t-1}, X_{t-2}, ..., X_0)$ depends only on X_{t-1}, that is*

$$(X_t | X_{t-1}, X_{t-2}, ..., X_0) \overset{D}{=} (X_t | X_{t-1}).$$

(ii) *the conditional distribution of X_t given X_{t-1}, has a positive conditional density $f(X_t|X_{t-1})$, $f(X_t|X_{t-1}) > 0$, which is continuous.*

Note that (i) implies that $(X_t)_{t=0,1,2,...}$ is a Markov chain on $\mathbb{R}^p$, that is a Markov chain with non-integer values, or what is called a Markov chain on a general state space. The condition (ii) of continuity, while simple to validate, is not needed and milder conditions on the conditional density can be applied in general.

Example I.3.7 *For the AR process x_t in Example I.2.4, x_t conditional on x_{t-1} has density*

$$f(x_t|x_{t-1}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_t-\rho x_{t-1})^2}{2\sigma^2}\right),$$

which is positive and continuous in x_t and x_{t-1}.

Example I.3.8 *For the ARCH process in (I.1), x_t conditional on x_{t-1} has the Gaussian density,*

$$f(x_t|x_{t-1}) = \frac{1}{\sqrt{2\pi\sigma_t^2}} \exp\left(-\frac{1}{2\sigma_t^2}x_t^2\right), \quad \sigma_t^2 = \sigma^2 + \alpha x_{t-1}^2$$

which again is continuous as desired.

We shall in the next use the follwing result from probability analysis:

Lemma I.3.3 *If a real variable X, $X \in \mathbb{R}$, has density $f(x)$, then $Y = cX$, with $c \neq 0$ a constant, has density $\frac{1}{\sqrt{c^2}} f\left(\frac{y}{c}\right)$. Moreover, if $VX < \infty$, then $EY = cEX$ and $VY = c^2 VX$.*

Example I.3.9 *In the ARCH process x_t in (I.1) the assumption of z_t i.i.d.$N(0,1)$ is sometimes replaced by the assumption that z_t is i.i.d.$D(0,1)$ where D is a t_v-distribution scaled by $\sqrt{\frac{v-2}{v}}$. Here $v > 2$ and denotes the degrees of freedom. An ARCH process defined this way satisfies Assumption I.3.1.*

To see this note first that if X is t_v-distributed with $v > 2$, then X has $EX = 0$ and $VX = v/(v-2)$. Moreover, X has density,

$$f(x) = \frac{\gamma(v)}{\sqrt{v\pi}} \left(1 + \frac{x^2}{v}\right)^{-\left(\frac{v+1}{2}\right)},$$

where the constant $\gamma(v) = \Gamma\left(\frac{v+1}{2}\right) / \sqrt{\Gamma\left(\frac{v}{2}\right)}$, with $\Gamma(\cdot)$ the so-called Gamma function.

As $VX = v/(v-2)$, *then using Lemma I.3.3,* $z_t = \left(\sqrt{\frac{v-2}{v}}\right) X$ *has* $V(z_t) = 1$ *and* $E(z_t) = 0$, *explaining the assumption on the i.i.d.D(0,1) innovations* z_t. *It follows likewise by simple insertion that* z_t *has density*

$$f(z) = \frac{\gamma(v)}{\sqrt{(v-2)\pi}} \left(1 + \frac{z^2}{(v-2)}\right)^{-\left(\frac{v+1}{2}\right)}.$$

Next, by the ARCH equation $x_t = \sigma_t z_t$, *and hence, using Lemma I.3.3 again,* x_t *conditional on* x_{t-1} *has density,*

$$f(x_t|x_{t-1}) = \frac{\gamma(v)}{\sqrt{\sigma_t^2(v-2)\pi}} \left(1 + \frac{x_t^2}{\sigma_t^2(v-2)}\right)^{-\left(\frac{v+1}{2}\right)}, \quad \sigma_t^2 = \sigma^2 + \alpha x_{t-1}^2.$$

I.3.3.2 Drift function

Next, we define the concept of *a drift function* for a process X_t satisfying Assumption I.3.1. With X_t a time series, a drift function for X_t is some function δ, where $\delta(X_t) \geq 1$ and which is not identically ∞. The choice of drift function is quite flexible, but a key example is the next.

Example I.3.10 *A much used drift function in the analysis of univariate AR and ARCH processes is*

$$\delta(X_t) = 1 + X_t^2.$$

If X_t *is a vector,* $X_t = (X_{1t}, ..., X_{pt})' \in \mathbb{R}^p$, *then one may apply,*

$$\delta(X_t) = 1 + X_t' X_t = 1 + \sum_{i=1}^{p} X_{it}^2.$$

The role of such a drift function is to measure the dynamics, or the *drift* of X_t, by studying the dynamics of $\delta(X_t)$ instead. We do this by considering the conditional expectation of $\delta(X_t)$ given X_{t-1} or some other past value of X_t, say X_{t-m}. That is we are interested in measuring

$$E(\delta(X_t)|X_{t-m}),$$

for some m, where typically $m = 1$ is used.

Example I.3.11 *Consider the AR(1) process x_t in Example I.2.4 and set $\delta(x_t) = 1 + x_t^2$. Then*

$$\begin{aligned} E(\delta(x_t)|x_{t-1}) &= E\left(1 + (\rho x_{t-1} + \varepsilon_t)^2 |x_{t-1}\right) \\ &= 1 + \rho^2 E\left(x_{t-1}^2|x_{t-1}\right) + 2\rho x_{t-1} E(\varepsilon_t|x_{t-1}) + E\left(\varepsilon_t^2|x_{t-1}\right) \\ &= 1 + \rho^2 x_{t-1}^2 + 2\rho x_{t-1} E(\varepsilon_t) + E\left(\varepsilon_t^2\right) \\ &= 1 + \sigma^2 + \rho^2 x_{t-1}^2 \\ &= \rho^2 \delta(x_{t-1}) + c, \end{aligned} \tag{I.23}$$

where the constant c is given by $c = (1 - \rho^2 + \sigma^2)$. Thus, apart from the constant c, we obtain what mimics a simple first order autoregression in $\delta(x_t)$. That is, we may write

$$\delta(x_t) = \rho^2 \delta(x_{t-1}) + c + \eta_t,$$

with $\eta_t = (\delta(x_t) - E(\delta(x_t)|x_{t-1}))$ and such that by defintion $E\eta_t = 0$. The equation for $\delta(x_t)$ is stable if $\rho^2 < 1$ and persistent if $\rho = 1$. That is, the dynamics of the drift function $\delta(x_t)$ reflects the dynamics of the AR process x_t.

Example I.3.12 *Consider the ARCH process x_t given by (I.1) and set $\delta(x_t) = 1 + x_t^2$. Then*

$$\begin{aligned} E(\delta(x_t)|x_{t-1}) &= E\left(1 + \sigma_t^2 z_t^2 |x_{t-1}\right) \\ &= 1 + \left(\sigma^2 + \alpha x_{t-1}^2\right) E\left(z_t^2|x_{t-1}\right) \\ &= 1 + \alpha x_{t-1}^2 + \sigma^2 \\ &= \alpha \delta(x_{t-1}) + c, \end{aligned}$$

where $c = (1 + \sigma^2 - \alpha)$. Thus as before in the AR example, we can interpretate this as a simple autoregression in $\delta(x_{t-1})$ with autoregressive coefficient α. This is stable if $\alpha < 1$ (as $\alpha \geq 0$ by definition of the ARCH process). That indeed the ARCH process x_t itself is a stable and (asymptotically) stationary process if $\alpha < 1$ is a key implication of the next.

In the above examples the dynamics of the drift function, or rather the conditional expectation of $\delta(X_t)$ given X_{t-m}, we used the concept of stability informally. More precisely, we shall make the following assumption:

Assumption I.3.2 *Assume that $(X_t)_{t=0,1,2,..}$, with $X_t \in \mathbb{R}^p$, satifies Assumption I.3.1. Assume further that there is a drift function δ, $\delta(X_t) \geq 1$,*

which for some lag m, there are positive constants M, C and ϕ, $\phi < 1$, such that,

$$\begin{array}{ll} (i) & E\left(\delta\left(X_t\right)|X_{t-m}\right) \leq \phi\delta\left(X_{t-m}\right) \quad \text{for } X'_{t-m}X_{t-m} > M, \\ (ii) & E\left(\delta\left(X_t\right)|X_{t-m}\right) \leq C \qquad\qquad\quad \text{for } X'_{t-m}X_{t-m} \leq M. \end{array}$$

If (X_t) satisfies Assumption I.3.2 then we say that X_t satisfies the drift criterion (with drift function δ).

Example I.3.13 *The AR process x_t in satisfies the drift criterion if $\rho^2 < 1$ with $\delta\left(x_t\right) = 1 + x_t^2$ with x_{t-1}^2 chosen large. To see this note first that the calculations in Example I.3.10 gave,*

$$E\left(\delta\left(x_t\right)|x_{t-1}\right) = \rho^2\delta\left(x_{t-1}\right) + c, \quad \text{with } c = \left(1 - \rho^2 + \sigma^2\right).$$

As $\rho^2 < 1$ we may choose ϕ smaller than 1, but larger than ρ^2. That is, we choose ϕ such that $\rho^2 < \phi < 1$. With this choice of ϕ, $(\phi - \rho^2) > 0$, and hence for x_{t-1}^2 large enough, $x_{t-1}^2 > M$ say,

$$c < \left(\phi - \rho^2\right)\delta\left(x_{t-1}\right).$$

We can therefore conclude that there for some large $M > 0$, and $x_{t-1}^2 > M$,

$$E\left(\delta\left(x_t\right)|x_{t-1}\right) = \rho^2\delta\left(x_{t-1}\right) + c < \phi\delta\left(x_{t-1}\right).$$

For $x_{t-1}^2 \leq M$, we have $\rho^2\delta\left(x_{t-1}\right) + c \leq \rho^2\delta\left(M\right) + c = C$.

Example I.3.14 *The ARCH process x_t satisfies the drift criterion if $\alpha < 1$ with $\delta\left(x_{t-1}\right) = 1 + x_{t-1}^2$. Using that by Example I.3.12,*

$$E\left(\delta\left(x_t\right)|x_{t-1}\right) = \alpha\delta\left(x_{t-1}\right) + c, \quad \text{with } c = \left(1 - \alpha + \sigma^2\right),$$

the considerations in Example I.3.13 carry over directly with $\alpha = \rho^2$.

We are now in position to state the following result from Tjøstheim (1990) and Jensen and Rahbek (2007):

Theorem I.3.2 *Assume that $(X_t)_{t=0,1,...}$ satisfies the drift criterion with drift function δ. Then the initial value, X_0, can be given a distribution such that $(X_t)_{t=0,1,2}$ is a stationary process. Moreover, the stationary process is weakly mixing and has finite moments, $E\left(\delta\left(X_t\right)\right) < \infty$.*

Finally, for any initial value X_0, the law of large numbers in (I.21) in Theorem I.3.1 applies to X_t.

Thus, if X_t satisfies the drift criterion, not only is the process stationary (by giving X_0 the correct distribution) and weakly mixing such that the law of large numbers apply, but as $E(\delta(X_t)) < \infty$, then any moments of X_t which are bounded by the drift function δ are finite.

Furthermore, the last statement in the theorem says that if X_0 does not have the correct stationary initial distribution, the importance of this will vanish. In fact, it will vanish exponentially fast, and the distribution of X_t for moderately large t will resemble the stationary distribution. This feature is often referred to as *geometric ergodicity.*

Example I.3.15 *By Example I.3.13 For the AR process we may conclude that $Ex_t^2 < \infty$, and that the law of large numbers apply to the stationary x_t by Theorem I.3.2. While this illustrates the results, this conclusion is not surprising as we already know that with $\rho^2 < 1$, then x_t has a stationary representation and since it is Gaussian actually $Ex_t^{2k} < \infty$ for any k. To give an understanding of the role of initial value, recall from Example I.3.3 that with x_0 fixed,*

$$x_t = \rho^t x_0 + \sum_{i=0}^{t-1} \rho^i \varepsilon_{t-i} \stackrel{D}{=} N\left(\rho^t x_0, \sigma^2 \frac{1-\rho^{2t}}{1-\rho^2}\right),$$

while for the stationary version,

$$\sum_{i=0}^{\infty} \rho^i \varepsilon_{t-i} \stackrel{D}{=} N\left(0, \sigma^2 \frac{1}{1-\rho^2}\right).$$

We observe that $\rho^t x_0 \to 0$ exponentially fast, and likewise $\frac{1-\rho^{2t}}{1-\rho^2} \to \frac{1}{1-\rho^2}$, and hence the initial value plays no role asymptotically.

Example I.3.16 *For the ARCH process,*

$$x_t = \sigma_t z_t, \quad \sigma_t^2 = \sigma^2 + \alpha x_{t-1}^2,$$

with z_t i.i.d.$N(0,1)$ we may from Example I.3.14 conclude that if $0 \le \alpha < 1$ then x_t has a stationary solution with $Ex_t^2 < \infty$. Hence any moments of order lower than 2 are finite, for example $E|x_t| < \infty$ since $|x_t| \le \delta(x_t) = 1 + x_t^2$.

We do not know if for example x_t has finite fourth order moments, $Ex_t^4 < \infty$. To find out under which conditions this holds we need to consider a drift function from which we can conclude this. An example is

$$\delta(x_t) = 1 + x_t^4.$$

With z_t *i.i.d.*$N(0,1)$, $E(z_t^4) = 3$ *and we find,*

$$\begin{aligned} E(\delta(x_t)|x_{t-1}) &= 1 + \left(\sigma^2 + \alpha x_{t-1}^2\right)^2 E\left(z_t^4\right) \\ &= 1 + 3\left(\sigma^4 + 2\alpha\sigma^2 x_{t-1}^2 + \alpha^2 x_{t-1}^4\right) \\ &= 3\alpha^2\left(1 + x_{t-1}^4\right) + \left(1 - 3\alpha^2 + 3\sigma^4\right) + 6\alpha\sigma^2 x_{t-1}^2 \\ &= 3\alpha^2\delta(x_{t-1}) + c\left(x_{t-1}^2\right), \end{aligned}$$

where $c\left(x_{t-1}^2\right) = c + 6\alpha\sigma^2 x_{t-1}^2$, *with* $c = \left(1 - 3\alpha^2 + 3\sigma^4\right)$. *We thus need to choose* α *so small that* $3\alpha^2 < 1$. *To see that the drift-criterion holds with such an* α, *choose* ϕ *such that* $3\alpha^2 < \phi < 1$. *Next we need to establish that for for some large* M *and* $x_{t-1}^2 > M$,

$$\left(\phi - 3\alpha^2\right)\delta(x_{t-1}) > c\left(x_{t-1}^2\right).$$

But this clearly holds as $x_{t-1}^4 > x_{t-1}^2$ *for large* x_{t-1}^2 *and* $(\phi - 3\alpha^2) > 0$.

Hence the conclusion is that while a stationary x_t *exists for* $\alpha < 1$ *and* $Ex_t^2 < \infty$ *in this case, we need to restrict* α *further to have fourth order moments. More precisely, and as already indicated, provided* $\alpha < 1/\sqrt{3} \simeq 0.56$ *then* $Ex_t^4 < \infty$.

We saw in Example I.3.16 that the value of α in the conditional variance $\sigma_t^2 = \sigma^2 + \alpha x_{t-1}^2$ was cruical for stationarity of x_t and also for finite moments of x_t. This is a typical feature of non-linear time series where parameter values have implications for interpretation in terms of both stationarity and finite moments.

In terms of the AR process x_t the autoregressive coefficient is the key parameter to an understanding of the dynamics of $x_t = \rho x_{t-1} + \varepsilon_t$ as is well-known. We know that if $|\rho| < 1$ then x_t is stationary and with ε_t Gaussian all moments are finite. At the same time we know that the restriction of $|\rho| < 1$ is cruical in the sense that if $\rho = 1$, as often found empirically, then x_t is a unit-root process and non-stationary. Surprisingly this is not so for the ARCH process, where $\alpha = 1$ still allows x_t to be stationary. However, with $\alpha = 1$, then only moments up to order one are finite, $E|x_t| < \infty$. That is, $\alpha = 1$ implies stationarity but for a process without variance.

More precisely we can make the following table where we can divide the interval for α as follows:

ARCH process x_t defined in (I.1):	
$x_t = \sigma_t z_t,\ \sigma_t = \sigma^2 + \alpha x_{t-1}^2$ and z_t *i.i.d.*N$(0,1)$.	
Stationary for $0 \leq \alpha < 3.56$	
Finite moments:	
$\frac{\pi}{2} \leq \alpha < 3.56$	no moments (fractional)
$1 \leq \alpha < \frac{\pi}{2}$	$E\|x_t\| < \infty$
$\frac{1}{\sqrt{3}} \leq \alpha < 1$	$Ex_t^2 < \infty$
$0 \leq \alpha_1 < \frac{1}{\sqrt{3}}$	$Ex_t^4 < \infty$

Actually, see Nelson (1990), the 3.56 is an approximation of the number $\frac{1}{2}\exp(-\Psi(\frac{1}{2}))$ where $\Psi(\cdot)$ is the Euler psi function with $\Psi(\frac{1}{2}) \cong -1.96351$. The number appears from considering the condition for stationaríty

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

which can be solved for z_t *i.i.d.*N$(0,1)$ as above. In fact, if z_t has another distribution such as the t_v distribution mentioned in Example I.3.9 the intervals above would change.

This as well as other examples of ARCH models will be discussed in exercises.

I.4 Central limit theory

As noted, a powerful implication of Theorem I.3.2 for X_t is that independently of the initial value, X_0 the LLN in Theorem I.3.1 applies. That is,

$$\frac{1}{T}\sum_{t=1}^{T} g\left(X_t, X_{t+1}, ..., X_{t+m}\right) \xrightarrow{P} E\left(g\left(X_t, X_{t+1}, ..., X_{t+m}\right)\right), \tag{I.24}$$

provided that $E|g\left(X_t, ..., X_{t+m}\right)| < \infty$.

The following theorem generalizes this to also hold for the central limit theorem (CLT) in Meyn and Tweedie (1993, ch. 17). More precisely, we have:

Theorem I.4.1 (Meyn and Tweedie, 1993, Theorem 17.0.1) *Assume that Theorem I.3.2 applies to $(X_t)_{t\geq 0}$ with X_t stationary. With $f\left(X_t, X_{t-1}, ..., X_{t-m}\right) \in \mathbb{R}$, assume that $E|f^2\left(X_t, ..., X_{t-m}\right)| < \infty$, and $Ef(X_t, X_{t-1}, ..., X_{t-m})| = 0$. Then,*

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} f(X_t, X_{t-1}, ..., X_{t-m}) \xrightarrow{D} N(0, \gamma),$$

where

$$\gamma = \lim_{T\to\infty} E\left[\frac{1}{T}\left(\sum_{t=1}^{T} f(X_t, X_{t-1}, ..., X_{t-m})\right)^2\right].$$

A different version of the CLT is from Brown (1971) for *Martingale differences* which is the one we will apply when discussing asymptotic normality later for estimators[1]:

Theorem I.4.2 (Corollary to Brown, 1971) *Assume that Theorem I.3.2 applies to* $(X_t)_{t\geq 0}$ *with* X_t *stationary. With* $f(X_t, X_{t-1}, ..., X_{t-m}) \in \mathbb{R}$*, assume that* $E|f^2(X_t, ..., X_{t-m})| < \infty$*, and* $E(f(X_t, X_{t-1}, ..., X_{t-m})|X_{t-1}, ..., X_{t-m}) = 0$*. Then the central limit theorem (CLT) applies as* $T \to \infty$*,*

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} f(X_t, X_{t-1}, ..., X_{t-m}) \xrightarrow{D} N(0, E\left(f^2(X_t, X_{t-1}, ..., X_{t-m})\right)).$$

As an example of the application of the CLT quoted, consider:

Example I.4.1 *The empirical autocovariance function of order one for the ARCH(1) process* x_t*, is given by,*

$$\frac{1}{T}\sum_{1}^{T} x_t x_{t-1}. \tag{I.25}$$

If $\alpha < 1$*, such that* $Ex_t^2 < \infty$*, the LLN indeed implies the obvious result that as* T *tends to* ∞*, then (I.25) will converge in probability to* $Ex_t x_{t-1} = 0$ *using* $g(x_t, x_{t-1}) = x_t x_{t-1}$*.*

Likewise, one would expect that multiplied by $\sqrt{T}$*, the CLT in Theorem I.4.2 would apply to (I.25). Set therefore,*

$$Y_t = f(x_t, x_{t-1}) = x_t x_{t-1}.$$

Then Y_t *is a function of* (x_t, x_{t-1}) *and* $E(Y_t|x_{t-1}) = 0$ *as desired. Moreover, if* $Ex_t^4 < \infty$*, or* $\alpha < 1/\sqrt{3}$*,*

$$E\left(Y_t^2\right) = E\left(x_t^2 x_{t-1}^2\right) \leq \sqrt{E\left(x_t^4\right) E\left(x_{t-1}^4\right)} < \infty,$$

[1]Note that under an additional regularity condition (often referred to as "Lindeberg") then the stationarity requirement may be omitted, see Brown (1971)

using Hölders inequality, which says $E|XY| \leq \sqrt{E\left|X\right|^2 E\left|Y\right|^2}$ *for general random variables. Moreover, we can compute the variance,*

$$E\left(x_t^2 x_{t-1}^2\right) = E\left(x_{t-1}^2 E\left(x_t^2 | x_{t-1}\right)\right) = E\left(x_{t-1}^2\left(\sigma^2 + \alpha x_{t-1}^2\right)\right) = E\left(x_{t-1}^2\right)\sigma^2 + \alpha E\left(x_{t-1}^4\right),$$

with $E\left(x_{t-1}^2\right)$ *and* $E\left(x_{t-1}^4\right)$ *given in (I.15) and (I.16) respectively.*

We thus conclude that while $\alpha < 1$ *is sufficient for*

$$\frac{1}{T}\sum x_t x_{t-1} \xrightarrow{P} E\left(x_t x_{t-1}\right) = 0,$$

we need the stronger assumption that $\alpha < 1/\sqrt{3}$ *for*

$$\frac{1}{\sqrt{T}}\sum x_t x_{t-1} \xrightarrow{D} N\left(0, E\left(x_t^2 x_{t-1}^2\right)\right).$$

Often in applied work the autocorrelation function is studied for x_t^2*, given by,*

$$\frac{1}{T}\sum_1^T \left(x_t^2 - \left(\frac{1}{T}\sum_{t=1}^T x_t^2\right)\right) x_{t-1}^2. \tag{I.26}$$

Considerations as above lead to the requirement that $Ex_t^4 < \infty$ *for convergence in probability, while by using Theorem I.4.1 the requirement is* $Ex_t^8 < \infty$ *for convergence in distribution. These are quite strong restrictions, and therefore to avoid such, often the autocorrelation function is given for* $|x_t|$ *instead.*

References

Bera and Higgins, 1993, ARCH Models: Properties and Estimation, Journal of Economic Surveys, 305-366

Bollerslev, T., R.Y. Chou and K.F. Kroner, 1992, ARCH modelling in Finance: A review of the theory and empirical evidence, Journal of Econometrics, 52, 5-61

Carrasco and Chen, 2002, β-mixing and Moment properties of Various GARCH, Stochastic Volatility and ACD Models, Econometric Theory

Engle, R.F., 1982, Autoregressive conditional heteroscedaticity with estimates of United Kingdom inflation, Econometrica, 50, 987-1007

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

Franses, P.H. and D. van Dijk, 2000, Nonlinear Time Series Models in Empirical Finance, Cambridge University Press

Jensen, S.T. and A. Rahbek, 2007, On the Law of Large Numbers for (Geometrically) Ergodic Processes, Econometric Theory (2007), 23, 761-766

Hansen, E. and A. Rahbek, 1998, Stationarity and Asymptotics of Multivariate ARCH Time Series with an Application to Robustness of Cointegration Analysis, preprint no.*12*, Department of Theoretical Statistics

Meyn, S.P. and R.L. Tweedie, 1993, Markov chains and stochastic stability, Communications and Control Engineering Series, Springer-Verlag, London ltd., London

Nelson, D., 1990, Stationarity and persistence in the GARCH(1,1) model, Econometric Theory, 6, 318-334

Pagan, A., 1996, The econometrics of financial markets, Journal of Empirical Finance, 3:15-102

Shephard, N., 1996, Statistical Aspects of ARCH and Stochastic Volatility, in "Time Series Models", ed. Cox, Hinkley and Barndorff-Nielsen, Chapman Hall, London.

Taylor, S.J., 2005, Asset Price Dynamics, Volatility and Prediction, Princeton University Press

Tjøstheim, D., 1990, Non-linear time series and markov chains, Advances in Applied Probability, 22, 587-611

Anders Rahbek
Rasmus Søndergaard Pedersen
University of Copenhagen

September 2023

Part II
General ARCH and GARCH

Similar to empirical analyses of data from macroeconomics based on autoregressive models, it is often needed in financial applications to allow for richer dynamics than the first order ARCH for example by adding further lagged values of squared returns in the conditional volatility.

We discuss in this part how the previous results can be extended and applied to such cases. It turns out that the concept of stationary, or stable, dynamics is closely related to that of the largest eigenvalue of matrices and some theory for calculus with matrices will therefore be introduced.

II.1 The ARCH(2) process

The ARCH(2) process can for $t = 2, 3, 4 \ldots$ be represented as

$$x_t = \sigma_t z_t \tag{II.1}$$
$$\sigma_t^2 = \sigma^2 + \alpha_1 x_{t-1}^2 + \alpha_2 x_{t-2}^2 \tag{II.2}$$

with initial values x_0, x_1 and where the z_t's are *i.i.d.*N(0,1). This way the distribution of x_t conditional on the information up to time $t-1$, given by the variables $(x_{t-1}, ..., x_0)$, depends on (x_{t-1}, x_{t-2}), unlike the ARCH(1).

II.1.1 Assumption ?? for the ARCH(2) process

Clearly x_t does not satisfy Assumption **??**. However, the so-called companion satisfies Assumption **??** (i) where the companion form is given by defining,

$$X_t = \begin{pmatrix} x_t \\ x_{t-1} \end{pmatrix}. \tag{II.3}$$

By this choice X_t conditional on the past values $(X_{t-1}, ..., X_0)$ depends only on $X_{t-1} = (x_{t-1}, x_{t-2})'$,

$$X_t = \begin{pmatrix} x_t \\ x_{t-1} \end{pmatrix} = \begin{pmatrix} \sigma_t z_t \\ x_{t-1} \end{pmatrix} = \begin{pmatrix} \sqrt{\left(\sigma^2 + \alpha_1 x_{t-1}^2 + \alpha_2 x_{t-2}^2\right)} z_t \\ x_{t-1} \end{pmatrix}.$$

However, if we condition on X_{t-1} we condition on x_{t-1} in particular such that the conditional density of X_t given X_{t-1} is not satisfying Assumption **??** (ii) as it is singular. We can see this directly by using the formula for conditional densities in (**??**), $f(x,y) = f(x|y) f(y)$, from which we would get

$$\begin{aligned} f(X_t|X_{t-1}) &= f\left((x_t, x_{t-1}) \,|\, (x_{t-1}, x_{t-2})\right) = f(x_t|x_{t-1}, x_{t-2}) f(x_{t-1}|x_{t-1}, x_{t-2}) \\ &= f(x_t|x_{t-1}, x_{t-2}) f(x_{t-1}|x_{t-1}), \end{aligned}$$

where $f(x_{t-1}|x_{t-1}, x_{t-2}) = f(x_{t-1}|x_{t-1})$, since x_{t-1} is fixed, and hence singular. That is, while indeed the first term on the right hand side $f(x_t|x_{t-1}, x_{t-2})$ is a continuous Gaussian density by definition of the ARCH(2),

$$f(x_t|x_{t-1}, x_{t-2}) = \frac{1}{\sqrt{2\pi\sigma_t^2}} \exp\left(-\frac{x_t^2}{2\sigma_t^2}\right),$$

the second term $f(x_{t-1}|x_{t-1})$ is singular.

By definition, the equation for x_{t-1} is given by,

$$x_{t-1} = \sigma_{t-1} z_{t-1} = \sqrt{\left(\sigma^2 + \alpha_1 x_{t-2}^2 + \alpha_2 x_{t-3}^2\right)} z_{t-1},$$

such that if we conditioned on x_{t-2} and x_{t-3} instead, that is $X_{t-2} = (x_{t-2}, x_{t-3})'$, the density of x_{t-1} given (x_{t-2}, x_{t-3}) would be well-defined and Gaussian.

Hence, it appears that while X_t conditional on X_{t-1} is not nice in the sense that it has a singular density, then X_t conditional on X_{t-2} would be satisfying a continuity condition.

To derive the density for X_t conditional on X_{t-2} we can use the formula (**??**) for conditional densities as follows,

$$\begin{aligned} f(X_t|X_{t-2}) &\overset{\text{(by definition)}}{=} f(x_t, x_{t-1}|x_{t-2}, x_{t-3}) \\ &\overset{\text{(by (??))}}{=} \frac{f(x_t, x_{t-1}, x_{t-2}, x_{t-3})}{f(x_{t-2}, x_{t-3})} \\ &\overset{\text{(multiply and divide)}}{=} \left(\frac{f(x_t, x_{t-1}, x_{t-2}, x_{t-3})}{f(x_{t-1}, x_{t-2}, x_{t-3})}\right)\left(\frac{f(x_{t-1}, x_{t-2}, x_{t-3})}{f(x_{t-2}, x_{t-3})}\right) \\ &\overset{\text{(by (??))}}{=} f(x_t|x_{t-1}, x_{t-2}) f(x_{t-1}|x_{t-2}, x_{t-3}). \end{aligned} \tag{II.4}$$

As noted, $f(x_t|x_{t-1}, x_{t-2})$ and $f(x_{t-1}|x_{t-2}, x_{t-3})$ are Gaussian densities, and hence $f(X_t|X_{t-2})$ would satisfy Assumption **??** (ii) as it is the product of two continuous densities.

To allow for the general case of more lags, k say, in the conditional variance and mean equations we can modify Assumption **??** as follows:

Assumption II.1.1 *Assume that for* $(X_t)_{t=0,1,2,...}$ *with* $X_t \in \mathbb{R}^p$ *it holds that:*

(i) *the conditional distribution of* X_t *given* $(X_{t-1}, X_{t-2}, ..., X_0)$ *depends only on* X_{t-1}*, that is*

$$(X_t|X_{t-1}, X_{t-2}, ..., X_0) \stackrel{D}{=} (X_t|X_{t-1}) .$$

(ii) *the conditional distribution of* X_t *given* X_{t-k}*, for some* $k \geq 1$*, has a positive conditional density* $f(X_t|X_{t-k})$*,* $f(X_t|X_{t-k}) > 0$*, which is continuous.*

And we note that Theorem **??** indeed holds with the drift criterion, that is Assumption **??**, satisfied and under Assumption II.1.1 instead of Assumption **??**.

Next we turn to the drift criterion for the ARCH(2) process.

II.1.2 The drift criterion for the ARCH(2)

The theory here highlights the reason why Assumption **??** allows for a general lag m when computing $E(\delta(X_t)|X_{t-m})$, where we in the previous examples used $m = 1$. It is closely related to our previous discussion on the modifed Assumption II.1.1 which was needed for the ARCH(2) process, and of course, ARCH(k) processes in general.

We start by considering the AR(2) process, and use this to introduce some few results for matrices. Recall first the concept of an eigenvalue:

Definition II.1.1 *With* A *a* $p \times p$ *matrix,*

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1p} \\ \vdots & & \vdots \\ a_{p1} & \cdots & a_{pp} \end{pmatrix},$$

where $a_{ij} \in \mathbb{R}$*, then the eigenvalues*[1] *of* A*,* λ_i*,* $i = 1, .., p$*, are found by solving,*

$$\det(\lambda I_p - A) = 0. \tag{II.5}$$

[1]Note that solving the equation in (II.5) amounts to find the roots of a p^{th} order polynomial, and that the roots λ may be real valued or complex valued, that is $\lambda \in \mathbb{R}$ or $\lambda \in \mathbb{C}$. If a root is complex valued one can write it as, $\lambda = a + \mathbf{i}b$, where $a, b \in \mathbb{R}$ while $\mathbf{i}^2 = -1$. An implication is that the absolute value of λ is simple to compute by,

$$|\lambda| = \sqrt{a^2 + b^2}.$$

We denote by γ_A the absolute value of the largest eigenvalue of A, that is

$$\gamma_A = \max_{i=1,\ldots,p} |\lambda_i|.$$

with the largest absolute value, $|\lambda|$, by γ_A.

Example II.1.1 *With,*

$$A = \begin{pmatrix} \rho_1 & \rho_2 \\ 1 & 0 \end{pmatrix}, \ \lambda I_2 - A = \begin{pmatrix} \lambda - \rho_1 & -\rho_2 \\ -1 & \lambda \end{pmatrix},$$

and we find

$$\det(\lambda I_2 - A) = \lambda(\lambda - \rho_1) - \rho_2 = \lambda^2 - \lambda\rho_1 - \rho_2. \tag{II.6}$$

Hence the two eigenvalues λ_1 and λ_2 of A are the two roots in this second order polynomial.

The main reason that we are interested in γ_A is that it plays a cruical role in linear as well as nonlinear time series when determining whether a time series with general lag structure is stationary.

Example II.1.2 *Consider the AR(2) process,*

$$x_t = \rho_1 x_{t-1} + \rho_2 x_{t-2} + \varepsilon_t$$

with ε_t i.i.d.N(0,σ^2). Recall from the well-known theory of VAR processes, that x_t has a stationary solution if the roots in the characteristic polynomial for x_t, written here as,

$$\lambda^2 - \lambda\rho_1 - \rho_2 = 0, \tag{II.7}$$

are smaller than one in absolute value.

Next, similar to the ARCH(2), consider the companion form of the AR(2) process with $X_t = (x_t, x_{t-1})'$, such that

$$X_t = \begin{pmatrix} \rho_1 & \rho_2 \\ 1 & 0 \end{pmatrix} X_{t-1} + \begin{pmatrix} \varepsilon_t \\ 0 \end{pmatrix} = AX_{t-1} + \begin{pmatrix} \varepsilon_t \\ 0 \end{pmatrix}. \tag{II.8}$$

From Example II.1.1, we see that it is the size of the eigenvalues of A, γ_A that defines the dynamics of X_t and hence of x_t. That is the roots of the characteristic polynomial, see (II.7) and (II.6).

Analogous to the ARCH(2) case, we also see that X_t satisfies Assumption II.1.1, see (II.4) where $f(X_t|X_{t-2}) = f(x_t|x_{t-1}, x_{t-2}) f(x_{t-1}|x_{t-2}, x_{t-3})$ with,

$$f(x_t|x_{t-1}, x_{t-2}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x_t - \rho_1 x_{t-1} - \rho_2 x_{t-2})^2\right).$$

Thus we have seen that the ARCH(2) and AR(2) processes satisfy Assumption II.1.1 and that the largest value of the eigenvalues of the companion form matrix for the AR(2) process determines the stationarity properties.

We next proceed to illustrate that the conclusion can be reached by using the drift criterion in Assumption **??**, and that we may therefore also discuss mixing, moments and stationarity at the same time.

We shall need a final result relating eigenvalues of a matrix A with the size, or norm $|A|$, of the matrix.

Definition II.1.2 *Define a matrix norm* $|A|$ *by*

$$|A| = \sqrt{\gamma_{A'A}}.$$

With this definition of matrix norm we have two key implications:

Lemma II.1.1 *With* A *a* $p \times p$ *dimensional matrix, then*

$$|A^m|^{1/m} \to \gamma_A,$$

as $m \to \infty$ *and where* γ_A *is defined in Definition II.1.1 and* $|A|$ *is defined in Definition II.1.2. Moreover, for any vector* x *of dimension* p,

$$|Ax| \leq |A||x|,$$

where for any vector $x = (x_1, x_2, ..., x_p)'$, $|x| = \sqrt{x_1^2 + ... + x_p^2}$.

Example II.1.3 *Consider* A, *given by*

$$A = \begin{pmatrix} 0 & a \\ 0 & 0 \end{pmatrix},$$

then as A *has two eigenvalues at* $\lambda = 0$, *the maximal* $\gamma_A = 0$. *This is independent of the magnitude of the entry* a, *such that indeed* a *could be huge, while still having small* γ_A. *Computing,* $\gamma_{A'A}$ *with*

$$A'A = \begin{pmatrix} 0 & 0 \\ a & 0 \end{pmatrix} \begin{pmatrix} 0 & a \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & a^2 \end{pmatrix},$$

we find $\gamma_{A'A} = a^2$ *is large if* a *is (in absolute value). Note that* $|a| = \sqrt{a^2} = \sqrt{\gamma_{A'A}}$.

Example II.1.4 *With A as in Example II.1.3, we have*

$$A = \begin{pmatrix} 0 & a \\ 0 & 0 \end{pmatrix}, \; A^2 = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} = A^3 = ... = A^m.$$

Thus while $|A| = \sqrt{a^2}$, after one iteration or multiplication, we have $|A^2| = \sqrt{\gamma_0} = 0 = \gamma_A$.

With $x = (x_1, x_2)'$,

$$Ax = \begin{pmatrix} ax_2 \\ 0 \end{pmatrix}, \; |Ax|^2 = a^2 x_2^2 = |A|^2 x_2^2 \leq |A|^2 \left(x_2^2 + x_1^2\right) = |A|^2 |x|^2.$$

Example II.1.5 *With A diagonal, we get*

$$A = \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix}, \quad \text{and } A^m = \begin{pmatrix} a^m & 0 \\ 0 & b^m \end{pmatrix}.$$

Suppose that $a > b \geq 0$ in which case $\gamma_A = a$. Next, we find

$$A'A = \begin{pmatrix} a^2 & 0 \\ 0 & b^2 \end{pmatrix} \text{ and } A^{m\prime} A^m = \begin{pmatrix} a^{2m} & 0 \\ 0 & b^{2m} \end{pmatrix}.$$

Hence $|A| = \sqrt{\gamma_{A'A}} = \sqrt{a^2} = a$, $|A^m| = a^m$ and hence for any m, $|A^m|^{1/m} = (a^m)^{1/m} = a = \gamma_A$.

Finally, observe that with $x = (x_1, x_2)'$, $Ax = (ax, bx)'$ such that

$$|Ax|^2 = a^2 x_1^2 + b^2 x_2^2 \leq a^2 \left(x_1^2 + x_2^2\right) = |A|^2 |x|^2.$$

We are now in position to use the drift criterion on the AR(2) process.

Example II.1.6 *Consider the AR(2) process in Example II.1.2 written on companion form,*

$$X_t = \begin{pmatrix} \rho_1 & \rho_2 \\ 1 & 0 \end{pmatrix} X_{t-1} + \begin{pmatrix} \varepsilon_t \\ 0 \end{pmatrix} = AX_{t-1} + e_t. \tag{II.9}$$

A natural drift function $\delta(Y_t)$ is given by

$$\begin{aligned} \delta(X_t) = \delta(x_t, x_{t-1}) &= 1 + x_{t-1}^2 + x_{t-2}^2 \\ &= 1 + X_t' X_t = 1 + |X_t|^2. \end{aligned}$$

We can then compute $E\left(\delta\left(X_t\right)|X_{t-1}\right)$ *as follows,*

$$\begin{aligned}
E(\delta(X_t)|X_{t-1}) &= 1 + E\left((AX_{t-1}+e_t)'\left(AX_{t-1}+e_t\right)|X_{t-1}\right) \\
&= 1 + X_{t-1}'A'AX_{t-1} + E\left(e_t'e_t\right) \\
&= 1 + X_{t-1}'A'AX_{t-1} + \sigma^2 \\
&= |AX_{t-1}|^2 + 1 + \sigma^2 \\
&\leq |A|^2|X_{t-1}|^2 + 1 + \sigma^2 \\
&= |A|^2\delta\left(X_{t-1}\right) + \left(1 + \sigma^2 - |A|^2\right),
\end{aligned}$$

where we have used Lemma II.1.1 for the inequality. Hence it appears that choosing $|A|^2 < 1$ *is an obvious choice for the drift criterion to be satistied. However,* $|A|^2 < 1$ *restricts the AR(2) process in a meaningless way. In particular, with* $\rho_1 = \rho_2 = 0$*, such that* $x_t = \varepsilon_t$ *and thus "as stationary as possible",* $|A|^2 = 1$ *which violates the condition.*

We thus conclude that the condition is too strong (actually meaningless). What should be considered instead is the recursion of the dynamic system for X_t*. More precisely, as in the univariate AR(1) case, we can do simple recursion in (II.9),*

$$\begin{aligned}
X_t &= AX_{t-1} + e_t = A^2X_{t-2} + Ae_{t-1} + e_t = \ldots \\
&= A^mX_{t-m} + \sum_{i=0}^{m-1} A^ie_{t-i}.
\end{aligned}$$

And we find,

$$E\left(\delta(X_t)|X_{t-m}\right) = 1 + X_{t-m}'A^{m\prime}A^mX_{t-1} + \tilde{\sigma}^2,$$

with $\tilde{\sigma}^2$ *proportional to* σ^2*. As before, we can use Lemma II.1.1,*

$$X_{t-m}'A^{m\prime}A^mX_{t-m} = |A^mX_{t-m}|^2 \leq |A^m|^2|X_{t-m}|^2,$$

such that the drift criterion is satisfied if $|A^m| < 1$*. Now we can apply Lemma II.1.1 from which it holds that,*

$$|A^m|^{1/m} \to \gamma_A.$$

We can therefore conclude that if $\gamma_A < 1$*, then also* $|A^m| < 1$ *for some* m *large enough. And the condition* $\gamma_A < 1$ *is exactly the well-known condition for stationarity of an AR(2) process, see Example II.1.2.*

II.1.2.1 The ARCH(2) process

Return to the ARCH(2) process as defined by (II.1),

$$x_t = \sigma_t z_t, \quad \sigma_t^2 = \sigma^2 + \alpha_1 x_{t-1}^2 + \alpha_2 x_{t-2}^2,$$

which has companion form for $X_t = (x_t, x_{t-1})'$ in (II.9).).

Lemma II.1.2 *With the ARCH(2) process given by (II.1), then $X_t = (x_t, x_{t-1})'$ satisfies Theorem* **??** *and has finite second order moments if either of the three equivalent conditions hold for $\alpha_1, \alpha_2 \geq 0$:*

(i) $\alpha_1 + \alpha_2 < 1$

(ii) $|\lambda^2 - \alpha_1 \lambda - \alpha_2| = 0 \Rightarrow |\lambda| < 1$

(iii) $\gamma_A < 1$ *with* $A = \begin{pmatrix} \alpha_1 & 1 \\ \alpha_2 & 0 \end{pmatrix}$

In particular, under either of (i), (ii) and (iii), x_0 and x_1 can be given an initial distribution such that x_t (and X_t) is stationary, weakly mixing and $Ex_t^2 < \infty$.

Proof of Lemma II.1.2:
Start by computing $E(\delta(X_t)|X_{t-m}) = E(\delta(X_t)|x_{t-m}, x_{t-m-1})$, for $m = 1$,

$$\begin{aligned} E(\delta(X_t)|X_{t-1}) &= E(1 + X_t' X_t | X_{t-1}) \\ &= 1 + E\left(\left(x_t^2 + x_{t-1}^2\right)|x_{t-1}, x_{t-2}\right) = E\left(\left(\sigma_t^2 z_t^2 + x_{t-1}^2\right)|x_{t-1}, x_{t-2}\right) \\ &= 1 + \sigma_t^2 + x_{t-1}^2 \\ &= 1 + \sigma^2 + (\alpha_1 + 1) x_{t-1}^2 + \alpha_2 x_{t-2}^2. \end{aligned} \tag{II.10}$$

Next, using Lemma **??** for conditional expectations, and (II.10), we see that for $m = 2$ we get,

$$\begin{aligned} E(\delta(X_t)|X_{t-2}) &= E(E(\delta(X_t)|X_{t-1}, X_{t-2})|X_{t-2}) \\ &= E(E(\delta(X_t)|X_{t-1})|X_{t-2}) \\ &= E\left(1 + \sigma^2 + (\alpha_1 + 1) x_{t-1}^2 + \alpha_2 x_{t-2}^2 | x_{t-2}, x_{t-3}\right) \\ &= 1 + \sigma^2 + (\alpha_1 + 1)\sigma_{t-1}^2 + \alpha_2 x_{t-2}^2, \end{aligned}$$

such that

$$E(\delta(X_t)|X_{t-2}) = 1 + \sigma^2 + (\alpha_1 + 1)\sigma^2 + ((\alpha_1 + 1)\alpha_1 + \alpha_2) x_{t-2}^2 + (\alpha_1 + 1)\alpha_2 x_{t-3}^2. \tag{II.11}$$

And, likewise for $m = 3$[2],

$$\begin{aligned} E\left(\delta\left(X_t\right)|X_{t-3}\right) &= E\left(E\left(\delta\left(X_t\right)|X_{t-2}\right)|X_{t-3}\right) \\ &= 1+\sigma^2+\left(\alpha_1+1\right)\sigma^2+\left(\left(\alpha_1+1\right)\alpha_1+\alpha_2\right)\sigma^2 \\ &+\left(\left(\left(\alpha_1+1\right)\alpha_1+\alpha_2\right)\alpha_1+\left(\alpha_1+1\right)\alpha_2\right)x_{t-3}^2+\left(\left(\alpha_1+1\right)\alpha_1+\alpha_2\right)\alpha_2 x_{t-4}^2 \end{aligned} \tag{II.13}$$

To define the coefficients appearing in the conditional expectations above we define,

$$\beta_0 = 1,\ \beta_1 = \alpha_2, \quad \beta_2 = (1+\alpha_1)$$

and next for $m = 2, 3, ...$

$$\beta_{2m} = \alpha_1\beta_{2m-2} + \beta_{2m-3}, \qquad \text{and} \quad \beta_{2m-1} = \alpha_2\beta_{2m-2}. \tag{II.14}$$

With this definition, for example for $m = 3$ in (II.13) we find immediately,

$$E\left(\delta\left(X_t\right)|X_{t-3}\right) = 1+\beta_0\sigma^2+\beta_2\sigma^2+\beta_4\sigma^2+\beta_6 x_{t-3}^2+\beta_5 x_{t-4}^2,$$

and for general m,

$$E\left(\delta\left(X_t\right)|X_{t-m}\right) = E\left(\delta\left(X_t\right)|x_{t-m}, x_{t-m-1}\right) = \beta_{2m}x_{t-m}^2+\beta_{2m-1}x_{t-m-1}^2+\tilde{\sigma}_m^2 \tag{II.15}$$

where $\tilde{\sigma}_m^2 = \sum_{i=0}^{m-1}\beta_{2i}\sigma^2$.

We can write the right hand side of (II.15) as follows,

$$\begin{aligned} E\left(\delta\left(X_t\right)|X_{t-m}\right) &= X_{t-m}'\begin{pmatrix}\beta_{2m} & 0\\ 0 & \beta_{2m-1}\end{pmatrix}X_{t-m}+\tilde{\sigma}_m^2 \\ &= |BX_{t-m}|^2+\tilde{\sigma}_m^2 \\ &\leq |B|^2|X_{t-m}|^2+\tilde{\sigma}_m^2, \end{aligned}$$

[2] Also, for $m = 4$,

$$\begin{aligned} E\left(\delta\left(X_t\right)|X_{t-4}\right) &= E\left(E\left(\delta\left(X_t\right)|X_{t-3}\right)|X_{t-4}\right) \\ &= 1+\sigma^2+\left(\alpha_1+1\right)\sigma^2+\left(\left(\alpha_1+1\right)\alpha_1+\alpha_2\right)\sigma^2+ \\ &\left(\left(\left(\alpha_1+1\right)\alpha_1+\alpha_2\right)\alpha_1+\left(\alpha_1+1\right)\alpha_2\right)E\left(x_{t-3}^2|x_{t-4}, x_{t-5}\right)+ \\ &\left(\left(\alpha_1+1\right)\alpha_1+\alpha_2\right)\alpha_2 x_{t-4}^2 \\ &= 1+\sigma^2+\left(\alpha_1+1\right)\sigma^2+\left(\left(\alpha_1+1\right)\alpha_1+\alpha_2\right)\sigma^2+ \\ &\left(\left(\left(\alpha_1+1\right)\alpha_1+\alpha_2\right)\alpha_1+\left(\alpha_1+1\right)\alpha_2\right)\sigma^2+ \\ &\left(\left(\left(\left(\alpha_1+1\right)\alpha_1+\alpha_2\right)\alpha_1+\left(\alpha_1+1\right)\alpha_2\right)\alpha_1+\left(\left(\alpha_1+1\right)\alpha_1+\alpha_2\right)\alpha_2\right)x_{t-4}^2 \\ &\left(\left(\left(\alpha_1+1\right)\alpha_1+\alpha_2\right)\alpha_1+\left(\alpha_1+1\right)\alpha_2\right)\alpha_2 x_{t-5}^2 \end{aligned} \tag{II.12}$$

where

$$B = \begin{pmatrix} \sqrt{\beta_{2m}} & 0 \\ 0 & \sqrt{\beta_{2m-1}} \end{pmatrix}.$$

Thus as in Example II.1.6 we need B to have eigenvalues smaller than 1. But the eigenvalues of B are simply $\lambda_1 = \sqrt{\beta_{2m}}$ and $\lambda_2 = \sqrt{\beta_{2m-1}}$, so what we need is $\beta_{2m} < 1$ and $\beta_{2m-1} < 1$.

To see that this holds, use the definition of the coefficients β_{2m} and β_{2m-1} in (II.14), and write these as,

$$v_m = \begin{pmatrix} \beta_{2m} \\ \beta_{2m-1} \end{pmatrix} = A \begin{pmatrix} \beta_{2m-2} \\ \beta_{2m-3} \end{pmatrix} = Av_{m-1}, \text{ with } A = \begin{pmatrix} \alpha_1 & 1 \\ \alpha_2 & 0 \end{pmatrix}.$$

Thus with $\gamma_A < 1$, then by Lemma II.1.1, $|A^m| < 1$ for m large and hence as

$$v_m = A^m e, \quad \text{with } e = (1,1)'$$

we get

$$|v_m|^2 = v_m' v_m = e' A^{m\prime} A^m e = |A^m e|^2 \leq |A^m|^2 |e|^2 = 2|A^m|^2 < 1 \text{ for } m \text{ large.}$$

So the v_m elements, that is the diagonal elements of B, are indeed smaller than 1 in absolute value as desired.

II.2 ARCH(k)

The ARCH(k) process can for $t = k, k+1, k+2 \ldots$ be represented as

$$x_t = \sigma_t z_t \tag{II.16}$$
$$\sigma_t^2 = \sigma^2 + \alpha_1 x_{t-1}^2 + ... + \alpha_k x_{t-k}^2 \tag{II.17}$$

with initial values $x_0, x_1, ..., x_{k-1}$ and where the z_t's are *i.i.d.*N(0,1). As in the ARCH(2) case, we can use the drift criterion to establish that the ARCH(k) process with $X_t = (x_t, x_{t-1}, ..., x_{t-k-1})'$ satisfies Theorem **??** and has finite second order moments if, $\alpha_i \geq 0$ and

$$\left|\lambda^k - \alpha_1 \lambda^{k-1} - ... - \alpha_k\right| = 0 \Rightarrow |\lambda| < 1. \tag{II.18}$$

This is equivalent to the condition that,

$$\alpha_1 + \alpha_2 + ... + \alpha_k < 1.$$

II.2.1 Multivariate ARCH(k)

Consider one natural extension of the univariate ARCH process to the $p-$dimensional process as given by

$$X_t = \Omega_t^{1/2} Z_t, \tag{II.19}$$
$$\Omega_t = \Omega + \sum_{i=1}^{k} A_i X_{t-i} X_{t-i}' A_i',$$

where $\Omega > 0$, (A_i) are any $p \times p$ matrices and Z_t $i.i.d. N_p(0, I_p)$. This is a simple example of the so-called BEKK process in Engle and Kroner (1995). In the scalar-ARCH the A_i matrices are replaced by scalars $\alpha_i (= A_i^2)$ which is sometimes applied in high-dimensional, or "vast", systems.

Other extensions include that at each lag $i = 1, ..., k$ further loadings may be added, such that with $s \leq p$, one may use the general ARCH,

$$X_t = \Omega_t^{1/2} Z_t, \tag{II.20}$$
$$\Omega_t = \Omega + \sum_{i=1}^{k} \sum_{j=1}^{s} A_{ij} X_{t-i} X_{t-i}' A_{ij}',$$

Here the summation over j for fixed i, allows for a more rich feed-back mechanism from X_{t-i} which in the univariate case vanishes.

Mimicking the just given calculations for the univariate ARCH(2) it follows that if

$$\left| I_{p^2} \lambda^k - \sum_{j=1}^{s} (A_{1j} \otimes A_{1j}) \lambda^{k-1} - ... \sum_{j=1}^{s} (A_{kj} \otimes A_{kj}) \right| = 0 \Rightarrow |\lambda| < 1,$$

holds, then X_t satisfies Theorem **??** with $\delta(X_t) = 1 + |X_t|^2$, has second order moments and can be given an initial distribution such that it becomes stationary.

Note that with A a $p \times p$ matrix,

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1p} \\ \vdots & & \vdots \\ a_{p1} & \cdots & a_{pp} \end{pmatrix},$$

then $A \otimes A$ is the $(p^2 \times p^2)-$dimensional matrix defined by,

$$A \otimes A = \begin{pmatrix} a_{11}A & \cdots & a_{1p}A \\ \vdots & & \vdots \\ a_{p1}A & \cdots & a_{pp}A \end{pmatrix}.$$

II.3 GARCH(1,1)

The generalized ARCH process, the GARCH(1,1) is probably the most used of all ARCH processes. The GARCH(1,1), or simply the GARCH, process for $t = 1, 2, ...$ can be represented as

$$x_t = \sigma_t z_t \tag{II.21}$$
$$\sigma_t^2 = \sigma^2 + \alpha x_{t-1}^2 + \beta \sigma_{t-1}^2, \tag{II.22}$$

with initial values x_0 and σ_0^2, and where the z_t's are *i.i.d.*N(0,1). Moreover, $\sigma^2 > 0$, $\alpha \geq 0$ and $\beta \geq 0$.

Loosely speaking, the GARCH process can be motivated by a few recursions:

$$\begin{aligned}
\sigma_t^2 &= \sigma^2 + \alpha x_{t-1}^2 + \beta \sigma_{t-1}^2 \\
&= \sigma^2 + \alpha x_{t-1}^2 + \beta \left(\sigma^2 + \alpha x_{t-2}^2 + \beta \sigma_{t-2}^2 \right) \\
&= (1 + \beta) \sigma^2 + \alpha x_{t-1}^2 + \beta \alpha x_{t-2}^2 + \beta^2 \sigma_{t-3}^2 \\
&= \left(1 + \beta + \beta^2 \right) \sigma^2 + \alpha x_{t-1}^2 + \beta \alpha x_{t-2}^2 + \beta^2 \alpha x_{t-3}^2 + \beta^3 \sigma_{t-3}^3 \\
&= ... \\
&= \sum_{i=0}^{t-1} \beta^i \sigma^2 + \sum_{i=0}^{t-1} \alpha \beta^i x_{t-1-i}^2 + \beta^t \sigma_0^2
\end{aligned}$$

Hence, it may be viewed as a way in just two parameters, α and β, to allow for an increasing (in t) number of lagged x_t^2 to enter the ARCH specification.

Now x_t conditional on the past information $(x_{t-1}, ..., x_0)$ and σ_0^2 is not a function of x_{t-1} but a function of the entire past $(x_{t-1}, x_{t-2}, ..., x_0)$ and also σ_0^2. So it is not a Markov chain. However, by setting for example $X_t = (x_t, \sigma_t)'$ then indeed X_t given past X_t is a function of X_{t-1}, and thus satisfying Assumption II.1.1 (i). Unfortunately defining X_t this way, Assumption II.1.1 is not satisfied. Instead the following approach can be applied, see Carrasco and Chen (2002) for example. Rewrite σ_t^2 as follows,

$$\begin{aligned}
\sigma_t^2 &= \sigma^2 + \alpha \sigma_{t-1}^2 z_{t-1}^2 + \beta \sigma_{t-1}^2 \\
&= \sigma^2 + \left(\alpha z_{t-1}^2 + \beta \right) \sigma_{t-1}^2.
\end{aligned}$$

Then we can apply the drift criterion to σ_t and next use, $x_t = \sqrt{\sigma_t^2} z_t$ to conclude that x_t inherits the properties of σ_t^2. More precisely, cf. Carrasco and Chen (2002, Proposition 5) (Meitz and Saikkonen, 2008, discuss modifications of Proposition 5 in Carrasco and Chen, 2002), if α and β are such that Theorem **??** holds for a drift function $\delta(\sigma_t^2)$, then σ_t^2 can be given an

initial distribution such that σ_t^2 and hence x_t are stationary and weakly mixing. Moreover, as $E\delta(\sigma_t^2) < \infty$, then if $\delta(\sigma_t^2) = 1 + (\sigma_t^2)^s$, $s > 0$, we also have $Ex_t^{2s} < \infty$.

Consider the simple example of $\delta(\sigma_t^2) = 1 + \sigma_t^2$. Then

$$\begin{aligned} E\left(\delta\left(\sigma_t^2\right)|\sigma_{t-1}^2\right) &= \sigma^2 + \alpha\sigma_{t-1}^2 + \beta\sigma_{t-1}^2 \\ &= (\alpha+\beta)\,\delta\left(\sigma_{t-1}^2\right) + \left(\sigma^2 - (\alpha+\beta)\right). \end{aligned}$$

Hence we immediately conclude that if $(\alpha+\beta) < 1$, then $Ex_t^2 < \infty$ and x_t and σ_t^2 are weakly mixing.

More generally, see also Nelson (1990, Theorem 3), we have the following similar to the ARCH(1) process:

GARCH(1,1) process x_t defined in (II.21):	
$x_t = \sigma_t z_t,\ \ \sigma_t = \sigma^2 + \alpha x_{t-1}^2 + \beta\sigma_{t-1}^2$ and z_t *i.i.d.*N$(0,1)$.	
Stationary for $E\log(\alpha z_t^2 + \beta) < 0$	
Finite moments:	
$E\log(\alpha z_t^2 + \beta) < 0$	no moments (fractional)
$E(\alpha z_t^2 + \beta)^{1/2} < 1$	$E\lvert x_t\rvert < \infty$
$E(\alpha z_t^2 + \beta) < 1$ or $(\alpha+\beta) < 1$	$Ex_t^2 < \infty$
$E(\alpha z_t^2 + \beta)^2$ or $\beta^2 + 3\alpha^2 + 2\alpha\beta < 1$	$Ex_t^4 < \infty$

In general, the condition is that $E\left(\sigma_t^{2k}\right) < \infty$ if and only if $E(\alpha z_t^2 + \beta)^k < 1$, see Nelson (1990) and Carrasco and Chen (2002). Explicit expressions for $E\log(\alpha z_t^2 + \beta)$ and $E(\alpha z_t^2 + \beta)^k$ in terms of α, β and so-called hypergeometric functions are given in Nelson (1990). Note also that the regions change if z_t is assumed to be *i.i.d.* t_v distributed, for example.

Lastly, an example of a multivariate GARCH is the simple BEKK-GARCH(1,1) process given by,

$$\Omega_t = \Omega + AX_{t-1}X_{t-1}'A' + B\Omega_{t-1}B',$$

with A and B $(p \times p)$-dimensional matrices.

References

Brown, B.M. (1971), Martingale Central Limit Theorems, The Annals of Mathematical Statistics 42:59-66.

Engle and Kroner,(1995), Multivariate Simultaneous Generalized ARCH,. Econometric Theory

Meitz and Saikkonen, (2008), Ergodicity, Mixing, and Existence of Moments of a Class of Markov Models with Applications to GARCH and ACD Models, Econometric Theory.

Meyn, S.P. and R.L. Tweedie, 1993, Markov chains and stochastic stability, Communications and Control Engineering Series, Springer-Verlag, London ltd., London

Nelson, D., 1990, Stationarity and persistence in the GARCH(1,1) model, Econometric Theory, 6, 318-334

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,

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

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 $\hat{\theta}_{(\text{q})\text{mle}} = \hat{\theta}$ 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,*

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

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,

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

provided that $i_T(\theta^*)$ 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*

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

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 $|\rho_0| < 1$,

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

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*

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

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

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

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,*

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

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,

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

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

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

Then we get

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

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,

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

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

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

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

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

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,

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

The condition (A.3) implies that the term $\frac{1}{T}\partial^3 \ell_T\left(\theta^*\right)/\partial\theta^3\left(\hat{\theta}-\theta_0\right) \xrightarrow{P} 0$ 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 $\theta_0 = \left(\sigma_0^2, \alpha_0\right)'$ can be chosen as,

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

where

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

Likewise, for $\theta = \left(\theta_1, ..., \theta_d\right)' \in \Theta$, 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,

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

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,

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

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,

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

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

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

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 $N(\theta_0)$ as defined in (III.15),

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

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:

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

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

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,

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

Hence for the Gaussian MLE, one can use that

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

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 $i.i.d.N(0,1)$, then $\hat{\theta} = (\hat{\sigma}^2, \hat{\alpha})'$ 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

$$\frac{1}{T} i_T(\hat{\theta}) = -\frac{1}{T} \frac{\partial^2}{\partial\theta\partial\theta'} \ell_T(\theta) \mid_{\theta=\hat{\theta}} \xrightarrow{P} \Omega_I.$$

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

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

Then the score equals,

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

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,

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

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,

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

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,

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

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

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

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,*

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

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

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

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,

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

and hence

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

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,

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

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

[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 $E z_t^4 < \infty$,

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

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

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

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,

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

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

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

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

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

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),

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

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,

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

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

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

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,

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

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,

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

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,

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

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 $E\log(\alpha_0 z_t^2) < 0$ 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 $\delta(X) = 1+|X|^2$ 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,

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

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,

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

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.*$\mathrm{N}(0,1)$.

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,

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

and Ω_S consistently by

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

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

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

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

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

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

III.6.4.2 On weakly mixing

Recall from Part II, that if

$$E\log\left(\beta_0+\alpha_0z_t^2\right)<0,$$

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 $\frac{\partial}{\partial\theta}\sigma_0^2(\theta)=0$, and the recursions give directly,

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

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

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

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),

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

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,

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

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

(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.

Anders Rahbek
Rasmus Søndergaard Pedersen
University of Copenhagen

October 2021

Part IV
Risk management with GARCH models

In this chapter we consider applications of GARCH models in relation to risk management. We define the *Value-at-Risk* and *Expected Shortfall* risk measures. The measures are important, for instance, from a regulatory perspective, as banks are obliged to disclose their estimates of such risk measures in relation to their holdings of risky assets (Basel Committee on Banking Supervision, 2013). We discuss how these measures are computed and estimated in the case that returns are *i.i.d.* Gaussian or are generated from (G)ARCH processes. One may note that the computation of the risk measures boils down to computing a quantile (Value-at-Risk) or a conditional expectation (Expected Shortfall) of some (conditional) distribution. In the *i.i.d.* Gaussian case these distributions are known, which makes the computations of the risk measures easy in the sense that we obtain closed-form expressions. For the GARCH models, on the other hand, conditional distributions are (in general) intractable, and we discuss how one may circumvent this issue by means of simulation-based estimation. We also discuss how the estimation uncertainty of the estimated risk measures is addressed.

IV.1 Value-at-Risk (VaR)

As in the previous chapters, we let x_{t+1} denote the log-return of some asset from t to $t+1$. We shall also need the h-period return, $h \geq 1$, which by definition is given by

$$x_{t+1,h} = \sum_{i=1}^{h} x_{t+i}.$$

A typical way of quantifying the risk of holding an asset over one period is to compute the so-called Value-at-Risk (VaR). Specifically, *the* 1*-period* VaR

at risk level $\kappa \in (0,1)$ (or, in short, the VaR) is denoted VaR_t^κ and satisfies

$$P(x_{t+1} < -\text{VaR}_t^\kappa | \mathcal{I}_t) = \kappa, \quad \text{VaR}_t^\kappa \in \mathcal{I}_t, \tag{IV.1}$$

where $\mathcal{I}_t$ denotes some information set available at time t (e.g. the series of previous returns). Note that $P(-x_{t+1} \leq \text{VaR}_t^\kappa | \mathcal{I}_t) = 1 - \kappa$, so that the VaR measures the maximum loss $(-x_{t+1})$ not exceeded with probability $1-\kappa$, or equivalently,VaR is the $1-\kappa$ percentile of the conditional loss distribution.[1] Note that, by construction, the VaR depends on the return process, the information set $\mathcal{I}_t$, as well as the *confidence level* $1-\kappa$. Typical values of κ in applications are 1%, 2.5%, and 5%.

In the following we consider some examples of VaR computation. Throughout, we assume that the information set contains only past values of the returns, i.e. $\mathcal{I}_t = \{r_i : i \leq t\}$. We emphasize that one could include additional variables to the information set, which would lead to careful considerations, and assumptions, about how these variables are related to the return process. We first consider the VaR in the case where the returns are driven by an *i.i.d.* Gaussian process.

Example IV.1.1 *(*VaR*: The Gaussian case I) Suppose that* $x_t \sim i.i.d.N(0,1)$*. Then, with* $\Phi(\cdot)$ *the cdf of the standard normal distribution and using that* x_{t+1} *is independent of* $\mathcal{I}_t$*,*

$$P(x_{t+1} < -\text{VaR}_t^\kappa | \mathcal{I}_t) = \Phi(-\text{VaR}_t^\kappa) = \kappa$$

Hence,

$$\text{VaR}_t^\kappa = -\Phi^{-1}(\kappa),$$

i.e. the VaR *is (negative) the* κ *percentile of the standard normal distribution.*

Since the returns are *i.i.d.* Gaussian the previous example is straightforward to extend to an arbitrary horizon h and volatility σ. Define initially the h-period VaR as

$$P\left(x_{t+1,h} < -\text{VaR}_{t,h}^\kappa | \mathcal{I}_t\right) = \kappa.$$

[1] Note that the definition above implicitly assumes that the VaR exists, which is indeed the case whenever the conditional return distribution is continuous. A more general definition that ensures that the VaR always exists is that $\text{VaR}_t^\kappa = \inf\{y \in \mathbb{R} : P(-x_{t+1} \leq y | \mathcal{I}_t) \geq 1-\kappa\}$. Some textbooks, such as the one by Francq and Zakoïan (2019), make the convention that the VaR must be non-negative, such that the VaR is given by $\max[0, \inf\{y \in \mathbb{R} : P(-x_{t+1} \leq y | \mathcal{I}_t) \geq 1-\kappa\}]$.

Example IV.1.2 *(*VaR*: The Gaussian case II) Suppose that* $x_t \sim i.i.d.N(0, \sigma^2)$*. It follows that* $x_{t+1,h} = \sum_{i=1}^{h} x_{t+i} \overset{D}{=} \sigma\sqrt{h}z$ *where* $z \sim N(0,1)$ *and independent of* $\mathcal{I}_t$*. Hence,*

$$\begin{aligned} P\left(x_{t+1,h} < -\text{VaR}_{t,h}^{\kappa} | \mathcal{I}_t\right) &= P\left(\sigma\sqrt{h}z < -\text{VaR}_{t,h}^{\kappa}\right) \\ &= P\left(z < -\text{VaR}_{t,h}^{\kappa}/(\sigma\sqrt{h})\right) \\ &= \Phi\left(-\text{VaR}_{t,h}^{\kappa}/(\sigma\sqrt{h})\right) = \kappa, \end{aligned}$$

such that

$$\text{VaR}_{t,h}^{\kappa} = -\sigma\sqrt{h}\Phi^{-1}(\kappa).$$

In practice, the volatility σ is typically unknown but may be estimated based on a sample of returns; see Section IV.3.

IV.2 Expected Shortfall (ES)

For several years, the VaR was viewed as an industry standard for quantifying the risk of asset portfolios. Unfortunately, VaR has some shortcomings.

In particular, VaR lacks the property of being subadditive: Let $x_{t+1}^{(1)}$ and $x_{t+1}^{(2)}$ denote the returns of two assets (Asset 1 and 2), and let $\text{VaR}_t^{\kappa}(x_{t+1}^{(1)})$ and $\text{VaR}_t^{\kappa}(x_{t+1}^{(2)})$ denote their respective VaR. Then it does not necessarily hold that

$$\text{VaR}_t^{\kappa}(x_{t+1}^{(1)} + x_{t+1}^{(2)}) \leq \text{VaR}_t^{\kappa}(x_{t+1}^{(1)}) + \text{VaR}_t^{\kappa}(x_{t+1}^{(2)}).$$

For instance, for certain distributions of $x_{t+1}^{(1)}$ and $x_{t+1}^{(2)}$, with $x_{t+1}^{(1)}$ and $x_{t+1}^{(2)}$ independent and identically distributed, it holds that

$$\text{VaR}_t^{\kappa}(x_{t+1}^{(1)}+x_{t+1}^{(2)}) > \text{VaR}_t^{\kappa}(x_{t+1}^{(1)})+\text{VaR}_t^{\kappa}(x_{t+1}^{(2)}) = 2\text{VaR}_t^{\kappa}(x_{t+1}^{(1)}) = \text{VaR}_t^{\kappa}(2x_{t+1}^{(1)}),$$

that is, one can reduce the overall risk (as measured by VaR) by holding two quantities of Asset 1 instead of holding one of each asset, even though the losses of the assets are independent. Hence, one may argue that the use of VaR may discourage diversification (see e.g. Ibragimov, 2009).

Moreover, recall that VaR is the maximum loss not exceeded with a given probability $1-\kappa$. The risk measure does not tell us how much we lose (or may expect to lose) given that the loss exceeds the VaR, and hence the VaR does not tell us anything about the risk exposure in the presence of a "tail-event" (that happens with probability κ). This has led to the so-called Expected Shortfall (ES) risk measure that, by definition, quantifies the expected loss

given that the loss exceeds the VaR: *The* 1*-period* ES *at risk level* $\kappa \in (0,1)$ is given by

$$\mathrm{ES}_t^\kappa = E[-x_{t+1}|x_{t+1} < -\mathrm{VaR}_t^\kappa, \mathcal{I}_t].$$

It follows that

$$\mathrm{ES}_t^\kappa = \kappa^{-1} \int_0^\kappa \mathrm{VaR}_t^u du,$$

and that $\mathrm{ES}_t^\kappa \geq \mathrm{VaR}_t^\kappa$.

Example IV.2.1 *(*ES*: The Gaussian case I) Suppose that* $x_t \sim i.i.d.N(0,1)$*. Recall that* $E[X|X \in A] = E[X\mathbb{I}\{X \in A\}]/P(X \in A)$ *(given that* $P(X \in A) > 0$*). Then*

$$\mathrm{ES}_t^\kappa = E[-x_{t+1}|x_{t+1} < -\mathrm{VaR}_t^\kappa, \mathcal{I}_t] = \frac{E[-x_{t+1}\mathbb{I}\{x_{t+1} < -\mathrm{VaR}_t^\kappa\}]}{P(x_{t+1} < -\mathrm{VaR}_t^\kappa)}.$$

Using that $P(x_{t+1} < -\mathrm{VaR}_t^\kappa) = \kappa$*, we have that*

$$\mathrm{ES}_t^\kappa = \kappa^{-1} \int_{\mathrm{VaR}_t^\kappa}^\infty x\phi(x)dx = \kappa^{-1} \int_{-\Phi^{-1}(\kappa)}^\infty x\phi(x)dx,$$

where $\phi(x) \equiv \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$ *is the pdf of the standard normal distribution. It holds that*

$$\frac{d}{dx}\phi(x) = -x\phi(x),$$

such that

$$\int x\phi(x)dx \quad \left(= -\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)\right) = -\phi(x).$$

Hence

$$\begin{aligned}\int_{-\Phi^{-1}(\kappa)}^\infty x\phi(x)dx &= [-\phi(y)]_{-\Phi^{-1}(\kappa)}^\infty \\ &= [\lim_{y\to\infty} -\phi(y)] - [-\phi(-\Phi^{-1}(\kappa))] \\ &= 0 + \phi(-\Phi^{-1}(\kappa)),\end{aligned}$$

and we have that

$$\mathrm{ES}_t^\kappa = \kappa^{-1}\phi(-\Phi^{-1}(\kappa)).$$

Example IV.2.2 *(*ES*: The Gaussian case II) Suppose that* $x_t \sim i.i.d.N(0,\sigma^2)$. *Let* $z \sim N(0,1)$ *and note that* $x_{t+1} \stackrel{d}{=} \sigma z$. *Then*

$$\begin{aligned}\mathrm{ES}_t^\kappa &= E[-x_{t+1}|x_{t+1} < -\mathrm{VaR}_t^\kappa, \mathcal{I}_t] \\ &= E[-\sigma z|\sigma z < \sigma\Phi^{-1}(\kappa)] \\ &= \sigma E[(-z)|z < \Phi^{-1}(\kappa)] \\ &= \sigma\kappa^{-1}\phi(-\Phi^{-1}(\kappa)),\end{aligned}$$

where we have used the arguments from the previous example. Likewise, for the h*-period* ES, *defined as*

$$\mathrm{ES}_{t,h}^\kappa = E[-x_{t+1,h}|x_{t+1,h} < -\mathrm{VaR}_{t,h}^\kappa, \mathcal{I}_t]$$

we have that

$$\mathrm{ES}_{t,h}^\kappa = \sigma\sqrt{h}\kappa^{-1}\phi(-\Phi^{-1}(\kappa)).$$

IV.3 VaR and ES for GARCH processes

By construction, the estimator of the volatility is subject to estimation uncertainty, which in turn implies that the estimated VaR and ES are subject to estimation uncertainty. Before turning to the GARCH case, consider initially the i.i.d. case.

IV.3.1 Estimation uncertainty

Given a sample $(x_t : t = 1,\ldots,T)$, with $x_t \sim i.i.d.N(0,\sigma^2)$, one may use the estimator $\hat{\sigma}_T = \sqrt{T^{-1}\sum_{t=1}^T x_t^2}$ for σ, in order to obtain an estimator of the h-period VaR:

$$\widehat{\mathrm{VaR}_{t,h}^\kappa} = -\hat{\sigma}_T\sqrt{h}\Phi^{-1}(\kappa).$$

By the LLN for i.i.d. processes, $\hat{\sigma}_T^2 \xrightarrow{p} \sigma^2$ as $T \to \infty$, and hence $\widehat{\mathrm{VaR}_{t,h}^\kappa}$ is consistent for $\mathrm{VaR}_{t,h}^\kappa$, i.e.

$$\widehat{\mathrm{VaR}_{t,h}^\kappa} \xrightarrow{p} \mathrm{VaR}_{t,h}^\kappa.$$

Next, in order to quantify the estimation uncertainty, we note that from the CLT for i.i.d. processes, $\sqrt{T}(\hat{\sigma}_T^2 - \sigma^2) \xrightarrow{d} N(0,\Sigma)$ as $T \to \infty$, where $\Sigma \equiv \mathrm{V}(r_t^2 - \sigma^2) = 2\sigma^4$. Moreover, since $x \mapsto \sqrt{x}$ is continuously differentiable

on the positive real axis, we have that $\sqrt{T}(\hat{\sigma}_T - \sigma) \xrightarrow{d} N(0, \sigma^2/2)$ [by the Δ-method]. Hence,

$$\sqrt{T}\left(\widehat{\mathrm{VaR}_{t,h}^{\kappa}} - \mathrm{VaR}_{t,h}^{\kappa}\right) = -\sqrt{h}\Phi^{-1}(\kappa)\sqrt{T}(\hat{\sigma}_T - \sigma) \xrightarrow{d} N(0, h\sigma^2\Phi^{-1}(\kappa)^2/2),$$

such that

$$\widehat{\mathrm{VaR}_{t,h}^{\kappa}} \overset{a}{\sim} N(\mathrm{VaR}_{t,h}^{\kappa}, h\sigma^2\Phi^{-1}(\kappa)^2/(2T)),$$

and one may report the (approximate) 95% error bands of the VaR as $\widehat{\mathrm{VaR}_{t,h}^{\kappa}} \pm 1.96\sqrt{h/(2T)}|\Phi^{-1}(\kappa)|\hat{\sigma}_T$. In order to take into account the estimation uncertainty, or the additional "estimation risk", one may for instance use the upper band,

$$\widehat{\mathrm{VaR}_{t,h}^{\kappa}} + 1.96\sqrt{h/(2T)}|\Phi^{-1}(\kappa)|\hat{\sigma}_T,$$

as the "estimation risk-adjusted VaR measure".
As for the VaR, we may obtain an ES estimate as

$$\widehat{\mathrm{ES}}_{t,h}^{\kappa} = \kappa^{-1}\hat{\sigma}_T\sqrt{h}\phi(-\Phi^{-1}(\kappa)),$$

and, as $T \to \infty$, we have that

$$\widehat{\mathrm{ES}_{t,h}} \xrightarrow{p} \mathrm{ES}_{t,h}^{\kappa},$$

and

$$\sqrt{T}\left(\widehat{\mathrm{ES}_{t,h}} - \mathrm{ES}_{t,h}^{\kappa}\right) \xrightarrow{d} N(0, \kappa^{-2}h\phi(-\Phi^{-1}(\kappa))^2\sigma^2/2).$$

IV.3.2 GARCH

Suppose that the returns are driven by a GARCH process such that

$$x_t = \sigma_t z_t, \quad z_t \sim i.i.d.(0,1),$$

and $\sigma_t^2 > 0$ some function of past returns.

Example IV.3.1 *(*VaR*: ARCH(1), Gaussian errors) Suppose that* $z_t \sim N(0,1)$*, and* $\sigma_t^2 = \omega + \alpha x_{t-1}^2$*. Then, similar to Example IV.1.2,*

$$\begin{aligned} P\left(x_{t+1} < -\mathrm{VaR}_t^{\kappa}|\mathcal{I}_t\right) &= P\left(\sigma_{t+1}z_{t+1} < -\mathrm{VaR}_{t,1}^{\kappa}|\mathcal{I}_t\right) \\ &= P\left(z_{t+1} < -\mathrm{VaR}_{t,1}^{\kappa}/\sigma_{t+1}|\mathcal{I}_t\right) \\ &= \Phi(-\mathrm{VaR}_{t,1}^{\kappa}/\sigma_{t+1}), \end{aligned}$$

where we have used that $\sigma_{t+1} \in \mathcal{I}_t$ and that z_{t+1} is independent of $\mathcal{I}_t$. We hence obtain that

$$\mathrm{VaR}_t^\kappa = -\sigma_{t+1}\Phi^{-1}(\kappa).$$

As discussed in Part III, we may obtain an estimator for the parameters $\theta = (\omega, \alpha)'$ by quasi-maximum likelihood given a sample of returns, $\hat{\theta}_T = (\hat{\omega}_T, \hat{\alpha}_T)'$. Based on this estimator we obtain an estimator for the conditional volatlity, given by

$$\hat{\sigma}_{t+1} = \sqrt{\hat{\omega}_T + \hat{\alpha}_T x_t^2},$$

and we have the estimated VaR*,*

$$\widehat{\mathrm{VaR}_t^\kappa} = -\hat{\sigma}_{t+1}\Phi^{-1}(\kappa).$$

Notice that (unlike the case of Gaussian returns in the previous examples) VaR_t^κ is random as it depends on x_t. In order to analyze the statistical properties of the VaR *estimator, it is customary to consider x_t as fixed and setting it equal to some fixed value, $x_t = x$. We then have that*

$$\widehat{\mathrm{VaR}_t^\kappa} - \mathrm{VaR}_t^\kappa = -(\hat{\sigma}_{t+1} - \sigma_{t+1})\Phi^{-1}(\kappa),$$

with $\hat{\sigma}_{t+1}^2 = \hat{\omega}_T + \alpha_T x^2$ and $\sigma_{t+1}^2 = \omega + \alpha x^2$. Recall that under certain conditions discussed in Part III, as $T \to \infty$,

$$\hat{\theta}_T \xrightarrow{p} \theta \quad \textit{and} \quad \sqrt{T}(\hat{\theta}_T - \theta) \xrightarrow{d} N(0, \Omega_I^{-1}\Omega_S\Omega_I^{-1}).$$

This implies that, by a first-order Taylor expansion (up to a negligible remainder term),

$$\hat{\sigma}_{t+1} - \sigma_{t+1} = \frac{1}{2\sqrt{\omega + \alpha x^2}}(1, x^2)(\hat{\theta}_T - \theta),$$

and we conclude that

$$\widehat{\mathrm{VaR}_t^\kappa} \xrightarrow{p} \mathrm{VaR}_t^\kappa,$$

and

$$\sqrt{T}(\widehat{\mathrm{VaR}_t^\kappa} - \mathrm{VaR}_t^\kappa) \xrightarrow{d} N\left(0, \frac{\Phi^{-1}(\kappa)^2}{4(\omega + \alpha x^2)}(1, x^2)\Omega_I^{-1}\Omega_S\Omega_I^{-1}(1, x^2)'\right).$$

Similar to Example IV.3.1, one may use this result to construct error bands for the estimated VaR*.*

Example IV.3.2 *(*VaR*: ARCH(1), non-Gaussian errors) Suppose that z_t has some unknown distribution with cdf F_z. Then as in the previous example, we have that*

$$\mathrm{VaR}_t^\kappa = -\sigma_{t+1} F_z^{-1}(\kappa),$$

where F_z^{-1} is the (generalized) inverse of F_z. In addition to estimating σ_{t+1}, we need an estimate of $F_z^{-1}(\kappa)$. We may obtain such an estimate by computing the empirical κ percentile of the standardized residuals. Specifically, we let

$$\hat{z}_t \equiv \frac{x_t}{\hat{\sigma}_t}, \quad \text{with } \hat{\sigma}_t^2 = \hat{\omega}_T + \hat{\alpha}_T x_t^2, \quad t = 1, \ldots, T,$$

and consider the ordered residuals $\hat{z}_{(1)} \leq \ldots \leq \hat{z}_{(T)}$ in order to obtain

$$\hat{F}_z^{-1}(\kappa) = \hat{z}_{(\max\{\lfloor T\kappa \rfloor, 1\})}, \tag{IV.2}$$

where $\lfloor x \rfloor$ is the integer part of $x \in \mathbb{R}$. The VaR *estimate is given by*

$$\widehat{\mathrm{VaR}_t^\kappa} = -\hat{\sigma}_{t+1} \hat{F}_z^{-1}(\kappa).$$

Example IV.3.3 *(*ES*, ARCH(1)) Suppose that $z_t \sim i.i.d.N(0,1)$. Note that $x_{t+1} = \sigma_{t+1} z_{t+1}$, and similar to Example IV.2.2 we have that*

$$\begin{aligned}
\mathrm{ES}_{t,1}^\kappa &= E[-x_{t+1} | x_{t+1} < -\mathrm{VaR}_t^\kappa, \mathcal{I}_t] \\
&= E[-\sigma_{t+1} z_{t+1} | \sigma_{t+1} z_{t+1} < \sigma_{t+1} \Phi^{-1}(\kappa), \mathcal{I}_t] \\
&= \sigma_{t+1} E[(-z_{t+1}) | z_{t+1} < \Phi^{-1}(\kappa)] \\
&= \kappa^{-1} \sigma_{t+1} \phi(-\Phi^{-1}(\kappa)),
\end{aligned}$$

If we, as in Example IV.3.2, instead assume that z_t has an unknown distribution,

$$\begin{aligned}
\mathrm{ES}_t^\kappa &= \sigma_{t+1} E[(-z_{t+1}) | z_{t+1} < F_z^{-1}(\kappa)] \\
&= \kappa^{-1} \sigma_{t+1} E[(-z_{t+1}) \mathbb{I}\{z_{t+1} < F_z^{-1}(\kappa)\}]
\end{aligned}$$

Note that the quantity $E[(-z_{t+1}) \mathbb{I}\{z_{t+1} < F_z^{-1}(\kappa)\}]$ may be estimated based on the standardized residuals and the estimator for $F_z^{-1}(\kappa)$ defined in (IV.2):

$$\hat{E}[-z_{t+1} \mathbb{I}\{z_{t+1} < F_z^{-1}(\kappa)\}] = \frac{1}{T} \sum_{t=1}^{T} (-\hat{z}_{t+1}) \mathbb{I}\{\hat{z}_{t+1} < \hat{F}_z^{-1}(\kappa)\}.$$

IV.4 Simulation-based estimation

Until now, we have focused on the computation of one-period ahead VaR and ES for GARCH processes. In this section, we consider the multi-period ahead risk measures. Such measures are much more burdensome to compute, because the multi-period ahead conditional distributions of GARCH processes are unknown, as illustrated below.

Consider the ARCH(1) process with Gaussian errors, i.e. $x_{t+1} = \sqrt{\omega + \alpha x_t^2} z_{t+1}$ with $z_t \sim i.i.d. N(0,1)$. Since the factor $\sqrt{\omega + \alpha x_t^2}$ is known given the information set $\mathcal{I}_t$, it holds that $x_{t+1}|\mathcal{I}_t \sim N(0, \omega + \alpha x_t^2)$ which we exploited in the previous section to obtain the VaR and ES. Now, suppose that we want to compute the two-period ahead VaR. This relies on computing the κ percentile of the conditional loss distribution, i.e. the conditional distribution of $-x_{t+1,2}$ given $\mathcal{I}_t$ with $x_{t+1,2} = (x_{t+1} + x_{t+2})$. By recursions,

$$x_{t+2} = \sqrt{\omega + \alpha x_{t+1}^2} z_{t+2} = \sqrt{\omega + \alpha(\omega + \alpha x_t^2) z_{t+1}^2} z_{t+2}.$$

Clearly, $x_{t+2}|\mathcal{I}_{t+1} \sim N(0, \omega + \alpha x_{t+1}^2)$, but it is not clear what the conditional distribution of x_{t+2} (given $\mathcal{I}_t$) is, since the factor $\sqrt{\omega + \alpha(\omega + \alpha x_t^2) z_{t+1}^2}$ is unknown given the information set $\mathcal{I}_t$. In particular, one can show that the conditional distribution is non-Gaussian.

Consequently, one may view the conditional distribution as intractable, and instead one may approximate the risk-measures by means of simulations, as outlined in the following algorithm.

Algorithm IV.4.1 *(Simulation-based* VaR *for ARCH(1) with Gaussian errors) Let* $(\omega, \alpha)'$ *and* x_t *be known and fixed.*

1. *For* $i = 1, \ldots, M$ *(with* $(1-\kappa)M \geq 1$*) draw* $z_{t+1}^{(i)}$ *and* $z_{t+2}^{(i)}$ *independently from* $N(0,1)$*, and compute*

$$x_{t+1,2}^{(i)} = (x_{t+1}^{(i)} + x_{t+2}^{(i)}),$$

with

$$x_{t+1}^{(i)} = \sqrt{\omega + \alpha x_t^2} z_{t+1}^{(i)},$$
$$x_{t+2}^{(i)} = \sqrt{\omega + \alpha (x_{t+1}^{(i)})^2} z_{t+2}^{(i)}.$$

2. *Consider the ordered returns* $x_{t+1,2}^{[M]} \leq \ldots \leq x_{t+1,2}^{[1]}$*. Using the definition of* VaR *in (IV.1), obtain the approximate* VaR *as the* $(1-\kappa)$ *percentile of the simulated losses, i.e.*

$$\mathrm{VaR}_{t,2}^{\kappa,\mathrm{sim}} = -(x_{t+1,2}^{[\lfloor (1-\kappa)M \rfloor]}),$$

Typically, M is chosen to be quite large, say 10^5 or 10^6 in order to increase the presicion of the VaR estimate. Note that the above algorithm may be extended or modified in several directions:

- One may extend to different GARCH models and to arbitrary horizon h.
- As is typical in practice, the values of $(\omega, \alpha)'$ are unknown, and one may in step 1 replace $(\omega, \alpha)'$ by estimates $(\hat{\omega}_T, \hat{\alpha}_T)'$.
- If the distribution of z_t is unknown, then one may in step 1 draw $z_{t+1}^{(i)}$ and $z_{t+2}^{(i)}$ with equal probability (and with replacement) from the aforementioned standardized residuals, $(\hat{z}_t : t = 1, \ldots, T)$.
- In a third step, one may compute (similar to Example IV.3.3 above) the ES as

$$\mathrm{ES}_{t,2}^{\kappa,\mathrm{sim}} = \kappa^{-1} \frac{1}{M} \sum_{i=1}^{M} (-x_{t+1,2}^{(i)}) \mathbb{I}\{x_{t+1,2}^{(i)} < -\mathrm{VaR}_{t,2}^{\kappa,\mathrm{sim}}\}.$$

Next, we consider how to address the estimation uncertainty of the above algorithm, in the case that we use estimated parameter values $(\hat{\omega}_T, \hat{\alpha}_T)'$ in step 1. Since we do not have an explicit expression for the ES in terms of the model parameters, it is not possible to apply Taylor expansions arguments as we did for the VaR estimator. Instead we address the uncertainty by means of an extra layer of simulations, where we exploit that the estimator $\hat{\theta}_T = (\hat{\omega}_T, \hat{\alpha}_T)'$ is approximately Gaussian. Recall from Part III, that (under suitable conditions), as $T \to \infty$, $\sqrt{T}(\hat{\theta}_T - \theta) \xrightarrow{d} N(0, \Omega_I^{-1} \Omega_S \Omega_I^{-1})$, such that

$$\hat{\theta}_T \overset{a}{\sim} N(\theta, T^{-1} \Omega_I^{-1} \Omega_S \Omega_I^{-1}), \tag{IV.3}$$

and there exist consistent estimators, $\hat{\Omega}_{S,T}$ and $\hat{\Omega}_{I,T}$, for Ω_S and Ω_I. The following relies on, in a first step, to draw parameter values from the distribution $N(\hat{\theta}_T, T^{-1} \hat{\Omega}_{I,T}^{-1} \hat{\Omega}_{S,T} \hat{\Omega}_{I,T}^{-1})$ in order to take into account the random variation in $\hat{\theta}_T$ and hence in the estimator for the VaR (see e.g. Blasques et al. 2016, for a similar approach in relation to the computation of forecasting error bands).

Algorithm IV.4.2 *(Simulation-based* VaR *for ARCH(1) with Gaussian errors under estimation uncertainty) Let x_t be known and fixed, and suppose that estimator $\hat{\theta}_T = (\hat{\omega}_T, \hat{\alpha}_T)'$ satisfies (IV.3) and that we have some estimators, $\hat{\Omega}_{S,T}$ and $\hat{\Omega}_{I,T}$, for Ω_S and Ω_I.*

1. *For $b = 1, \ldots, B$ (say, $B = 999$), draw $\theta^{(b)} = (\omega^{(b)}, \alpha^{(b)})'$ from*

$$N(\hat{\theta}_T, T^{-1}\hat{\Omega}_{I,T}^{-1}\hat{\Omega}_{S,T}\hat{\Omega}_{I,T}^{-1}). \tag{IV.4}$$

2. *For a given b, for $i = 1, \ldots, M$ draw $z_{t+1}^{(i,b)}$ and $z_{t+2}^{(i,b)}$ independently from $N(0,1)$, and compute*

$$x_{t+1,2}^{(i,b)} = (x_{t+1}^{(i,b)} + x_{t+2}^{(i,b)}),$$

with

$$x_{t+1}^{(i,b)} = \sqrt{\omega^{(b)} + \alpha^{(b)} x_t^2} z_{t+1}^{(i,b)},$$
$$x_{t+2}^{(i,b)} = \sqrt{\omega^{(b)} + \alpha^{(b)} (x_{t+1}^{(i,b)})^2} z_{t+2}^{(i,b)}.$$

3. *For a given b, consider the ordered returns $x_{t+1,2}^{[M,b]} \leq \ldots \leq x_{t+1,2}^{[1,b]}$. Using the definition of* VaR *in (IV.1), compute the approximate* VaR *as the $(1-\kappa)$ percentile of the simulated losses, i.e.*

$$\mathrm{VaR}_{t,2}^{\kappa,\mathrm{sim},b} = -(x_{t+1,2}^{[\lfloor (1-\kappa)M \rfloor, b]}).$$

4. *Consider the ordered B* VaR *estimates, $\mathrm{VaR}_{t,2}^{\kappa,\mathrm{sim},(1)} \leq \ldots \leq \mathrm{VaR}_{t,2}^{\kappa,\mathrm{sim},(B)}$. The 95% error interval of the* VaR *is given by*

$$[\mathrm{VaR}_{t,2}^{\kappa,\mathrm{sim},(B\times 0.025)}, \mathrm{VaR}_{t,2}^{\kappa,\mathrm{sim},(B\times 0.975)}].$$

Remark IV.4.1 *In step 1, one may end up with a draw of $\omega^{(b)}$ and/or $\alpha^{(b)}$ that is negative, and hence outside of the parameter space. In this case one may make a new draw of $\theta^{(b)}$. Alternatively, one may reparametrize the model, e.g. by estimating $\tilde{\omega} = \log(\omega)$ and $\tilde{\alpha} = \log(\alpha)$ that are freely varying.*

Remark IV.4.2 *As an alternative to the algorithm outlined above, one may address the estimation uncertainty by means of bootstrap methods. In particular, instead of drawing each $\theta^{(b)}$ from the "asymptotic" distribution in (IV.4), one may estimate each $\theta^{(b)}$ from a bootstrap sample generated from the assumed data generating process with true parameter values given by $\hat{\theta}_T = (\hat{\omega}_T, \hat{\alpha}_T)'$. One advantage of such an approach is that the draws $(\theta^{(b)} : b = 1, \ldots, B)$ may be closer to mimicking draws from the unknown finite-sample distribution of $\hat{\theta}_T$. See e.g. Cavaliere, Pedersen and Rahbek (2018) for more details on bootstrap methods in relation to (G)ARCH processes.*

References

Basel Committee on Banking Supervision, 2013, "Fundamental review of the trading book: A revised market risk framework", BIS, Basel, Switzerland, http://www.bis.org/publ/bcbs265.pdf

Blasques, F., Koopmann, S.J. , Lasak, K., & Lucas, A., 2016, "In-sample confidence bands and out-of-sample forecast bands for time-varying parameters in observation-driven models", *International Journal of Forecasting*, vol. 32, pp. 875–887.

Cavaliere, G., Pedersen, R.S., & Rahbek, A., 2018, "The fixed volatility bootrap for a class of ARCH processes", *Journal of Time Series Analysis*, vol. 39, pp. 920-941.

Francq, C. & Zakoïan, J.-M., 2019, *GARCH Models: Structure, Statistical Inference and Financial Applications*, 2nd edition. Wiley.

Ibragimov, R., 2009, "Portfolio diversification and value at risk under thick-tailedness", *Quantitative Finance*, vol. 9, pp. 565–580.

Anders Rahbek
Rasmus Søndergaard Pedersen
University of Copenhagen

October 2021

Part V
Option Pricing Based on GARCH Models

V.1 Introduction

In this note, we consider pricing of financial derivatives based on GARCH models. We focus on the pricing of European call options, but emphasize that some of the outlined methods can be applied to other derivatives, in particular ones with payoff at a known future date T. Similar to the Expected Shortfall, discussed in the previous lecture note, the computation of the derivatives prices – in essence – relies on computing the mean of a (conditional) payoff distribution. For some special cases we obtain closed-form pricing formulas, but in general we rely on simulation-based estimation.

A European call option gives the right - but not the obligation - to buy a stock for a specified strike price K on a specified date T. Hence, with S_T the price of the stock at time T, the payoff of the option is given by

$$\pi_T^{\text{Call}}(K) = \begin{cases} S_T - K & \text{if } S_T \geq K \\ 0 & \text{otherwise} \end{cases}, \tag{V.1}$$

i.e. $\pi_T^{\text{Call}}(K) = \max(S_T - K, 0)$. Due to its nonnegative payoff, the option must have a nonnegative price at time $t \leq T$. Techniques for determining the prices of options is a vast research field dating back to seminal work by Bachelier (1900) as well as the work of Black, Scholes, and Merton (1973). As mentioned, we will discuss how such options may be priced under the more realistic assumption that the returns on the underlying stock follows a GARCH process.

V.1.1 On derivatives pricing

We recall some important concepts from finance theory. Let $r \geq 0$ denote the (continuously compounded) risk-free interest rate. The price, B_t, at time

$t \in \mathbb{N}$ of one monetary unit invested at time $t = 0$ at the risk-free rate is by definition given by

$$B_t = B_{t-1}e^r, \quad t \in \mathbb{N}, \quad B_0 = 1. \tag{V.2}$$

Let $\mathcal{I}_t$ denote some information set available at time t, and let P_t denote the price at time t of some asset with payoff π_{t+1} at time $t+1$. The absence of arbitrage in a market is equivalent to the existence of a so-called stochastic discount factor (SDF) :

Definition V.1.1 (One-period SDF) *The stochastic discount factor (SDF), $m_{t,t+1} \geq 0$, satisfies*

$$P_t = E[m_{t,t+1}\pi_{t+1}|\mathcal{I}_t]. \tag{V.3}$$

For additional details, we refer to Back (2017).[1] Ideally, equation (V.3) holds for all assets, and hence an SDF can be used to price all assets in a market. We give a few important examples: Let S_t denote the price at time t of the stock that does not pay out dividends at time $t+1$. By definition, the payoff at time $t+1$ of this stock is simply the stock price at time $t+1$ ($P^S_{t+1} = S_{t+1}$), so that

$$S_t = E\left[m_{t,t+1}S_{t+1}|\mathcal{I}_t\right]. \tag{V.4}$$

Moreover, in light of the price of the risk-free bank account in (V.2), we have that

$$B_t = E[m_{t,t+1}B_{t+1}|\mathcal{I}_t] \quad \Leftrightarrow \quad 1 = E[m_{t,t+1}\underbrace{(B_{t+1}/B_t)}_{e^r}|\mathcal{I}_t],$$

so that

$$E[m_{t,t+1}|\mathcal{I}_t] = e^{-r}. \tag{V.5}$$

Lastly, we have that the European call option expiring at time $t+1$ has price

$$P^{\text{Call}}_t(t+1, K) = E[m_{t,t+1}\pi^{\text{Call}}_{t+1}(K)|\mathcal{I}_t] = E[m_{t,t+1}\max(S_{t+1} - K, 0)|\mathcal{I}_t].$$

Note that, conveniently, we may also define the $(T-t)$ period SDF:

Definition V.1.2 (Multi-period SDF) *The $(T-t)$ period SDF $m_{t,T}$ is given by*

$$m_{t,T} = \prod_{i=t}^{T-1} m_{i,i+1}, \tag{V.6}$$

where $m_{i,i+1}$ is the one-period SDF defined in (V.3) for $i = t, \ldots, T-1$.

[1] Note that if investors were risk neutral one would simply discount payoffs according to the risk-free rate, i.e. $S_t = e^{-r}E[S_{t+1}|\mathcal{I}_t]$, as done in Section V.3. This kind of discounting only takes into account the time value of money – but it does not account for risk. The SDF is intended to capture both the time value of money and the discounting due to risk.

The SDF is sometimes refered to as the *pricing kernel*.

A straightforward application of the law of iterated expectations and (V.4)-(V.5) imply that

$$E[m_{t,T}|\mathcal{I}_t] = e^{-r(T-t)} \quad \text{and} \quad E\left[m_{t,T}S_T|\mathcal{I}_t\right] = S_t. \tag{V.7}$$

Likewise, the price (at time t) of the European call option with payoff (V.1) is

$$P_t^{\text{Call}}(T,K) = E[m_{t,T}\pi_T^{\text{Call}}(K)|\mathcal{I}_t] = E[m_{t,T}\max(S_T - K, 0)|\mathcal{I}_t].$$

and we consider the computation of this quantity under various assumptions about the joint dynamics of the SDF $(m_{t,t+1})$ and the stock price (S_t).

V.2 The normal and log-normal distributions

In this section we provide a few results for the normal and log-normal distributions that will show up to be useful for the derivations of the option prices.

By definition, a positive real-valued variable $X > 0$ is log-normal, if $Y = \log(X)$ is Gaussian. Thus with Y $\mathrm{N}(\mu, \sigma^2)$-distributed, then $X = \exp(Y)$ has density[2]

$$f(x) = \frac{1}{x}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}\left(\log x - \mu\right)^2\right), \qquad x > 0. \tag{V.8}$$

We say that X is log-normal with parameters (μ, σ^2), with density $f(x)$ in (V.8). The kth moment of X is

$$EX^k = \exp\left(\sigma^2\frac{k^2}{2} + k\mu\right). \tag{V.9}$$

In particular,

$$EX = \exp\left(\frac{\sigma^2}{2} + \mu\right) \text{ and } VX = \left(\exp\left(\sigma^2\right) - 1\right)\exp\left(\sigma^2 + 2\mu\right),$$

and moreover, as will be useful for computation of kurtosis and skewness of returns later,

$$EX^2 = \exp\left(2\sigma^2 + 2\mu\right), \quad EX^4 = \exp\left(8\sigma^2 + 4\mu\right) \text{ and } EX^3 = \exp\left(\sigma^2\frac{9}{2} + 3\mu\right).$$

The following lemma states an important property of a log-normal random variable.

[2] To see this, note that $h(y) = \exp(y) = x$, we have $h^{-1}(x) = \log x$ with derivative $\frac{1}{x}$. Hence if Y has density $g(y)$, then x by the formula for transformation of distributions, has density $f(x) = \frac{1}{x}g(\log(x))$. That is, with $h(y) = x$, $f(x) = g\left(h^{-1}(x)\right)\left|\partial h^{-1}(x)\right|$.

Lemma V.2.1 *Let X be log-normal with parameters (μ, σ^2), and let K be a non-negative constant. Then*

$$E[\max(X-K,0)] = E[X]\Phi\left(-u+\sigma\right) - K\Phi\left(-u\right), \quad u := \frac{\log(K)-\mu}{\sigma}$$

where $\Phi(\cdot)$ is the cdf of a standard normal distribution.

Lastly, the next lemma shows up to be useful for deriving the famous Black-Scholes-Merton pricing formula.

Lemma V.2.2 *Let X,$Y \in \mathbb{R}$ be jointly normal with covariance given by σ_{XY}. Then for any continuous function $g : \mathbb{R} \to \mathbb{R}$,*

$$E\left[\exp(X)g(Y)\right] = E[\exp(X)]E[g(Y+\sigma_{XY})].$$

V.3 Non-stochastic discounting

In this section, we consider a "simple" pricing approach, where the stock price is assumed to be log-normal and the SDF is constant. Note that if the SDF, $m_{t,t+1}$, is constant and satisfies (V.4)-(V.5), then we have that

$$m_{t,t+1} = e^{-r} \tag{V.10}$$

and

$$E(S_{t+1}/S_t|\mathcal{I}_t) = e^r. \tag{V.11}$$

The condition in (V.11) holds under the assumption that S_{t+1}/S_t is conditionally log-normal with parameters (μ, σ^2) with

$$\mu + \sigma^2/2 = r, \tag{V.12}$$

which can be seen by recalling that $E(S_{t+1}/S_t|\mathcal{I}_t) = \exp(\mu+\sigma^2/2)$ in light of (V.9). Recall that the price of the call option expiring at time $t+1$ is

$$\begin{aligned} P_t^{\text{Call}}(t+1,K) &= E[m_{t,t+1}\max(S_{t+1}-K,0)|\mathcal{I}_t] \\ &= S_t E[m_{t,t+1}\max(S_{t+1}/S_t - K/S_t,0)|\mathcal{I}_t] \\ &= S_t e^{-r} E[\max(S_{t+1}/S_t - K/S_t,0)|\mathcal{I}_t], \end{aligned}$$

Then under the aforementioned conditional log-normality of S_{t+1}/S_t, we may apply Lemma V.2.1 to evaluate $E[\max(S_{t+1}/S_t - K/S_t,0)|\mathcal{I}_t]$. Consequently, we obtain the following theorem.

Theorem V.3.1 (One-period Black-Scholes) *Suppose that the SDF, $m_{t,t+1}$, is constant and given by (V.10), and assume that the $\mathcal{I}_t$-conditional distribution of the log-return $\log(S_{t+1}/S_t)$ is normal with constant mean μ and variance σ^2 satisfying (V.12). Then the price of the call option expiring at time $t+1$ is*

$$\begin{aligned} P_t^{Call}(t+1,K) &= E[m_{t,t+1}\max(S_{t+1}-K,0)|\mathcal{I}_t] \\ &= S_t\Phi\left(-u_{t,t+1}+\sigma\right) - e^{-r}K\Phi\left(-u_{t,t+1}\right), \end{aligned}$$

where

$$u_{t,s} := \frac{\log(Ke^{-(r-\sigma^2/2)(s-t)}) - \log(S_t)}{\sigma\sqrt{s-t}}, \quad t<s. \tag{V.13}$$

The above theorem may be extended to an arbitrary time-horizon. In particular, using the definition of the SDF, it is straightforward to show that for $T \geq t+1$,

$$m_{t,T} = e^{-r(T-t)} \quad \text{and} \quad E[S_T/S_t|\mathcal{I}_t] = e^{r(T-t)}.$$

Again, the latter condition holds under the assumption that S_T/S_t is conditionally log-normal with

$$E[\log(S_T/S_t)|\mathcal{I}_t] + V[\log(S_T/S_t)|\mathcal{I}_t]/2 = r(T-t). \tag{V.14}$$

For instance, this condition holds if daily log-returns, $\{\log(S_{i+1}/S_i)\}_{i=t}^{T-1}$, are (conditional on $\mathcal{I}_t$) i.i.d. normal with mean μ and variance σ^2, so that $E[\log(S_T/S_t)|\mathcal{I}_t] = (T-t)\mu$, $V[\log(S_T/S_t)|\mathcal{I}_t] = \sigma^2(T-t)$, and (V.14) holds under the condition that $\mu + \sigma^2/2 = r$. We have the following theorem.

Theorem V.3.2 (Multi-period Black-Scholes) *Suppose that the SDF, $m_{t,t+1}$, is constant and satisfies (V.10), and let $m_{t,T}$ be given by (V.6). Assume that the $\mathcal{I}_t$-conditional distribution of the $(T-t)$-period log-return, $\log(S_T/S_t)$, is normal with mean $\mu(T-t)$ and variance $\sigma^2(T-t)$, where μ and σ^2 are constants satisfying (V.12). Then the price of the call option expiring at time T is*

$$\begin{aligned} P_t^{Call}(T,K) &= E[m_{t,T}\max(S_T-K,0)|\mathcal{I}_t] \\ &= S_t\Phi\left(-u_{t,T}+\sigma\sqrt{T-t}\right) - \exp(-r(T-t))K\Phi\left(-u_{t,T}\right), \end{aligned}$$

where $u_{t,T}$ is defined in (V.13).

We note that the above pricing formulas made use of a constant SDF, which (due to no arbitrage) in turn implied that stock price was expected to grow at the risk-free rate, $E[S_T/S_t|\mathcal{I}_t] = e^{r(T-t)}$, which appears to be restrictive. In particular, we may want to price options written on stocks with other expected growth rates. In order to do so, we consider pricing under a truly random SDF, which is the topic of the next section.

V.4 Stochastic discounting

In this section we consider pricing when the SDF is random. Since the SDF is a positive random variable, a candidate (conditionally) distribution is the log-normal. Hence, in line with the previous section, where we considered log-normal stock prices, we follow the approach of, e.g., Garcia et al. (2010, Section 2.4) and assume that the log-returns and the logarithm of the SDF are conditionally jointly normal. Specifically, we assume that

$$\begin{pmatrix} \log(S_{t+1}/S_t) \\ \log m_{t,t+1} \end{pmatrix} |\mathcal{I}_t \sim N_2 \left(\begin{pmatrix} \mu \\ \mu_m \end{pmatrix}, \begin{pmatrix} \sigma^2 & \sigma_{m,r} \\ \sigma_{m,r} & \sigma_m^2 \end{pmatrix} \right). \tag{V.15}$$

At first glance, this assumption looks quite general in the sense that we have many parameters to specify the joint conditional distribution. Moreover, we may note that as we have conditioned on $\mathcal{I}_t$, the parameters may not be constant, but should simply be adapted with respect to $\mathcal{I}_t$. For now we suppress the potential time-dependence of the parameters and view them as constants. We return to the case of time-varying parameters in Section V.5. Note that the parameters cannot be chosen freely, as the no-arbitrage conditions in (V.4)-(V.5) impose some structure on the model in (V.15). First, since $m_{t,t+1}$ is (conditionally) log-normal with parameters (μ_m, σ_m^2), we have, in light of (V.9), that the first part of (V.5) implies that $\mu_m + \sigma_m^2/2 = -r$. Next, from (V.4) we have that

$$E\left[\exp\left\{\log m_{t,t+1}\right\} \times \exp\left\{\log(S_{t+1}/S_t)\right\} |\mathcal{I}_t\right] = 1, \tag{V.16}$$

and we apply the fact that $(\log(S_{t+1}/S_t), \log m_{t,t+1})$ is jointly normal and Lemma V.2.2 in order show that the no arbitrage conditions impose additional structure on the model parameters. We have the following result, which is proved in the appendix.

Lemma V.4.1 *Suppose that (V.15) holds. The no arbitrage conditions in (V.4)-(V.5) imply that*

$$\mu_m + \frac{\sigma_m^2}{2} = -r \tag{V.17}$$

and

$$\sigma_{m,r} + \mu = -\frac{\sigma^2}{2} + r. \tag{V.18}$$

We now turn to pricing of a call option under the joint normality assumption in (V.15). Recall that the price of a call option expiring at time $t+1$ is

$$\begin{aligned} P_t^{\text{Call}}(t+1, K) &= E[m_{t,t+1} \max(S_{t+1} - K, 0) | \mathcal{I}_t] \\ &= S_t E[\exp\left\{\log m_{t,t+1}\right\} \max(\exp\{\log(S_{t+1}/S_t)\} - K/S_t, 0) | \mathcal{I}_t]. \end{aligned}$$

Using (V.15) and noting that $\max(x-y,0)$ is continuous in x for a fixed y, it follows by Lemma V.2.2, that

$$\begin{aligned} P_t^{\text{Call}}(t+1,K) &= S_t E[\exp(\log m_{t,t+1})\,|\mathcal{I}_t] \\ &\quad \times E[\max(\exp\{\log(S_{t+1}/S_t)+\sigma_{m,r}\}-K/S_t,0)|\mathcal{I}_t] \\ &= S_t E[m_{t,t+1}|\mathcal{I}_t]E[\max(e^{\sigma_{m,r}}(S_{t+1}/S_t)-K/S_t,0)|\mathcal{I}_t] \\ &= S_t e^{-r} E[\max(e^{\sigma_{m,r}}(S_{t+1}/S_t)-K/S_t,0)|\mathcal{I}_t], \end{aligned} \tag{V.19}$$

where the last equality follows from (V.5). An explicit formula for $P_t^{\text{Call}}(t+1,K)$ is now obtained by (i) realizing that $e^{\sigma_{m,r}}(S_{t+1}/S_t)$ is conditionally log-normal with parameters $(\mu+\sigma_{m,r},\sigma^2)$, (ii) applying Lemma V.2.1 to $E[\max(e^{\sigma_{m,r}}(S_{t+1}/S_t)-K/S_t,0)|\mathcal{I}_t]$, and (iii) exploiting the parameter constraints stated in Lemma V.4.1. We have the following result.

Theorem V.4.1 *Suppose that the joint conditional distribution of the log-return and the log-SDF is given by (V.15), and suppose that the no-arbitrage conditions in (V.4)-(V.5) hold. Then the price of the call option expiring at time $t+1$ is*

$$\begin{aligned} P_t^{Call}(t+1,K) &= E[m_{t,t+1}\max(S_{t+1}-K,0)|\mathcal{I}_t] \\ &= S_t\Phi(-u_{t,t+1}+\sigma)-e^{-r}K\Phi(-u_{t.t+1}), \end{aligned}$$

where $u_{t,t+1}$ is defined in (V.13). That is, the price is equivalent to the one stated in Theorem V.3.1.

We refer to the appendix for a proof. Note that the pricing formula is identical to the one in Theorem V.3.1 where the SDF was assumed constant. The reason is that, under (conditional) log-normality of the SDF, the formula in Theorem V.4.1 only depends on the volatility parameter σ appearing in the dynamics equation (V.15). In particular, we note that a constant SDF is applicable with the dynamics in (V.15) by setting $\sigma_{m,r}=\sigma_m^2=0$ and $\mu_m=e^{-r}$, such that the log-SDF has a degenerate normal distribution with unit mass at e^{-r}. Hence, under conditionally Gaussian returns, it does not matter if the SDF is constant or log-normal when determining the price of an option.

Similar to Theorem V.3.2, we may obtain a pricing formula for the option expiring at an arbitrary time point $T\geq t+1$. We make the following assumption about the $(T-t)$-period conditional distribution of the log-returns and the log-SDF:

$$\begin{pmatrix}\log\frac{S_T}{S_t}\\ \log m_{t,T}\end{pmatrix}|\mathcal{I}_t \sim N_2\left((T-t)\begin{pmatrix}\mu_r\\ \mu_m\end{pmatrix},(T-t)\begin{pmatrix}\sigma^2 & \sigma_{m,r}\\ \sigma_{m,r} & \sigma_m^2\end{pmatrix}\right). \tag{V.20}$$

The dynamics in (V.20) hold if the joint process $\{(\log(S_{i+1}/S_i), \log m_{i,i+1})\}_{i=t}^{T-1}$ is (conditional on $\mathcal{I}_t$) i.i.d. bivariate normal with mean and covariance matrix given in (V.15). We emphasize that more advanced dynamics could be allowed, which we will consider in more detail in the next section. The assumption in (V.20) together with the no arbitrage condition in (V.7) yield the famous Black-Scholes-Merton formula stated in the following theorem. Its proof is similar to the one of Theorem V.4.1 and is left as an exercise.

Theorem V.4.2 *(Black-Scholes-Merton formula) Suppose that the $(T-t)$ log-return and log-SDF satisfy (V.20). Moreover, assume that the no arbitrage condition in (V.7) holds. Then the price of the call option expiring at time T is*

$$\begin{aligned} P_t^{Call}(T,K) &= E[m_{t,T}\max(S_T-K,0)|\mathcal{I}_t] \\ &= S_t\Phi\left(-u_{t,T}+\sigma\sqrt{T-t}\right) - \exp(-r(T-t))K\Phi\left(-u_{t,T}\right), \end{aligned}$$

where $u_{t,T}$ is defined in (V.13).

Remark V.4.1 *An equivalent way of stating the price, often found in textbooks, is that*

$$P_t^{Call}(T,K) = S_t\Phi(d_1) - \exp(-r(T-t))K\Phi(d_2),$$

where

$$d_1 = \frac{\log(S_t/K) + (r+\sigma^2/2)(T-t)}{\sigma\sqrt{T-t}} \quad \textit{and} \quad d_2 = d_1 - \sigma\sqrt{T-t}.$$

V.5 Pricing based on GARCH

In this section we consider option pricing when allowing for time-varying "parameters", such as time-varying conditional volatility. A natural extension of (V.15) is to allow the conditional mean and covariance matrix to be time dependent, i.e. we may assume that

$$\begin{pmatrix} \log(S_{t+1}/S_t) \\ \log m_{t,t+1} \end{pmatrix} |\mathcal{I}_t \sim N_2\left(\begin{pmatrix} \mu_{r,t+1} \\ \mu_{m,t+1} \end{pmatrix}, \begin{pmatrix} \sigma_{t+1}^2 & \sigma_{m,r,t+1} \\ \sigma_{m,r,t+1} & \sigma_{m,t+1}^2 \end{pmatrix}\right), \tag{V.21}$$

where $\sigma_{m,r,t+1}$ denotes the (potentially time-dependent) conditional covariance between $\log m_{t,t+1}$ and $\log(S_{t+1}/S_t)$. Note that the conditional distribution in (V.21) is consistent with the assumption that

$$\begin{pmatrix} \log(S_{t+1}/S_t) \\ \log m_{t,t+1} \end{pmatrix} = \begin{pmatrix} \mu_{r,t+1} \\ \mu_{m,t+1} \end{pmatrix} + \varepsilon_{t+1}, \quad \varepsilon_{t+1} = \Omega_{t+1}^{1/2} z_{t+1}, \tag{V.22}$$

where the process $\{z_{t+1}\}$ is *i.i.d.* with a bivariate standard normal distribution, i.e.

$$z_{t+1} \sim N_2\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\right),$$

and $\Omega_{t+1}^{1/2}$ is the (positive definite) square-root of

$$\Omega_{t+1} = \begin{pmatrix}\sigma_{t+1}^2 & \sigma_{m,r,t+1}\\ \sigma_{m,r,t+1} & \sigma_{m,t+1}^2\end{pmatrix}.$$

By definition, Ω_{t+1} is a function of $\mathcal{I}_t$. In particular, this allows for the log-returns to be driven by a (G)ARCH process, e.g. $\sigma_{t+1}^2 = \omega + \alpha(x_t - \mu_{r,t+1})^2$ with $x_t := \log(S_{t+1}/S_t)$. Note that under the no arbitrage conditions in (V.4)-(V.5), the entries of the conditional mean and covariance matrix in (V.21) are subject to constraints similar to the ones given in Lemma V.4.1 above. By arguments similar to the ones that we used to prove Theorem V.4.1, we have the following result.

Theorem V.5.1 *Suppose that the log-returns and the log-SDF are conditionally jointly normal according to (V.21). Under the no arbitrage condition (V.4)-(V.5), the price of a call option expiring at time $t+1$ is given by*

$$P_t^{Call}(t+1, K) = S_t\Phi\left(-u_t + \sigma_{t+1}\right) - e^{-r}K\Phi\left(-u_t\right),$$

where

$$u_t := \frac{\log(Ke^{-(r-\sigma_{t+1}^2/2)}) - \log(S_t)}{\sigma_{t+1}}. \tag{V.23}$$

As mentioned, the above pricing formula for an option that expires at time $t+1$ is compatible with log-returns following a GARCH process, such as $\log(S_{t+1}/S_t)|\mathcal{I}_t \sim N(0, \sigma_{t+1}^2)$. In particular, if $\log(S_{t+1}/S_t)$ follows a GARCH process, the option price can be obtained by estimating the GARCH parameters and plugging the estimated conditional variance, $\hat{\sigma}_{t+1}^2$, into the pricing formula. Similar to the risk measures, discussed in the previous chapter, one may choose to report error bands of the estimated option price, taking into account the estimation uncertainty of the model parameters.

V.6 Simulation-based methods

As we already noted in the previous chapter, if $\log(S_{t+1}/S_t)|\mathcal{I}_t \sim N(0, \sigma_{t+1}^2)$ it will (in general) not be the case that $\log(S_{t+h}/S_t)$ is conditionally Gaussian for $h > 1$ (see also Drost and Nijman, 1993, for additional details). Hence,

it is not obvious how to obtain closed-form GARCH-based prices for options expiring more than one day ahead. Instead, one may estimate the prices by means of simulation. We here consider a special case of a recent approach by Zhu and Ling (2015). Specifically, with $x_t := \log(S_t/S_{t-1})$ the log-return at time t, Zhu and Ling (2015) consider an SDF given by

$$m_{t,t+1} = \frac{\exp\left(\theta_{t+1}x_{t+1}\right)}{E[\exp\{(1+\theta_{t+1})x_{t+1}\}|\mathcal{I}_t]},$$

for some $\theta_{t+1} \in \mathcal{I}_t$ that depends on the conditional distribution of x_t and that is subject to constraints imposed by the no-arbitrage conditions in (V.4)-(V.5).[3] For instance, suppose that x_t follows a GARCH(1,1)-type process,

$$x_t = \mu_t + \sigma_t z_t, \quad z_t \sim i.i.d.N(0,1), \tag{V.24}$$

$$\sigma_t^2 = \omega + \alpha(x_{t-1} - \mu_{t-1})^2 + \beta\sigma_{t-1}^2, \tag{V.25}$$

where $\mu_t \in \mathcal{I}_{t-1}$ is the (known) conditional mean of x_t. Using that z_t is Gaussian and hence that $\exp(x_{t+1})$ is conditionally log-normal with parameters $(\mu_{t+1}, \sigma_{t+1}^2)$, we have that in light of (V.9),

$$E[\exp\{(1+\theta_{t+1})x_{t+1}\}|\mathcal{I}_t] = \exp\left\{(1+\theta_{t+1})\mu_{t+1} + \frac{(1+\theta_{t+1})^2\sigma_{t+1}^2}{2}\right\}.$$

This, combined with the no-arbitrage condition in (V.5), implies that

$$m_{t,t+1} = \frac{\exp\left(\theta_{t+1}x_{t+1}\right)}{\exp\left\{(1+\theta_{t+1})\mu_{t+1} + \frac{(1+\theta_{t+1})^2\sigma_{t+1}^2}{2}\right\}} \quad \text{with } \theta_{t+1} = \frac{r-\mu_{t+1}}{\sigma_{t+1}^2} - \frac{1}{2}. \tag{V.26}$$

We refer to Zhu and Ling (2015) for many more details and derivations. Recall that $m_{t,T} = \prod_{i=t+1}^{T} m_{i-1,i}$ and that

$$P_t^{\text{Call}}(T,K) = E\left[m_{t,T}\max(S_T - K, 0)|\mathcal{I}_t\right]. \tag{V.27}$$

One may estimate $P_t^{\text{Call}}(T,K)$ by the following algorithm.

Algorithm V.6.1 *Suppose that the returns on the stock follows the GARCH(1,1) process in (V.24)-(V.25), and that the SDF is given by (V.26). Moreover assume that $\mu_{t+1} = g(x_t, \mu_t, \sigma_t^2)$ for some known function $g : \mathbb{R}\times\mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}$.*[4]

[3] Note that this choice of SDF is mathematically tractable. As such, this SDF does not contain any deep financial-economic insight. In particular, note that the no-arbitrage condition in (V.4) is always satisfied for this choice of SDF, since the condition is equivalent to $E[m_{t,t+1}\exp\{x_{t+1}\}|\mathcal{I}_t] = 1$.

[4] Note that this choice of function $g(\cdot)$ is quite flexible, in the sense that is allows for GARCH-in-mean ($\mu_{t+1} = \gamma\sigma_{t+1}^2 = \gamma(\omega+\alpha(x_t-\mu_t)^2+\beta\sigma_t^2)$), auto-regressive ($\mu_{t+1} = \phi x_t$), and moving average ($\mu_{t+1} = \varphi(x_t - \mu_t)$) terms, or a combination hereof. Obviously, one could also allow for additional lags of the quantities (x_t, μ_t, σ_t^2).

1. *For $i = 1, \ldots, M$ (for some large M) draw $z_{t+1}^{(i)}, \ldots, z_T^{(i)}$ independently from $N(0,1)$, and compute for $j = t+1, \ldots, T$,*

$$\begin{aligned}\mu_j^{(i)} &= g(x_{j-1}^{(i)}, \mu_{j-1}^{(i)}, \sigma_{j-1}^{(i)2}),\\ \sigma_j^{(i)2} &= \omega + \alpha(x_{j-1}^{(i)} - \mu_{j-1}^{(i)}) + \beta\sigma_{j-1}^{(i)2},\\ x_j^{(i)} &= \mu_j^{(i)} + z_j^{(i)}\sqrt{\sigma_j^{2(i)}},\end{aligned}$$

where $\mu_t^{(i)} = \mu_t \in \mathcal{I}_t$, $\sigma_t^{(i)2} = \sigma_t^2 \in \mathcal{I}_t$ and $x_t^{(i)} = x_t \in \mathcal{I}_t$. Moreover, compute

$$\xi_T^{(i)} := m_{t,T}^{(i)} \max(S_T^{(i)} - K, 0),$$

where

$$m_{t,T}^{(i)} = \prod_{j=t+1}^{T} m_{j-1,j}^{(i)},$$

with, for $j = t+1, \ldots, T$,

$$m_{j-1,j}^{(i)} := \frac{\exp(\theta_j^{(i)} x_j^{(i)})}{\exp\left[(1+\theta_j^{(i)})\mu_j^{(i)} + (1+\theta_j^{(i)})^2\sigma_j^{(i)2}/2\right]}, \quad \theta_j^{(i)} := \frac{r - \mu_j^{(i)}}{\sigma_j^{(i)2}} - \frac{1}{2},$$

and

$$S_T^{(i)} = S_t \exp(\sum_{j=t+1}^{T} x_j^{(i)}).$$

2. *The price, $P_t^{Call}(T,K)$, in (V.27) may be approximated by*

$$P_t^{Call,sim}(T,K) = \frac{1}{M}\sum_{i=1}^{M} \xi_T^{(i)}.$$

V.7 Data on option prices

Table 1 contains prices of European call options written on the S&P 500 Index (SPX) for different strikes and maturities.

The data are retrieved from Bloomberg. The options are traded at the Chicago Board Options Exchange (CBOE), and we refer to the CBOE webpage for additional details about the contracts. The table contains mid prices obtained after the closing of the exchange on September 17th 2015. The first row of the table contains the expiration dates of the options, whereas the second row states the number of trading days between September 17th, 2015

Table 1: S&P 500 Call Option Prices, September 17, 2015

Strike	18-09-2015 1	16-10-2015 21	20-11-2015 45	19-12-2015 64	15-01-2016 82	18-03-2016 125
1600			384.2	385.4	389.05	392.25
1650		335.2	336.3	338.5	370.5	347.9
1700	285.7	286.2	289.05	292.5	297.8	304.65
1750	236.45	237.8	242.9	247.6	253.8	266.78
1800	186.45	190.5	198.2	209	211.5	222.55
1850	136.7	144.95	155.5	162.95	171.3	184.25
1900	86.85	102.1	115.5	124.3	139.25	165
1950	37	63.45	79.05	90.6	104.35	114.85
2000	2.45	31.3	47.65	57	67.47	90.8
2050	0.08	10	23.5	33	45.3	63.75
2100	0.05	1.6	8.6	14.85	24	43.5
2150		0.3	1.85	5.79	10.1	24.1
2200		0.1	0.6	2.15	3.4	12.5
2250			0.37	0.6	0.9	5.75
2300			0.1	0.15	0.5	2.1
2350				0.1		1.35
2400						0.7

and the expiration days. It should be mentioned that the liquidity (measured in terms of the trading volume) is low for some of the contracts, meaning that the quoted prices may not be that precise. In particular, this is the case for certain contracts with long maturity and for "deep out of the money contracts" (i.e. contracts with K much greater than S_t) with short maturity. The closing price (S_t) of the SPX Index was 1990.20 on September 17th 2015.

Example V.7.1 *Consider the European call option with strike* $K = 1950$ *and expiration date on October 16, 2015. Using a sample from January 4, 2010 - September 17, 2015 of S&P 500 log-returns, we get an estimate* $\hat{\sigma} = 0.010050$ *based on the sample variance of the returns. With the risk-free rate* $r = 0.003/251$ *and* $T - t = 21$ *trading days to expiration, we get an option price of 60.11, based on the standard Black-Scholes-Merton pricing formula in Theorem V.4.2.*
Estimating a GARCH(1,1) model as in (V.24)-(V.25) with constant $\mu_t \equiv \gamma \in \mathbb{R}\ \forall t$ *based on the same data series, we get* $(\hat{\omega}, \hat{\alpha}, \hat{\beta}, \hat{\gamma}) = (0.041367 \times 10^{-4}, 0.14645, 0.81185, 0.072782 \times 10^{-2})$. *Using Algorithm V.6.1 (with* $B = 10^6$*) we obtain an option price of 62.47, which is closer to the actual market price of 63.45.*
Considering all options expiring on October 16, 2015 with strikes ranging from 1650-2050, we find that the Black-Scholes-Merton formula yields an av-

erage absolute relative pricing error[5] *of 7.3%, whereas the GARCH-approach yields a relative error of 9.5%.*

V.8 Additional literature

The pricing of options based on GARCH models dates back to the seminal work by Duan (1995), see also Taylor (2005, Chapter 14). Typically, the GARCH-based pricing methods rely on choosing a suitable SDF in order to find the dynamics of the log-returns under the equivalent martingale measure Q, see Appendix 1. Under certain conditions, the Q-dynamics are also GARCH-type, which allows for simulating the log-returns under Q, so that

$$P_t^{\text{Call}}(T,K) = E^Q[e^{-r(T-t)}\max(S_T - K, 0)|\mathcal{I}_t]$$

can be computed by simulations.

The literature on GARCH-based option pricing is huge and typically focus on various extensions of the GARCH model, such as incorporating skewness in the conditional return distributions (Christoffersen et al., 2006), jumps in underlying stock price (Christoffersen et al., 2008), and incorporating realized volatility (based on high frequency data) in the GARCH model (Christoffersen et al., 2014). A recent book on GARCH-based pricing is written by Chorro et al. (2015).

A rigorous, technical treatment of discrete-time finance is given in Föllmer and Schied (2011), see also Back (2017). For a derivation of the Black-Scholes formula in continuous time we refer to Taylor (2005) and Björk (2009).

[5] With $\{P_t^{\text{mkt}}(T,K_i) : i = 1,\dots,N\}$ the actual market prices of options with strikes $K_1,\dots,K_N$ and $\{P_t^{\text{model}}(T,K_i) : i = 1,\dots,N\}$ the model-based prices, the average absolute relative pricing error of the model is defined as $AARPE^{\text{model}} = N^{-1}\sum_{i=1}^{N}|(P_t^{\text{mkt}}(T,K_i) - P_t^{\text{model}}(T,K_i))/P_t^{\text{mkt}}(T,K_i)|$

References

Bachelier, L., 1900, *Théorie de la Spéculation*, PhD thesis, École normale supérieure.

Back, K.E., 2017, *Asset Pricing and Portfolio Choice Theory*, 2nd edition, Oxford University Press.

Björk, T., 2009, *Arbitrage Theory in Continuous Time*, 3rd edition, Oxford University Press.

Black, F.S. and Scholes, M.S., 1973, "The pricing of options and corporate liabilities", *Journal of Political Economy*, vol. 81, pp. 637–654.

Chorro, C., Guégan, D. & Ielpo, F., 2015, *A Time Series Approach to Option Pricing: Models, Methods, and Empirical Performances*, Springer-Verlag.

Christoffersen, P., Feunou, B., Jacobs, K. & Meddahi, N., 2014, "The economic value of realized volatility: Using high-frequency returns for option valuation", *Journal of Financial and Quantitative Analysis*, vol. 49, pp. 663-697.

Christoffersen, P, Heston, S. & Jacobs, K., 2006, "Option valuation with conditional skewness", *Journal of Econometrics*, vol. 131, pp. 253-284.

Christoffersen, P., Jacobs, K., Ornthanalai, C. & Wang, Y., 2008, "Option valuation with long-run and short-run volatility components", *Journal of Financial Economics*, vol. 90, pp. 272-297.

Drost, F.C. & Nijman, T.E., 1993, "Temporal aggregation of GARCH processes", *Econometrica*, vol. 61, pp. 909-927.

Duan, J.-C., 1995, "The GARCH option pricing model", *Mathematical Finance*, vol. 5, pp. 13-32.

Föllmer, H. & Schied, A., 2011, *Stochastic Finance: An Introduction in Discrete Time*, 3rd edition, Walter de Gruyter.

Garcia, R., Ghysels, E. & Renault, E., (2010), "The econometrics of option pricing" in *Handbook of Financial Econometrics: Tools and Techniques*, edited by L. P. Hansen and Y. Aït-Sahalia, ch. 9, pp. 479-552, North-Holland.

Merton, R.C., 1973, "Theory of rational option pricing", *The Bell Journal of Economics and Management Science*, Vol. 4, pp. 167–179.

Taylor, S.J., 2005, *Asset Price Dynamics, Volatility and Prediction*, Princeton University Press.

Zhu, K. & Ling, S., 2015, "Model-based pricing for financial derivatives", Journal of Econometrics, vol. 187, pp. 447-457.

Appendix 1: On P and Q

In light of the assumption about the joint conditional distribution of $\lambda_{t,t+1} := \log(m_{t,t+1})$ and $x_{t+1} := \log(S_{t+1}/S_t)$ in (V.21)-(V.22), consider their joint pdf $f_P(\lambda, x)$ conditional on $\mathcal{I}_t$, where P indicates that $\lambda_{t,t+1}$ and x_{t+1} are distributed according to the distribution P. In order to ease the notation we here suppress that the conditional distribution depends on t. Then by definition of the SDF

$$\begin{aligned}
1 &= E\left[\exp(\lambda_{t,t+1})\exp(x_{t+1})|\mathcal{I}_t\right] \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\exp(\lambda)\exp(x)f_P(\lambda,x)d\lambda dx \\
&= \int_{-\infty}^{\infty} e^{-r}\exp(x)\left(e^r\int_{-\infty}^{\infty}\exp(\lambda)f_P(\lambda,x)d\lambda\right)dx \\
&= \int_{-\infty}^{\infty} e^{-r}\exp(x)f_Q(x)dx
\end{aligned}$$

where $f_Q(x) := \left(e^r\int_{-\infty}^{\infty}\exp(\lambda)f_P(\lambda,x)d\lambda\right)$. Notice that $f_Q(x) > 0$ for all x and

$$\begin{aligned}
\int_{-\infty}^{\infty} f_Q(x)dx &= e^r\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\exp(\lambda)f_P(\lambda,x)d\lambda dx = e^r E[\exp(\lambda_{t,t+1})|\mathcal{I}_t] \\
&= e^r E[m_{t,t+1}|\mathcal{I}_t] = 1.
\end{aligned}$$

Hence the function $f_Q(x)$ can be viewed as a pdf of x_{t+1} under some probability measure Q such that

$$\int_{-\infty}^{\infty} e^{-r}x f_Q(x)dx = E^Q[e^{-r}x_{t+1}|\mathcal{I}_t].$$

We have that

$$E^Q\left[e^{-r}\frac{S_{t+1}}{S_t}|\mathcal{I}_t\right] = 1 \quad \Leftrightarrow \quad E^Q[e^{-r(t+1)}S_{t+1}|\mathcal{I}_t] = e^{-rt}S_t,$$

which means that the discounted stock price $e^{-rt}S_t$ is a martingale with respect to $\mathcal{I}_t$ under the measure Q. Due to this property Q is typically referred to as the "equivalent martingale measure" (or risk-neutral measure). In a vast proportion of the financial literature people choose to work with this measure instead of the SDF.
Moreover, notice that

$$\begin{aligned}
&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\exp(\lambda)\exp(x)f_P(\lambda,x)d\lambda dx \\
&=\int_{-\infty}^{\infty}\exp(x)\left(\int_{-\infty}^{\infty}\exp(\lambda)f_P(\lambda|x)d\lambda\right)f_P(x)dx \\
&=\int_{-\infty}^{\infty}\exp(x)m(x)f_P(x)dx,
\end{aligned}$$

where $m(x) := \left(\int_{-\infty}^{\infty}\exp(\lambda)f_P(\lambda|x)d\lambda\right)$, and in light of the derivations above,

$$m(x) = e^{-r}\frac{f_Q(x)}{f_P(x)}.$$

Hence we can write the SDF as

$$m_{t,t+1} = m(x_{t+1}) = e^{-r}\frac{f_Q(x_{t+1})}{f_P(x_{t+1})}.$$

It should be mentioned that in discrete time (as considered here) the stochastic discount factor as well as the equivalent martingale measure are typically not unique, meaning that there exist several prices of the European call option compatible with absence of arbitrage (technically: markets are incomplete). Hence, given the collection of all stochastic discount factors, one could compute an interval of prices spanning the no-arbitrage bounds. A more standard way to proceed is, as here, to pick a single measure Q or SDF compatible with absence of arbitrage.

Appendix 2: Technical proofs

V.8.1 Proof of Lemma V.2.1

Note that $X \overset{d}{=} s\exp(\varepsilon\sigma)$, where $s := \exp(\mu)$ and $\varepsilon \overset{d}{=} N(0,1)$. Hence,

$$\begin{aligned}
&E[\max(X-K,0)] = E[\max\{s\exp(\varepsilon\sigma)-K,0\}] \\
&= sE[\max\{\exp(\varepsilon\sigma)-K/s,0\}] = s\int_{-\infty}^{\infty}\max\{\exp(u\sigma)-K/s,0\}\phi(u)du,
\end{aligned}$$

where $\phi(\cdot)$ is the pdf of the standard normal distribution. Next, by straightforward integration,

$$
\begin{aligned}
& s \int_{-\infty}^{\infty} \max\{\exp(u\sigma) - K/s, 0\}\phi(u)du \\
&= s \int_{\log(K/s)/\sigma}^{\infty} [\exp(\sigma u) - K/s]\phi(u)du \\
&= s \left[\int_{\log(K/s)/\sigma}^{\infty} \exp(\sigma u)\phi(u)du - \int_{\log(K/s)/\sigma}^{\infty} (K/s)\phi(u)du \right] \\
&= s \left[\int_{\log(K/s)/\sigma}^{\infty} \exp(\sigma^2/2)\phi(u-\sigma)du - \int_{\log(K/s)/\sigma}^{\infty} (K/s)\phi(u)du \right] \\
&= s \left[\exp(\sigma^2/2) \int_{\log(K/s)/\sigma-\sigma}^{\infty} \phi(u)du - (K/s) \int_{\log(K/s)/\sigma}^{\infty} \phi(u)du \right] \\
&= s \exp(\sigma^2/2) \left(\int_{-\infty}^{\infty} \phi(u)du - \int_{-\infty}^{\log(K/s)/\sigma-\sigma} \phi(u)du \right) \\
&\qquad - s(K/s) \left(\int_{-\infty}^{\infty} \phi(u)du - \int_{-\infty}^{\log(K/s)/\sigma} \phi(u)du \right) \\
&= s \left[\exp(\sigma^2/2)(1 - \Phi(\log(K/s)/\sigma - \sigma)) - (K/s)(1 - \Phi(\log(K/s)/\sigma)) \right] \\
&= s \left[\exp(\sigma^2/2)\Phi(-\log(K/s)/\sigma + \sigma) - (K/s)\Phi(-\log(K/s)/\sigma) \right]
\end{aligned}
$$

where we have used that $\exp(\sigma u)\phi(u) = \exp(\sigma^2/2)\phi(u-\sigma)$ and $\Phi(x) = 1 - \Phi(-x)$ for $x \in \mathbb{R}$. The result now follows by collecting terms and using that, by definition, $s = \exp(\mu)$, and that $E[X] = \exp(\mu + \sigma^2/2)$. $\square$

V.8.2 Proof of Lemma V.2.2

By the law of iterated expectations,

$$
\begin{aligned}
E\left[\exp(X)g(Y)\right] &= E\left[E[\exp(X)g(Y)|Y]\right] = E\left[g(Y)E[\exp(X)|Y]\right] \\
&= E\left[g(Y) \exp\left(\mu_{X|Y} + \frac{1}{2}\sigma^2_{X|Y} \right) \right],
\end{aligned}
$$

where the last equality follows by Lemma V.8.1 below, and with $\mu_{X|Y} = \mu_X + \omega(Y - \mu_Y)$, $\omega := \sigma_{XY}/\sigma_Y^2$, and $\sigma^2_{X|Y} := \sigma^2_X - \omega\sigma_{XY}$. Next, we note that

$$
\begin{aligned}
&E\left[g(Y)\exp\left(\mu_{X|Y}+\frac{1}{2}\sigma^2_{X|Y}\right)\right]\\
&=\int_{-\infty}^{\infty} g(y)\exp\left(\mu_{X|y}+\frac{1}{2}\sigma^2_{X|Y}\right)f(y)dy\\
&=\int_{-\infty}^{\infty} g(y)\exp\left(\mu_X-\omega\mu_Y+\omega y+\frac{1}{2}\sigma^2_{X|Y}\right)f(y)dy\\
&=\exp\left(\frac{1}{2}\sigma^2_{X|Y}+\mu_X-\omega\mu_Y\right)\int_{-\infty}^{\infty} g(y)\exp\left(\omega y\right)f(y)dy\\
&=\exp\left(\frac{1}{2}\sigma^2_{X|Y}+\mu_X-\omega\mu_Y\right)\int_{-\infty}^{\infty} g(y)\frac{1}{\sqrt{2\pi\sigma^2_Y}}\exp\left(-\frac{(y-\mu_Y)^2}{2\sigma^2_Y}+\omega y\right)dy.
\end{aligned}
$$

Straightforward derivations yield that

$$
\exp\left(-\frac{(y-\mu_Y)^2}{2\sigma^2_Y}+\omega y\right)=\exp\left(-\frac{(y-\mu_Y-\sigma_{XY})^2}{2\sigma^2_Y}\right)\exp\left(\frac{\sigma^2_{XY}+2\mu_Y\sigma_{XY}}{2\sigma^2_Y}\right),
$$

which implies that

$$
\begin{aligned}
E\left[g(Y)\exp\left(\mu_{X|Y}+\frac{1}{2}\sigma^2_{X|Y}\right)\right]&=\exp\left(\frac{1}{2}\sigma^2_{X|Y}+\mu_X-\omega\mu_Y+\frac{\sigma^2_{XY}+2\mu_Y\sigma_{XY}}{2\sigma^2_Y}\right)\\
&\quad\times\int_{-\infty}^{\infty} g(y)\frac{1}{\sqrt{2\pi\sigma^2_Y}}\exp\left(-\frac{(y-\mu_Y-\sigma_{XY})^2}{2\sigma^2_Y}\right)dy\\
&=\exp\left(\frac{1}{2}\sigma^2_X+\mu_X\right)\\
&\quad\times\int_{-\infty}^{\infty} g(y)\frac{1}{\sqrt{2\pi\sigma^2_Y}}\exp\left(-\frac{(y-\mu_Y-\sigma_{XY})^2}{2\sigma^2_Y}\right)dy\\
&=E[\exp(X)]E\left[g(Y+\sigma_{XY})\right],
\end{aligned}
$$

where the last equality holds by observing that

$$
\frac{1}{\sqrt{2\pi\sigma^2_Y}}\exp\left(-\frac{(y-\mu_Y-\sigma_{XY})^2}{2\sigma^2_Y}\right)
$$

is the pdf of $Y+\sigma_{XY}\sim\mathrm{N}(\mu_Y+\sigma_{XY},\sigma^2_Y)$, and that $E[\exp(X)]=\exp\left(\frac{1}{2}\sigma^2_X+\mu_X\right)$ by (V.9). □

Lemma V.8.1 *Let $(X,Y)'$ be bivariate $N_2(\mu,\Omega)$-distributed with mean $\mu=E(X,Y)'$ and covariance matrix $\Omega=Var(X,Y)'$ given by,*

$$
\mu=\begin{pmatrix}\mu_X\\ \mu_Y\end{pmatrix},\ \Omega=\begin{pmatrix}\sigma^2_X & \sigma_{XY}\\ \sigma_{YX} & \sigma^2_Y\end{pmatrix}.
$$

Then

$$E[\exp(X)|Y = y] = \exp(\mu_{X|y} + \frac{1}{2}\sigma^2_{X|Y}),$$

where $\mu_{X|y} = \mu_X + \omega(y - \mu_Y)$, $\omega := \sigma_{XY}/\sigma^2_Y$, *and* $\sigma^2_{X|Y} := \sigma^2_X - \omega\sigma_{XY}$.

Proof: Recall from Chapter I that the conditional distribution of X given $Y = y$ is $N(\mu_{X|y}, \sigma^2_{X|Y})$. We conclude that $\exp(X)$ is conditionally (on $Y = y$) log-normal with parameters $(\mu_{X|y}, \sigma^2_{X|Y})$, and hence using a conditional version of (V.9) for $k = 1$, we have that $E[\exp(X)|Y = y] = \exp(\mu_{X|y} + \frac{1}{2}\sigma^2_{X|Y})$. □

V.8.3 Proof of Theorem V.3.1

We have that

$$P^{\text{Call}}_t(t + 1, K) = S_t e^{-r} E[\max(S_{t+1}/S_t - K/S_t, 0)|\mathcal{I}_t]. \tag{V.28}$$

Note that (S_{t+1}/S_t) is conditionally log-normal with mean e^r under the no-arbitrage condition (V.12). A straightforward application of Lemma V.2.1 yields that

$$\begin{aligned} E[\max(S_{t+1}/S_t - K/S_t, 0)|\mathcal{I}_t] &= E[S_{t+1}/S_t|\mathcal{I}_t]\Phi(-u_{t,t+1} + \sigma) - \frac{K}{S_t}\Phi(-u_{t,t+1}), \\ &= e^r\Phi(-u_{t,t+1} + \sigma) - \frac{K}{S_t}\Phi(-u_{t,t+1}), \end{aligned} \tag{V.29}$$

where $u_{t,t+1}$ is defined in (V.13). The result follows by combining (V.29) and (V.28). □

V.8.4 Proof of Lemma V.4.1

Equation (V.17) is immediate. Turning to proving (V.18), recall that under no arbitrage we have that (V.16) holds. Using that (V.15) and applying Lemma V.2.2, we have that

$$\begin{aligned} &E\left[\exp\left\{\log m_{t,t+1}\right\} \times \exp\left\{\log(S_{t+1}/S_t)\right\}|\mathcal{I}_t\right] \\ &= E\left[\exp\left\{\log m_{t,t+1}\right\}|\mathcal{I}_t\right] E[\exp\left\{\log(S_{t+1}/S_t)\right\}|\mathcal{I}_t] \\ &= E\left[m_{t,t+1}|\mathcal{I}_t\right] \exp(\sigma_{m,r}) E[S_{t+1}/S_t|\mathcal{I}_t] \\ &= e^{-r} \exp(\sigma_{m,r}) \exp(\mu + \sigma^2/2), \end{aligned}$$

where we have used (V.5), the fact that S_{t+1}/S_t is conditionally log-normal with parameters (μ, σ^2), and (V.9). Substituting into (V.16) and taking logs yield that

$$-r + \sigma_{m,r} + \mu + \frac{\sigma^2}{2} = 0,$$

and we conclude that (V.18) holds. $\square$

V.8.5 Proof of Theorem V.4.1

From (V.16), we have that

$$P_t^{\text{Call}}(t+1, K) = S_t e^{-r} E[\max(e^{\sigma_{m,r}}(S_{t+1}/S_t) - K/S_t, 0)|\mathcal{I}_t].$$

Note that $e^{\sigma_{m,r}}(S_{t+1}/S_t)$ is conditionally log-normal with parameters $(\mu + \sigma_{m,r}, \sigma^2)$. By Lemma V.2.1 we have that

$$E[\max(e^{\sigma_{m,r}}(S_{t+1}/S_t) - K/S_t, 0)|\mathcal{I}_t] = E[e^{\sigma_{m,r}}(S_{t+1}/S_t)|\mathcal{I}_t]\Phi(-\tilde{u}+\sigma) - \frac{K}{S_t}\Phi(-\tilde{u}),$$

where

$$\tilde{u} := \log\left(\frac{K/S_t}{\exp(\mu + \sigma_{m,r})}\right)/\sigma$$

and

$$E[e^{\sigma_{m,r}}(S_{t+1}/S_t)|\mathcal{I}_t] = \exp(\mu + \sigma_{m,r} + \sigma^2/2).$$

By Lemma V.4.1, we have that

$$\tilde{u} = \log\left(\frac{K/S_t}{\exp(r - \sigma^2/2)}\right)/\sigma = \log\left(\frac{Ke^{-(r-\sigma^2/2)}}{S_t}\right)/\sigma = u_{t,t+1},$$

with $u_{t,t+1}$ defined in (V.13), and that

$$E[e^{\sigma_{m,r}}(S_{t+1}/S_t)|\mathcal{I}_t] = e^r.$$

Collecting terms gives that

$$P_t^{\text{Call}}(t+1, K) = S_t\Phi(-u_{t,t+1} + \sigma) - \frac{K}{e^r}\Phi(-u_{t,t+1}). \quad \square$$

Anders Rahbek and Rasmus Søndergaard Pedersen
University of Copenhagen
October 2020

Part VI
Stochastic Volatility: An Introduction

VI.1 Introduction

The models we have discussed so far have all been defined by specifying explicitly the conditional distribution of the present observation (log-return), x_t, given past observations $x_{t-1}, \ldots, x_1$ as in the rich class of GARCH models, where

$$x_t = \sigma_t z_t,$$

with z_t *i.i.d.* $(0,1)$ and σ_t^2 a function of the past observations $x_{t-1}, \ldots, x_1$.

Now we focus on stochastic volatility (SV) models where σ_t (or σ_t^2) is an unobservable, or so-called latent, exogenous process. Typically we will assume that (σ_t) satisfies

$$(\sigma_t | \sigma_{t-1}, \sigma_{t-2}, ...) \stackrel{d}{=} (\sigma_t | \sigma_{t-1}), \tag{VI.1}$$

that is, it is a Markov chain. We will assume further that the volatility process (σ_t) is independent of the z_t's in the specification of x_t. Note that this excludes modeling the leverage effect as in e.g. the GJR-ARCH model, where negative returns are allowed to have a different effect on σ_t than the positive.

Example VI.1 *Given an i.i.d.* $(0,1)$ *sequence* (z_t)*, we say that* (x_t) *is a stochastic volatility (SV) model if*

$$x_t = \sigma_t z_t,$$

where (σ_t) *is a stochastic process that is independent of* (z_t) *and* $Var(\sigma_t | x_{t-1}, ...) > 0$*. The volatility* σ_t *satisfies (VI.1) and* σ_t *is positive and real-valued (*$\sigma_t \in \mathbb{R}_+$*), or* σ_t *takes values in some finite and discrete set, say* $A = \{h_1, h_2\}$*. Note that ARCH processes do not fit into this definition as* σ_t *depends on* x_{t-1} *and therefore on* z_{t-1}*.* □

In the following sections we go through some issues related to the use of SV models using that SV models are examples of nonlinear time series driven by exogenous Markov chains as described above. The positive message is that dependence (weak mixing, geometric ergodicity) properties of the observable series x_t may usually be addressed only by studying the latent chain (σ_t). This makes it relatively easy to derive asymptotic properties of estimators based on method of moments.

However, as we shall see, the likelihood function is not directly computable and therefore maximum likelihood analysis for these kind of models is not straightforward. We discuss some general filtering techniques that allow us to evaluate the log-likelihood function (approximately).

VI.2 Stochastic volatility models

We list below some consequences of the definition of the SV-model given in Example VI.1. It is assumed that the innovations (z_t) follow a distribution with a strictly positive density, f, on $\mathbb{R}$ as in the case of Gaussian or Student's t-distributed z_t.

Property 1: If (σ_t) is (strictly) stationary so is (x_t) as by definition $x_t = \sigma_t z_t$ and (σ_t) and (z_t) are independent.

Property 2: If (σ_t) is stationary then x_t follows a mixture distribution of the density f with respect to the invariant distribution of (σ_t), cf. Example VI.2.2.

Property 3: The observations (x_t) are uncorrelated – provided of course that correlations are well-defined, or equivalently, second order moments exist.

Property 4: Given, or conditional on, $(\sigma_T, \ldots, \sigma_1)$, then x_t and x_{t+h} are independent. The conditional distribution of x_t given $(\sigma_T, \ldots, \sigma_1)$, denoted $x_t|\sigma_T, \ldots, \sigma_1$, is equal to the distribution of $x_t|\sigma_t$.

These properties characterize the *(hidden Markov) process* (x_t, σ_t).

Our main focus is on two classical examples of a stochastic volatility models, introduced by Taylor and discussed extensively in Taylor (1986, 2005). Much like the development of ARCH models, the purpose was to construct a time series model with a changing volatility that was able to match marginal distributions and correlation structures found in typical financial time series. Taking as offset the multiplicative model of the form

$$x_t = \sigma_t z_t,$$

modelling (σ_t) as an autoregressive model turned out to be a suitable compromise between simplicity and the property of replicating empirical findings of real data. In addition, the case where σ_t can take only two (or more) different values corresponding to states, is also of interest, both from a theoretical and an empirical point of view.

VI.2.1 The log-normal SV model

The first example of a stochastic volatility model is the log-normal SV model as introduced in Taylor (1986). Here the observations are given by

$$x_t = \sigma_t z_t = \exp(\log(\sigma_t))z_t, \tag{VI.2}$$

where (z_t) is $i.i.d.\mathrm{N}(0,1)$ and independent of the volatility process (σ_t), which is unobservable. It is further assumed that $h_t = \log(\sigma_t)$ follows a first-order autoregressive (AR(1)) process,

$$h_t = (1-\phi)\alpha + \phi h_{t-1} + \eta_t,$$

with (η_t) $i.i.d.\mathrm{N}\big(0,\sigma_\eta^2\big)$, $\sigma_\eta^2 = \beta^2(1-\phi^2)$, and $\alpha, \beta \in \mathbb{R}$, $\phi \in (-1,1)$. Since the distribution of h_t is Gaussian (when h_1 is Gaussian) then the volatility $\sigma_t = \exp(h_t)$ marginally follows a log-normal distribution leading to the terminology *log-normal SV* model, or simply *log-SV*, for the model defined by (VI.2). We refer to Part V for additional details about and properties of the log-normal distribution. □

VI.2.2 A two-state SV model

Consider again the process in Example VI.1 given by

$$x_t = \sigma_t z_t, \quad t = 1, 2, \ldots$$

where (z_t) is $i.i.d.\mathrm{N}(0,1)$ and independent of the volatility process (σ_t). In the log-normal SV model, the unobserved volatility σ_t is continuous taking positive real values, $\sigma_t \in \mathbb{R}_+$, whereas in the two-state SV model it can take only two positve values, say h_1 and h_2, corresponding to *states* 1 and 2 respectively. In general, SV models can be formulated for any finite number of states, but to keep things simple (and because the two-state case is widely applied) we focus on the two-state case here.

Let s_t be a process that takes values in $\{1,2\}$. A classic *mixture Gaussian model* is obtained by letting

$$\sigma_t = h_{s_t}$$

and letting s_t be *i.i.d.* with $P(s_t = 1) = 1 - P(s_t = 2) = p$, that is

$$(s_t|s_{t-1}, ..., s_1) \stackrel{d}{=} s_t.$$

However, this does not replicate the realizations of real-world log-returns x_t very well, as the model essentially rules out volatility clustering due to the memoryless state variable (and hence volatility). Instead we consider the popular Markov switching SV model where s_t is a Markov chain (see below), in which

$$(s_t|s_{t-1}, ..., s_1) \stackrel{d}{=} (s_t|s_{t-1}).$$

A different class of models such as the threshold or the so-called ACR (autoregressive conditional root) volatility models, is given by letting s_t depend on past returns, that is for example,

$$(s_t|s_{t-1}, ..., s_1, x_{t-1}, ...x_1) \stackrel{d}{=} (s_t|x_{t-1}).$$

However, such observation switching models violates our initial set-up where σ_t is independent of z_t and we discuss this elsewhere.

VI.2.3 A note on Markov chains

Let s_t be a random variable that takes values in $\{1, 2\}$. The process $(s_t)_{t=1,2,...}$ is a Markov chain,

$$(s_t|s_{t-1}, ..., s_1) \stackrel{d}{=} (s_t|s_{t-1}), \tag{VI.3}$$

and we specify the switches between states 1 and 2 by the parameters $(p_{ij})_{i,j=1,2}$

$$p_{ij} := P(s_t = j|s_{t-1} = i) = "P_{\text{begin, end state}}".$$

By construction, it follows that

$$p_{11} + p_{12} = P(s_t = 1|s_{t-1} = 1) + P(s_t = 2|s_{t-1} = 1) = 1,$$

and likewise $p_{22} + p_{12} = 1$. When discussing the properties of s_t it is convenient to collect the transition probabilities in the so-called transition matrix P,

$$P = \begin{pmatrix} p_{11} & p_{21} \\ p_{12} & p_{22} \end{pmatrix} = \begin{pmatrix} p_{11} & 1 - p_{22} \\ 1 - p_{11} & p_{22} \end{pmatrix}.$$

VI.2.3.1 Some simple considerations and calculations

A way to interpret this, which may also be useful for simulations of s_t, given a known p_{ij}, is to let S_t (as opposed to s_t) be a bivariate variable with two values corresponding to s_t :

$$S_t = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \text{ if } s_t = 1, \quad \text{and} \quad S_t = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \text{ if } s_t = 2. \tag{VI.4}$$

Then, with say $s_t = 1$, one finds by definition,

$$\begin{aligned} P(s_{t+1} = 1 | s_t = 1) &= p_{11} \\ P(s_{t+1} = 2 | s_t = 1) &= p_{12} = 1 - p_{11} \end{aligned}$$

or simply,

$$\begin{pmatrix} P(s_{t+1} = 1 | s_t = 1) \\ P(s_{t+1} = 2 | s_t = 1) \end{pmatrix} = PS_t = \begin{pmatrix} p_{11} & 1 - p_{22} \\ 1 - p_{11} & p_{22} \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} p_{11} \\ 1 - p_{11} \end{pmatrix}$$

Thus if $s_t = 1$, then $s_{t+1} = 1$ with probability p_{11} and $s_{t+1} = 2$ with probability $1 - p_{11} = p_{12}$.

Next, using repeatedly that

$$P(X \in A | Y \in B) = \frac{P(X \in A, Y \in B)}{P(Y \in B)}$$

together with (VI.3), we can find the probability that s_t when initiated in $s_t = 1$ ends in state 1 after two steps,

$$\begin{aligned} &P(s_{t+2} = 1 | s_t = 1) \\ &= P(s_{t+2} = 1, s_t = 1) / P(s_t = 1) \\ &= \left(P(s_{t+2} = 1, s_{t+1} = 1, s_t = 1) + P(s_{t+2} = 1, s_{t+1} = 2, s_t = 1) \right) / P(s_t = 1) \\ &= P(s_{t+2} = 1 | s_{t+1} = 1) P(s_{t+1} = 1 | s_t = 1) + P(s_{t+2} = 1 | s_{t+1} = 2) P(s_{t+1} = 2 | s_t = 1) \\ &= p_{11} p_{11} + p_{21} p_{12}. \end{aligned}$$

Similarly,

$$P(s_{t+2} = 2 | s_t = 1) = p_{12} p_{11} + p_{22} p_{12}.$$

We note that the same could have been obtained by simply using the transition matrix P as follows,

$$\begin{pmatrix} P(s_{t+2} = 1 | s_t = 1) \\ P(s_{t+2} = 2 | s_t = 1) \end{pmatrix} = P(PS_t) = P^2 S_t.$$

Likewise, for any k number of steps,

$$\begin{pmatrix} P(s_{t+k}=1|s_t) \\ P(s_{t+k}=2|s_t) \end{pmatrix} = P^k S_t \tag{VI.5}$$

with S_t defined in (VI.4). That is, we can compute the probability for s_{t+k} taking values 1 or 2 given any starting value for s_t by simple *multiplication* of the transition matrix P.

VI.2.3.2 Weak mixing and Geometric Ergodicity

The regularity conditions we used for establishing that the ARCH processes, x_t say, were weakly mixing, were that the transition density, $f(x_t|x_{t-1})$ was sufficiently well-behaved (i.e. continuous or in the case of threshold models, positive and bounded on intervals). Analogously when studying the properties of the finite state Markov chain s_t we need the transition matrix P to have certain properties. These properties, as implied by the transition matrix P for the finite state Markov chain, namely *irreducibility* and *aperiodicity* (both explained below), are exactly the ones implied by the conditions on the transition density $f(\cdot|\cdot)$ for the x_t Markov chain with a 'continuum of states'.

First of all, we need that s_t when entering state 1 (or 2) is not staying there. This would be implied by for example $p_{11} = 1$ such that $p_{12} = 0$, in which case the chain is said to have an absorbing state 1 (as it would never enter state 2 again) and s_t is a *reducible* Markov chain. We need the chain to be *irreducible,* that is given any starting state $s_t = i$, then for some k,

$$P(s_{t+k} = j|s_t = i) > 0,$$

for any $j = 1, 2$. This is implied by the condition that,

$$p_{11},\ p_{22} < 1. \tag{VI.6}$$

Secondly, we need that the irreducible Markov chain is not *periodic* as would be the case if $p_{21} = p_{12} = 1$, or

$$P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

In this case with $s_t = 1$, then $s_{t+2} = s_{t+4} = ... = 1$ while $s_{t+1} = s_{t+3} = ... = 2$. That is, s_t with a 'period of 2' returns to 1 (and to 2 with the same period). For the irreducible chain to also be *aperiodic,* we need in addition to (VI.6) that,

$$p_{11} + p_{22} > 0.$$

Next, application of the drift-criterion to the irreducible and aperiodic chain is vacuous (as it would be also for the case where s_t has any finite number of states) and as a result the two conditions imply that s_t is weakly mixing (all moments finite), or geometrically ergodic.

Alternatively, a simple way to state the conditions is that, given that the irreducibility condition (VI.6) holds, the eigenvalues of P satisfy $\lambda_1 = 1$ and $|\lambda_2| < 1$. To see this, observe that the characteristic equation of P is

$$\det(\lambda I_2 - P) = (\lambda - 1)(\lambda + 1 - (p_{11} + p_{22})) = 0.$$

Note that the fact that each column of P sums to one, implies that one eigenvalue equals unity.

VI.2.3.3 Stationary distribution

Recall that weakly mixing implies that there exists a stationary distribution for the process considered. As in the AR(1) case, where this can be found directly, this can also be found directly for the 2-state Markov chain. The stationary distribution of s_t is characterized by the unconditional probabilities,

$$v = \begin{pmatrix} P(s_t = 1) \\ P(s_t = 2) \end{pmatrix} = \begin{pmatrix} P(s_t = 1) \\ 1 - P(s_t = 1) \end{pmatrix} = \begin{pmatrix} P(s_{t+k} = 1) \\ P(s_{t+k} = 2) \end{pmatrix},$$

for all k. In particular, we have by definition (recall also the way we compute moments for the AR(1) process when we impose stationarity),

$$\begin{aligned} \begin{pmatrix} P(s_{t+1} = 1) \\ P(s_{t+2} = 2) \end{pmatrix} &= \begin{pmatrix} P(s_{t+1} = 1|s_t = 1) P(s_t = 1) + P(s_{t+1} = 1|s_t = 2) P(s_t = 2) \\ P(s_{t+1} = 2|s_t = 1) P(s_t = 1) + P(s_{t+1} = 2|s_t = 2) P(s_t = 2) \end{pmatrix} \\ &= \begin{pmatrix} p_{11} & p_{21} \\ p_{12} & p_{22} \end{pmatrix} \begin{pmatrix} P(s_t = 1) \\ P(s_t = 2) \end{pmatrix} \\ &= P \begin{pmatrix} P(s_t = 1) \\ P(s_t = 2) \end{pmatrix}, \end{aligned}$$

or with $v = (P(s_t = 1), P(s_t = 2))'$,

$$v = Pv.$$

That is v, is the eigenvector of P corresponding to the eigenvalue $\lambda_1 = 1$ of P. Tedious calculations give that

$$v = \frac{1}{2 - p_{11} - p_{22}} \begin{pmatrix} 1 - p_{22} \\ 1 - p_{11} \end{pmatrix}. \tag{VI.7}$$

This can be useful for defining starting values in an algorithm for obtaining the MLE of p_{ij}.

Note that it can be shown that also (the Perron–Frobenius theorem),

$$\lim_{k\to\infty} P^k = v\,(1,1)\,,$$

which is a (2×2) matrix of rank one, with each column containing the stationary invariant distribution.

VI.2.3.4 Markov chain with N states

We end by briefly stating a summary for the case where $s_t \in \{1,2,...,N\}$ for some finite N. For this general case we define irreducibility as follows: For any $i,j = 1,..,N$ there exists an integer m_{ij} (potentially depending on i and j) such that

$$P\left(s_{t+m_{ij}} = i | s_t = j\right) > 0.$$

Again defining the transition matrix, P by (observe the transpose),

$$P' = (p_{ij})_{i,j=1,2,...,N}\,,$$

then s_t is weakly mixing provided that it is irreducible and that the eigenvalues of P satisfy, $\lambda_1 = 1$, and $|\lambda_i| < 1$ for $i = 2,...,N$. The stationary distribution is similarly given by the eigenvector v corresponding to $\lambda_1 = 1$, or equivalently,

$$\lim_{k\to\infty} P^k = v\,(1,1,...,1)\,,$$

where $v = (P(s_t = 1)\,,...P(s_t = N))'$. Specifically, with $0_{N\times 1} = (0,\ldots,0)'$ and $\iota = (1,\ldots,1)'$ respectively N-dimensional column vectors of zeros and ones, and noting that $\iota' v = 1$, it holds that

$$v = (A'A)^{-1}A'\begin{pmatrix} 0_{N\times 1} \\ 1 \end{pmatrix}, \tag{VI.8}$$

where $A = (I_N - P', \iota)'$.

VI.3 Weakly mixing and the drift criterion for SV models

In order to discuss the properties of the return process $x_t = \sigma_t z_t$, we first observe that x_t itself is not a Markov chain while for the joint evolution of the return process x_t and the volatility σ_t, we have by assumption,

$$((x_t,\sigma_t)\,|\,(x_{t-1},\sigma_{t-1})\,,...,(x_1,\sigma_1)) \overset{d}{=} ((x_t,\sigma_t)\,|\,(x_{t-1},\sigma_{t-1}))\,.$$

And moreover, the joint distribution of (x_t, σ_t) given (x_{t-1}, σ_{t-1}) is characterized by the distributions of

$$(x_t|\sigma_t) \quad \text{and} \quad (\sigma_t|\sigma_{t-1}) .$$

In the case where both have densities, as in the log-normal SV model, this can be written in terms of densities as,

$$f\left((x_t, \sigma_t) \,|\, (x_{t-1}, \sigma_{t-1})\right) = f\left(x_t|\sigma_t\right) f\left(\sigma_t|\sigma_{t-1}\right) .$$

Thus if we establish that σ_t is weakly mixing, then x_t should 'inherit the properties'. More precisely, the joint process (x_t, σ_t) should be weakly mixing, provided that σ_t is.

VI.3.1 Two-state SV model

Turn first the case where σ_t is switching between two states as given by the s_t process. We saw that verification of weak mixing for the two-state – and indeed N-state – case of σ_t was straightforward reflecting that σ_t could only take a finite number of values.

It then follows by Chen and Carrasco (2002, Proposition 4) that the joint Markov process, (x_t, σ_t) is weakly mixing if P satisfies the regularity conditions just discussed, see also Genon-Catalot et al. (2002) and Meitz and Saikkonen (2008).

VI.3.2 Log-SV models

The alternative log-normal SV model is an example of a SV model where σ_t is real and positive, $\sigma_t \in \mathbb{R}_+ = (0, \infty)$, as opposed to the finite state case. We could proceed as for the two-state SV model and quote results in the just mentioned reference Carrasco and Chen (2002) and also Meitz and Saikkonen (2008) where it is shown that under general conditions that if σ_t satisfies some drift criterion, so does (x_t, σ_t).

However, we can establish this ourselves directly using our developed theory. That is, we can apply the drift criterion in Part I, Assumptions I.3.1-I.3.2 and Theorem I.3.2. In order to allow for some flexibility in the choice of models for σ_t we shall formulate the result such that the log-normal case is a specific example. Hence we restate the SV model as:

Assumption VI.1 *Consider the SV model as given by*

$$x_t = \sigma_t z_t, \qquad t = 1, 2, ...T.$$

Assume (i) that z_t is i.i.d. $(0,1)$ with some continuous density f_z. Assume furthermore, (ii) that the unobservable σ_t is real and positive, $\sigma_t \in \mathbb{R}_+$, with

$$(\sigma_t|\sigma_{t-1}, \sigma_{t-2}, ...) \stackrel{d}{=} (\sigma_t|\sigma_{t-1}),$$

and such that the transition density $f_\sigma(\sigma_t|\sigma_{t-1})$ is continuous. Finally, assume (iii) that σ_t and $(z_1, \ldots, z_t)$ are independent.

Now we shall assume that we have used a drift criterion with a drift function δ to establish that σ_t is weakly mixing - this is often straightforward as we shall also see below for the log-normal case.

Theorem VI.1 *Let $x_t = \sigma_t z_t$ be a SV model satisfying Assumption VI.1, with σ_t a Markov chain on $\mathbb{R}_+ = (0, \infty)$ which satisfies a drift criterion with drift function δ. The drift function is assumed to be bounded on closed intervals of the form $[b, B]$ in $\mathbb{R}_+$ and such that*

$$\lim_{\sigma\to\infty} \delta(\sigma) = \infty \qquad \textit{and} \qquad \lim_{\sigma\to 0} \delta(\sigma) = \infty.$$

Then (x_t, σ_t) satisfies the drift criterion with drift function,

$$d(x, \sigma) = \delta(\sigma) + x^2/\sigma^2. \tag{VI.9}$$

Hence (x_t, σ_t) is weakly mixing and geometric ergodic, provided Assumption I.3.1 in Part I[1] holds for the joint process (x_t, σ_t).

Proof:
First of all, by Assumption VI.1, (x_t, σ_t) satisfies Assumption I.3.1 since

$$((x_t, \sigma_t) \,|\, (x_{t-1}, \sigma_{t-1}), ..., (x_1, \sigma_1)) \stackrel{d}{=} ((x_t, \sigma_t) \,|\, (x_{t-1}, \sigma_{t-1})),$$

and the joint distribution of (x_t, σ_t) given (x_{t-1}, σ_{t-1}) is characterized by

$$f((x_t, \sigma_t) \,|\, (x_{t-1}, \sigma_{t-1})) = f(x_t|\sigma_t) f_\sigma(\sigma_t|\sigma_{t-1}) = \frac{1}{|\sigma_t|} f_z\left(\frac{x_t}{\sigma_t}\right) f_\sigma(\sigma_t|\sigma_{t-1}).$$

Next, with $\delta : \mathbb{R} \to [1, \infty]$ the drift function for σ_t, we note that for there are $0 < m < M < \infty$, $\in \mathbb{R}_+$, such that $E[\delta(\sigma_t)|\sigma_{t-1} = \sigma] \leq C$ for $\sigma \in [m, M]$, while

$$E[\delta(\sigma_t)|\sigma_{t-1} = \sigma] \leq \phi\delta(\sigma),$$

[1] or a similar condition as Assumption II.1.1

for $\sigma \notin [m, M]$ and $\phi < 1$, see Assumption I.3.2.

Applying the proposed drift function $d(\cdot)$ for (x_t, σ_t) we get,

$$\begin{aligned} & E[d(x_t, \sigma_t)|(x_{t-1}, \sigma_{t-1}) = (x, \sigma)] \\ & = E[\delta(\sigma_t) + x_t^2/\sigma_t^2|(x_{t-1}, \sigma_{t-1})] \\ & = E[\delta(\sigma_t)|\sigma_{t-1} = \sigma] + E\left(z_t^2\right) \\ & \leq \phi\delta(\sigma) + C + 1 \\ & = \left[\frac{\phi\delta(\sigma) + C + 1}{\delta(\sigma) + x^2/\sigma^2}\right] d(x, \sigma). \end{aligned} \tag{VI.10}$$

We need to establish that the term

$$\left[\frac{\phi\delta(\sigma) + C + 1}{\delta(\sigma) + x^2/\sigma^2}\right] \leq \rho$$

for some $\rho < 1$, and with $\sigma \notin [r, R]$ and $x \notin [-X, X]$, where $r, R, X > 0$.

For $\phi < \rho < 1$ we can choose $r, R > 0$ such that for $\sigma \notin [r, R]$ it holds that

$$\frac{\phi\delta(\sigma) + C + 1}{\delta(\sigma) + x^2/\sigma^2} \leq \frac{\phi\delta(\sigma) + C + 1}{\delta(\sigma)} = \phi + \frac{C + 1}{\delta(\sigma)} \leq \rho.$$

Similarly, since δ is bounded on compact subsets of $\mathbb{R}_+$ we may choose $X > 0$ such that for $\sigma \in [r, R]$ and $x_t^2 > X^2$ we have,

$$\frac{\phi\delta(\sigma) + C + 1}{\delta(\sigma) + x^2/\sigma^2} \leq \rho.$$

Finally, for $(\sigma, x) \in [r, R] \times [-X, X]$,

$$E[d(x_t, \sigma_t)|(\sigma_{t-1}, y_{t-1})] \leq K.$$

□

Examples of drift functions δ that satisfy the assumptions of the theorem are for example,

$$\delta(\sigma) = 1 + \sigma^2 + \sigma^{-2} \quad \text{and} \quad \delta(\sigma) = 1 + \left(\log \sigma^2\right)^2. \tag{VI.11}$$

VI.3.2.1 The log-normal SV

For the log-normal case, the volatility (σ_t) is a Markov chain with transition density which is continuous and hence bounded on closed intervals in $\mathbb{R}_+$ since $\sigma_t|\sigma_{t-1}$ is log-normal distributed.

With drift function $\delta(\sigma) = 1 + (\log \sigma_t)^2 = 1 + h_t^2$, where $h_t = \log(\sigma_t)$, follows a first order autoregressive AR(1)) process, $h_t = \gamma + \phi h_{t-1} + \eta_t$ with $\gamma := (1-\phi)\alpha$, $\sigma_\eta^2 := \beta^2(1-\phi^2)$, and $\phi \in (-1,1)$, we get directly that

$$\begin{aligned} E(\delta(\sigma_t)|\sigma_{t-1} = \sigma) &= 1 + \gamma^2 + \phi^2 h^2 + \sigma_\eta^2 + 2\gamma\phi h \\ &= \left(\frac{1 + \gamma^2 + \phi h^2 + \sigma_\eta^2 + 2\gamma\phi h}{1 + h^2} \right) \delta(\sigma) \end{aligned}$$

with $h = \log(\sigma^2)$. Hence, since $|\phi| < 1$, then for $\sigma \notin [r, R]$, $r, R > 0$ and hence $h \notin [\log r^2, \log R^2]$,

$$\left(\frac{1 + \gamma^2 + \phi^2 h^2 + \sigma_\eta^2 + 2\gamma\phi h}{1 + h^2} \right) \leq \rho,$$

with $|\phi| < \rho < 1$. From Theorem VI.1 we deduce that (x_t, σ_t) is geometrically ergodic in this case and hence satisfies the LLN and CLT for weakly mixing processes.

As a consequence the marginal process, (x_t), admits a stationary distribution. To find the invariant distribution remember that (x_t) is stationary if (σ_t), or equivalently (h_t), is stationary. From well known facts about the AR(1) process we know that in the Gaussian case, the invariant distribution for h_t is a Gaussian distribution with

$$E(h_t) = \alpha, \quad V(h_t) = \beta^2. \tag{VI.12}$$

Moreover, (h_t) is a Gaussian process with covariances given by

$$Cov(h_t, h_{t+k}) = \phi^k \beta, \quad k \geq 0.$$

Thus the stationary or invariant distribution of (x_t, σ_t) has density (see the discussion on the log-normal distribution),

$$\begin{aligned} f(\sigma, y) &= f(y|\sigma) f(\sigma) \\ &= (\frac{1}{\sigma} f_z(y/\sigma)) \left(\frac{1}{\sigma} \frac{1}{\sqrt{2\pi\beta}} \exp\left(-\frac{1}{2\beta} (\log(\sigma) - \alpha)^2 \right) \right). \end{aligned}$$

and the stationary version of (x_t) satisfies the LLN.

In particular, the marginal process (x_t) is stationary with an invariant distribution given by the density

$$\phi(x) = \int_{\mathbb{R}} f(\sigma, x) d\sigma.$$

For future reference we note that the $k-$th order moment of x_t exists if $E|z_t|^k < \infty$ in which case it holds that (using that σ_t is log-normal)

$$Ex_t^k = \exp\left(\beta \frac{k^2}{2} + k\alpha \right) Ez_t^k.$$

References

Carrasco, M. and X. Chen, 2002, "β-mixing and Moment properties of Various GARCH, Stochastic Volatility and ACD Models", *Econometric Theory*,

Taylor, S.J., 2005, *Asset Price Dynamics, Volatility and Prediction*, Princeton University Press

Taylor, S.J., 1986, *Modelling Financial Time Series*, John Wiley, Chichester.

Genon-Catalot, V., T. Jeantheau, and C. Laredo, 2000, "Stochastic Volatility Models as Hidden Markov Models and Statistical Applications", *Bernoulli* 6, 1051-1079.

Meitz, M. and P. Saikkonen, 2008, "Ergodicity, Mixing and Existence of Moments for a class of Markov Models with Applications to GARCH and ACD Models", *Econometric Theory*.

Anders Rahbek
Rasmus Søndergaard Pedersen
University of Copenhagen

October 2021

Part VII
Estimation of Finite State SV Models

In this chapter we discuss how to obtain likelihood-based estimators in SV models, where the switching driven by a finite-state first-order Markov chain, such as the two-state SV model:

Example VII.1 *The two-state SV model is given by,*

$$x_t = \sigma_t z_t,$$

for $t = 1, 2, ..., T$ and where the innovations z_t are i.i.d.$N(0,1)$ and independent of the unobserved volatility process (σ_t). The volatility is given by the unobserved switching process (the Markov chain) $s_t \in \{1, 2\}$, such that

$$\sigma_t = \begin{cases} h_1 & \text{if } s_t = 1 \\ h_2 & \text{if } s_t = 2 \end{cases},$$

and where the switching between the two states is governed by the transition matrix P given by

$$P = \begin{pmatrix} p(s_t = 1|s_{t-1} = 1) & p(s_t = 1|s_{t-1} = 2) \\ p(s_t = 2|s_{t-1} = 1) & p(s_t = 2|s_{t-1} = 2) \end{pmatrix} = \begin{pmatrix} p_{11} & p_{21} \\ p_{12} & p_{22} \end{pmatrix} = \begin{pmatrix} p_{11} & 1 - p_{22} \\ 1 - p_{11} & p_{22} \end{pmatrix}.$$

Hence the parameters we wish to estimate are

$$\theta = \{h_1, h_2, p_{11}, p_{22}\},$$

with $h_1, h_2 \in (0, \infty)$ and $p_{11}, p_{22} \in (0, 1)$. Much like in the GARCH model where the initial value σ_0^2 is unknown, the initial value s_0 (and hence σ_0^2) is not known.

Example VII.2 *For the general N-state case where $x_t = \sigma_t z_t$, the unobserved volatility sequence can take N values $h_1, ..., h_N$ respectively, according*

to the finite state switching process s_t on $\{1, ..., N\}$. The distribution of the one-step transition of the Markov chain s_t is given by the transition matrix

$$P = \begin{pmatrix} p_{11} & \cdots & p_{N1} \\ \vdots & \ddots & \vdots \\ p_{1N} & \cdots & p_{NN} \end{pmatrix},$$

where $p_{ij} = P(s_t = j \mid s_{t-1} = i)$. The parameters to be estimated are therefore $\theta = (h_1, ..., h_N, p_{11}, ..., p_{NN})$, where for $(p_{11}, ..., p_{NN})'$ we note the restriction that each column of the transition matrix must sum to one, that is for example $1 = p_{11} + ... + p_{1N}$.

Even though the two-state model is simple, it is widely applied, and moreover, as we shall see, somewhat involved to do estimation in. By definition the two-state SV model is a special case of the general finite (N-)state SV model and we shall formulate the results in a way such that also this model can be handled. In fact, the finite state SV models can also be viewed as examples of the wide class of so-called Hidden Markov Models (HMM) and/or Markov Switching (MS) models, which are widely applied – also in the field of macroeconometrics and microeconometrics, see also Section VII.5 below.

In terms of estimation, we note that writing down the log-likelihood function for the observations,

$$(x_1, ..., x_T),$$

is slightly involved, due to the fact that the path of $(s_t : t = 1, \ldots, T)$ is unobserved to the econometrician, which we will discuss in the following section. Further references for general MS models include Hamilton (1994, ch.22), who discuss their use in macro time series, MacDonald and Zucchini (1997), who provide a general statistical introduction, and Lange and Rahbek (2008), who focus on financial time series.

VII.1 Maximum likelihood estimation (MLE)

First, we make some initial considerations about the likelihood function and introduce some useful notation. Recall that we observe (the log-returns),

$$X_{0:T} = (x_0, ..., x_T),$$

while we do not observe the volatility and the switching variables,

$$S_{0:T} = (s_0, ..., s_T).$$

VII.1.1 A side note: Treating s_t as observed

If the switching variables, or the states, $S_{0:T}$ were observed the likelihood function could be computed as follows:

First note that as in autoregressive models, we can factorize the density conditional on (x_0, s_0) as follows (using $f(\cdot)$ to denote generic densities),

$$\begin{aligned} f(X_{1:T}, S_{1:T}|x_0, s_0) &= f(x_1, s_1, ..., s_T, x_T|x_0, s_0) \\ &= \prod_{t=1}^{T} f(x_t, s_t|x_{t-1}, s_{t-1}, ..., x_0, s_0) \\ &= \prod_{t=1}^{T} f(x_t|s_t) f(s_t|x_{t-1}, s_{t-1}, ..., x_0, s_0) \\ &= \prod_{t=1}^{T} f(x_t|s_t) p_{s_{t-1}s_t}, \end{aligned} \tag{VII.1}$$

where we have used that s_t depends only on s_{t-1} and not the remaining past values. Here, as z_t are $i.i.d.\mathrm{N}(0,1)$, we have

$$f(x_t|s_t) = \frac{1}{\sqrt{2\pi h_{s_t}^2}} \exp\left(-\frac{x_t^2}{2h_{s_t}^2}\right),$$

while to emphasize that s_t is characterized by probabilities rather than a density, we apply the notation $p_{s_{t-1}s_t}$ from the definition of the transition matrix P, that is

$$p_{s_{t-1}s_t} = p_{ij} \quad \text{if } (s_{t-1}, s_t) = (i, j).$$

Likewise, we use the notation

$$h_{s_t} = \begin{cases} h_1 & \text{if } s_t = 1 \\ h_2 & \text{if } s_t = 2 \end{cases}.$$

To get the full density, which is not conditional on (x_0, s_0), we can write,

$$f(X_{0:T}, S_{0:T}) = f(x_1, s_1, \ldots, , x_T, s_T|x_0, s_0) f(x_0, s_0).$$

Now in the two-state case s_1 can be either 1 or 2. Recall that the invariant probabilities were computed in Part VI as

$$v = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} p(s_1 = 1) \\ p(s_1 = 2) \end{pmatrix} = \begin{pmatrix} \frac{1}{2-p_{11}-p_{22}}(1-p_{22}) \\ \frac{1}{2-p_{11}-p_{22}}(1-p_{11}) \end{pmatrix}, \tag{VII.2}$$

and we could therefore write

$$f(x_0, s_0) = f(x_0|s_0)\, v_{s_0}.$$

This way we have the full likelihood function in terms of $\theta = (h_1, h_2, p_{11}, p_{22})'$. However, it is simpler to set $v := (v_1, 1 - v_1)$ and extend θ to include also v_1, or even to set for example $v_1 = \frac{1}{2}$ (that is, fixed and known). With large samples these choices matter very little.

Likewise, for the N-state case where $v = (v_1, ..., v_N)'$ one often instead of using the invariant probabilities expressed in terms of parameters in θ, choose to just let the parameters to be estimated to be (θ, v), or alternatively fix v as for example $v = \left(\frac{1}{N}, ..., \frac{1}{N}\right)'$ as in the two-state case. The latter is obviously the easiest.

In any case, collecting all terms we get the full log-likelihood function to be,

$$L_T(X_{0:T}, S_{0:T}; \theta) = \log\left(v_{s_0} f(x_0|s_0) \prod_{t=1}^{T} p_{s_{t-1}s_t} f(x_t \mid s_t)\right), \qquad \text{(VII.3)}$$

where $v = (v_1, ..., v_N)'$.[1]

VII.1.2 Treating s_t as unobserved

As s_t is not observed, the likelihood function must be evaluated based solely on the sequence $(x_1, ..., x_T)$. The direct way to do this is by simply summing (VII.3) over all possible paths for the state process s_t, that is

$$L_T(X_{0:T}; \theta) = \sum_{(s_0,...,s_T)\in\{1,...,N\}^{T+1}} L_T(X_{0:T}, S_{0:T}; \theta).$$

Unfortunately, this formulation leads to an exponentially growing number of terms, N^{T+1}, and is therefore infeasible even for moderate sample sizes. One could actually overcome this and making it grow linearly in T by formulating the likelihood function in terms of matrix products. However, the likelihood function is non-linear in θ and there are complex restrictions limiting the parameter space (primarily coming from the restrictions on the parameters

[1] Alternatively, one could consider the conditional log-likelihood based on (VII.1),

$$L_{T,c}(x_0, ..., x_T, s_0, ..., s_T; \theta) = \log\left(\prod_{t=1}^{T} p_{s_{t-1}s_t} f(x_t \mid s_t)\right).$$

in the transition matrix P), which means that we will take the alternative route of applying a so-called filtering algorithm as is quite standard in the literature, and which avoids these problems.

For the case of $N = 2$ states, we note that the density of $x_t|X_{0:t-1}$ is a weighted mixture of the two possible states,

$$\begin{aligned} f\left(x_t|X_{0:t-1}\right) &= f\left(x_t, s_t = 1|X_{0:t-1}\right) + f\left(x_t, s_t = 2|X_{0:t-1}\right) \\ &= f\left(x_t|s_t = 1, X_{0:t-1}\right) P\left(s_t = 1|X_{0:t-1}\right) \\ &\quad + f\left(x_t|s_t = 2, X_{0:t-1}\right) P\left(s_t = 2|X_{0:t-1}\right) \\ &= f\left(x_t|s_t = 1\right) P\left(s_t = 1|X_{0:t-1}\right) \\ &\quad + f\left(x_t|s_t = 2\right) P\left(s_t = 2|X_{0:t-1}\right) \end{aligned}$$

with weights given by the predicted probabilities, $P\left(s_t = i|X_{0:t-1}\right)$, $i = 1, 2$. By construction, the predicted probability, $P\left(s_t = i|X_{0:t-1}\right)$, is the best guess of the probability of regime i at time t given the observations until time $t-1$, $X_{0:t-1}$. The log-likelihood function (conditional on some known x_0) is hence given by

$$\begin{aligned} L_T(x_1, ..., x_T; \theta) &= \sum_{t=1}^{T} \log f\left(x_t|X_{0:t-1}; \theta\right) \\ &= \sum_{t=1}^{T} \log \left(\sum_{i=1}^{2} f\left(x_t|s_t = i; \theta\right) P\left(s_t = i|X_{0:t-1}; \theta\right) \right). \end{aligned} \tag{VII.4}$$

Note that

$$f\left(x_t|s_t = i; \theta\right) = \frac{1}{\sqrt{2\pi h_i^2}} \exp\left(-\frac{x_t^2}{2h_i^2}\right) \tag{VII.5}$$

is straightforward to evaluate given θ, and the challenging part of the estimation is to evaluate the predicted probabilities $P\left(s_t = i|X_{0:t-1}; \theta\right)$. We will consider an algorithm that evaluates the probabilities recursively.

VII.1.3 Filtering Algorithm

Suppose that we already have some values of the conditional regime probabilities at time $t-1$, $P\left(s_{t-1} = i|X_{0:t-1}\right)$, which is typically labelled the *filtered*

probability. Then the predicted probabilities are simply given by

$$\begin{aligned} P\left(s_t = i | X_{0:t-1}; \theta\right) &= P\left(s_t = i, s_{t-1} = 1 | X_{0:t-1}; \theta\right) + P\left(s_t = i, s_{t-1} = 2 | X_{0:t-1}; \theta\right) \\ &= P\left(s_t = i | s_{t-1} = 1, X_{0:t-1}; \theta\right) P(s_{t-1} = 1 | X_{0:t-1}; \theta) \\ &\quad + P\left(s_t = i | s_{t-1} = 2, X_{0:t-1}; \theta\right) P\left(s_{t-1} = 2 | X_{0:t-1}; \theta\right) \\ &= P\left(s_t = i | s_{t-1} = 1; \theta\right) P(s_{t-1} = 1 | X_{0:t-1}; \theta) \\ &\quad + P\left(s_t = i | s_{t-1} = 2; \theta\right) P\left(s_{t-1} = 2 | X_{0:t-1}; \theta\right) \\ &= \sum_{j=1}^{2} p_{ji} P\left(s_{t-1} = j | X_{0:t-1}; \theta\right). \quad \textbf{(prediction step)} \end{aligned} \tag{VII.6}$$

In order to make the algorithm recursive, we need a way to filtered probability at time t, $P\left(s_t = i | X_{0:t}; \theta\right)$, based on the predicted probability $P\left(s_t = i | X_{0:t-1}; \theta\right)$. To do so, we make use of the definition of conditional probability:

$$P\left(s_t = i | X_{0:t}; \theta\right) = \frac{f\left(s_t = i, x_t | X_{0:t-1}; \theta\right)}{f\left(x_t | X_{0:t-1}; \theta\right)}.$$

The numerator, given by the joint conditional density of x_t and $\{s_t = i\}$, equals

$$\begin{aligned} f\left(s_t = i, x_t | X_{0:t-1}; \theta\right) &= f\left(x_t | s_t = i, X_{0:t-1}; \theta\right) P\left(s_t = i | X_{0:t-1}; \theta\right) \\ &= f\left(x_t | s_t = i; \theta\right) P\left(s_t = i | X_{0:t-1}; \theta\right), \end{aligned}$$

and hence is a product of the conditional density in (VII.5) and predicted probability in (VII.6). Likewise, the denominator is the likelihood contribution in (VII.4). Hence, we have that the filtered probability is

$$P\left(s_t = i | X_{1:t}; \theta\right) = \frac{f\left(x_t | s_t = i; \theta\right) P\left(s_t = i | X_{1:t-1}; \theta\right)}{\sum_{i=1}^{2} f\left(x_t | s_t = i; \theta\right) P\left(s_t = i | X_{1:t-1}; \theta\right)}, \quad \textbf{(filtering step)}$$

and we emphasize that all terms are straightforward to evaluate. Given an initial value $P\left(s_0 = i | x_0; \theta\right)$ for $i = 1, 2$, e.g. $P\left(s_0 = i | x_0; \theta\right) = P(s_0 = i)$, we note that all predicted and filtered probabilities at times $t = 1, \ldots, T$ are determined recursively by the outlined steps, which are typically referred to as the *Hamilton algorithm*.

VII.1.4 Asymptotic Distribution and Inference

Asymptotic theory for the maximum likelihood estimators in HMMs is an active area of research, see, e.g., Kasahara and Shimotsu (2019) for a recent overview and the references herein. Proving asymptotic normality of

the maximum likelihood estimator, $\hat{\theta}_T$, in HHMs is typically much more demanding compared to the proofs for the GARCH-type models considered in Parts I-III. However, there exist high level conditions for asymptotic normality implying that

$$\sqrt{T}\left(\hat{\theta}_T - \theta_0\right)$$

has the usual asymptotic distribution in terms of the inverse of the information, see, e.g., Bickel et al. (1999) and Douc et al. (2004). The conditions are essentially that s_t is weakly mixing for our N-state SV models, and that θ_0 is an interior point of the parameter space.

An empirically important hypothesis to test is how many regimes are needed in order to characterize a given data series, e.g. if only $N = 1$ regime is needed instead of $N = 2$ regimes. Testing such a hypothesis is complicated due to the fact that parameter(s) of state 2, say the volatility h_2, is not identified under the null hypothesis of only one state. This violates the standard conditions of likelihood ratio testing, and the LR statistic does not have a standard χ^2 distribution. Several non-standard testing methods have been suggested in order to circumvent this issue, see e.g. Davies (1987), Andrews and Ploberger (1994), Hansen (1996), Cho and White (2007), and Carrasco et al. (2014).

VII.2 Predicted, Filtered, and Smoothed probabilities

Note that a by-product of the filtering algorithm, that evaluates the log-likelihood function, is that it produces the predicted probability $P\left(s_t = i|X_{0:t-1};\theta\right)$, i.e. the probability of being in state i in the next period. Such a probability is typically interesting for prediction purposes. As an example, one may quantify the probability of being in a high volatility state tomorrow given the information available today. Predicted probabilities extend to longer horizons, i.e. $P\left(s_{t+k} = i|X_{0:t};\theta\right)$ for any $k > 1$. Likewise, the filtered probability, $P\left(s_t = i|X_{0:t};\theta\right)$, may be used for addressing the probability of being in a high volatility regime today taking into account today's return y_t. This is sometimes referred to as real-time regime classification.

Lastly, it might be of interest to characterize the regime probabilities for a period given the information up to today, i.e. to compute the so-called smoothed probabilities $P\left(s_t = i|X_{0:T};\theta\right)$, where we note that the conditioning is based on all observations $X_{0:T} = (x_0, \ldots, x_T)$. Whereas the predicted and filtered probabilities are given recursively forward in time, the smoothed

probabilities are given by combining forward and backward recursions, as outlined in the following.

For $t = 1, ..., T$ define for $j = 1, 2$ the densities

$$\begin{aligned} a_t(j) &= f(x_0, ..., x_t, s_t = j; \theta), \\ b_t(j) &= f(x_{t+1}, ..., x_T \mid s_t = j; \theta), \end{aligned}$$

such that $a_t(j)$ denotes the joint density of the first t observations and of s_t. Similarly, $b_t(j)$ denotes the density of $x_{t+1}, \ldots, x_T$ given $s_t = j$. In terms of a_t and b_t, the smoothed probabilities $P(s_t = i|X_{0:T}; \theta)$ for $i = 1, 2$ can be computed as

$$\begin{aligned} &P(s_t = i|X_{0:T}; \theta) \\ &= \frac{f(x_0, ..., x_T, s_t = i; \theta)}{f(x_0, ..., x_T; \theta)} \\ &= \frac{f(x_0, ..., x_T, s_t = i; \theta)}{\sum_{j=1}^{2} f(x_0, ..., x_T\ , s_t = j; \theta)} \\ &= \frac{f(x_{t+1}, ..., x_T|x_0, ..., x_t, s_t = i; \theta) f(x_0, ..., x_t, s_t = i; \theta)}{\sum_{j=1}^{2} f(x_{t+1}, ..., x_T|x_1, ..., x_t, s_t = j; \theta) f(x_0, ..., x_t, s_t = j; \theta)} \\ &= \frac{f(x_{t+1}, ..., x_T|s_t = i; \theta) f(x_0, ..., x_t, s_t = i; \theta)}{\sum_{j=1}^{2} f(x_{t+1}, ..., x_T|s_t = j; \theta) f(x_0, ..., x_t, s_t = j; \theta)} \\ &= \frac{b_t(i) a_t(i)}{\sum_{j=1}^{2} b_t(j) a_t(j)}. \end{aligned} \tag{VII.7}$$

Note that we in particular used that

$$f(x_{t+1}, ..., x_T|x_1, ..., x_t, s_t = i) = f(x_{t+1}, ..., x_T|s_t = i),$$

which follows by the definition of the SV model with s_t the underlying Markov chain.

The sequence $a_t(\cdot)$ can be computed recursively using the *forward* algorithm given by

$$\begin{aligned} a_t(j) &= f(s_t = j, x_1, ..., x_t; \theta) \\ &= \sum_{i=1}^{2} f(s_{t-1} = i, s_t = j, x_1, ..., x_t) \\ &= \sum_{i=1}^{2} f(x_t|s_t = j; \theta) p_{ij} a_{t-1}(i). \end{aligned} \tag{VII.8}$$

The algorithm is initiated by putting $a_0(j) = f(x_0|s_0 = j;\theta)v_j$ for some fixed distribution $v = (v_1, v_2)$ on $\{1, 2\}$ for example $v = \left(\frac{1}{2}, \frac{1}{2}\right)$.

Finally the sequence $b_t(\cdot)$ can be computed using the *backward* algorithm given by

$$\begin{aligned} b_t(i) &= f(x_{t+1}, ..., x_T|s_t = i;\theta) \\ &= \sum_{j=1}^{2} f(s_{t+1} = j, x_{t+1}, ..., x_T|s_t = i;\theta) \\ &= \sum_{j=1}^{2} f\left(x_{t+1}|s_{t+1} = j;\theta\right) f\left(x_{t+2}, ..., x_T|s_{t+1} = j;\theta\right) f\left(s_{t+1} = j|s_t = i;\theta\right) \\ &= \sum_{j=1}^{2} f(x_{t+1}|s_{t+1} = j;\theta) b_{t+1}(j) p_{ij}. \end{aligned} \tag{VII.9}$$

The algorithm can be initiated by putting $b_T(j) = 1$ for all $j = 1, 2$.

VII.3 The General Case

In the following we give a brief outline of recursions in relation to the general N-state model introduced in Example VII.2. Most of the quantities are written on vector form, which is convenient for implementation in a programming language. For $s_t \in \{1, \ldots, N\}$, define the N-dimensional vector of conditional densities,

$$\eta_{t,\theta} = \begin{pmatrix} f(x_t|s_t = 1, X_{0:t-1};\theta) \\ \vdots \\ f(x_t|s_t = N, X_{0:t-1};\theta) \end{pmatrix},$$

and the predicted and filtered probabilities.

$$\xi_{t|t-1,\theta} = \begin{pmatrix} P(s_t = 1|X_{0:t-1};\theta) \\ \vdots \\ P(s_t = N|X_{0:t-1};\theta) \end{pmatrix} \quad \text{and} \quad \xi_{t|t,\theta} = \begin{pmatrix} P(s_t = 1|X_{0:t};\theta) \\ \vdots \\ P(s_t = N|X_{0:t};\theta) \end{pmatrix}.$$

The log-likelihood function for the observations $x_1, \ldots, x_T$ is then given by

$$\begin{aligned} L_T(\theta) &= \sum_{t=1}^{T} \log\left(\sum_{i=1}^{N} f\left(x_t|s_t = i;\theta\right) P\left(s_t = i|X_{0:t-1};\theta\right)\right) \\ &= \sum_{t=1}^{T} \log\left(\eta'_{t,\theta}\xi_{t|t-1,\theta}\right). \end{aligned}$$

Moreover, the filtering algorithm for $t = 1, \ldots, T$ is

$$\xi_{t|t,\theta} = \frac{1}{\eta'_{t,\theta}\xi_{t|t-1,\theta}} \left(\eta_{t,\theta} \odot \xi_{t|t-1,\theta}\right) \quad \text{(Filtering step)},$$
$$\xi_{t+1|t,\theta} = P\xi_{t|t,\theta} \quad \text{(Prediction step)},$$

where $\odot$ denotes element-by-element (so-called Hadamard) multiplication. The recursions may be initiated at some given probabilities such as the uniform $\xi_{1|0,\theta} = (N^{-1}, \ldots, N^{-1})'$ or at the stationary probabilities $\xi_{1|0,\theta} = v$, with v defined in (VI.10) in Part VI.

Moreover, it is straightforward to show that the k-step predicted probabilities are given by

$$\xi_{t+k|t,\theta} = \begin{pmatrix} P(s_{t+k} = 1|X_{0:t};\theta) \\ \vdots \\ P(s_{t+k} = N|X_{0:t};\theta) \end{pmatrix} = P^k \xi_{t|t,\theta},$$

and, lastly, the smoothed probabilities are given backwards by

$$\xi_{t|T,\theta} = \begin{pmatrix} P(s_t = 1|X_{0:T};\theta) \\ \vdots \\ P(s_t = N|X_{0:T};\theta) \end{pmatrix} = \xi_{t|t,\theta} \odot \left(P' \left[\xi_{t+1|T,\theta} \oslash \xi_{t+1|t,\theta}\right]\right),$$

where $\oslash$ denotes element-by-element division (see e.g. Hamilton, 1994, ch.22). The backward recursions may be initiated at the time T filtered probabilities, $\xi_{T|T,\theta}$.

VII.4 An Alternative Estimation Approach

An alternative to maximizing the log-likelihood function is the so-called Expectation-Maximization (EM) algorithm. The main idea is that if the latent process, $(s_1, \ldots, s_T)$, was observed, then it would be straightforward to maximize the joint log-likelihood function, as derived in (VII.3). The EM algorithm relies on maximizing the expected log-likelihood with the expectation taken conditionally over the observed sample $(x_1, \ldots, x_T)$, and this can be shown to yield the MLE. The expected log-likelihood function is derived via an algorithm similar to the one used for obtaining the aforementioned smoothed probabilities. This is known as the expectation step (E-step) of the algorithm.

VII.5 Extensions

The MS SV model can be extended in many directions. First, the classical AR and ARCH models considered in the first chapters may have regime switching parameters. Specifically, we may consider an MS-AR model with stochastic volatility where for $x_t \in \mathbb{R}$,

$$x_t = \mu_{s_t} + \rho_{s_t} x_{t-1} + \sigma_{s_t} z_t, \quad z_t \ i.i.d. N(0,1),$$

with $(\mu_{s_t}, \rho_{s_t}, \sigma_{s_t}) = (\mu_i, \rho_i, h_i) \in \mathbb{R} \times \mathbb{R} \times (0, \infty)$ when $s_t = i$, $i = 1, \ldots, N$, and (s_t) an N-state Markov chain, and where (s_t) and (z_t) are independent. Obviously, we could let z_t have another distribution, e.g. a Student's t-distribution (potentially with regime-switching degrees of freedom, in order to allow for time-varying tail heaviness). Likewise, we may as in Cai (1994) consider an MS-ARCH model, with

$$\begin{aligned} x_t &= \sigma_t z_t, \quad z_t \ i.i.d. N(0,1) \\ \sigma_t^2 &= \omega_{s_t} + \alpha_{s_t} x_{t-1}^2, \end{aligned}$$

with $(\omega_{s_t}, \alpha_{s_t}) = (\omega_i, \alpha_i) \in (0, \infty) \times [0, \infty)$ when $s_t = i$, $i = 1, \ldots, N$, and with (s_t) and (z_t) independent .

We may also consider multivariate extensions, such as MS-VAR models with stochastic covolatility, where for $x_t \in \mathbb{R}^d$

$$x_t = \mu_{s_t} + A_{s_t} x_{t-1} + \Sigma_{s_t}^{1/2} z_t, \quad z_t \ i.i.d. N(0, I_d),$$

with μ_{s_t} and A_{s_t} are respectively $(d \times 1)$ and $(d \times d)$ matrices and $\Sigma_{s_t}^{1/2}$ is the matrix square-root of a $(d \times d)$ positive definite matrix Σ_{s_t} – all allowed to be regime switching. Such a model has been widely used in empirical macro and finance, such as in Guidolin and Timmermann (2004) who apply the model to detect "bull" and "bear" states in UK stock and bond markets, and to construct optimal portfolio allocation, taking into account potential switching in economic regimes, see also Barberis (2000). Note that all of these extensions are, by definition, not SV models. Estimation in such models are more or less straightforward, and the filtering algorithms remains (up to some additional conditioning) essentially unchanged.

Lastly, one may allow for (strictly) exogenous covariates in the models and allow the transition probabilities to be time-varying, e.g. by letting the state probabilities depend on lagged observations, $x_1, \ldots, x_{t-1}$, or covariates; see e.g. Bec et al. (2008).

References

Andrews, D.W.K. and W. Ploberger, 1994, "Optimal Tests when a Nuisance Parameter is Present Only Under the Alternative", *Econometrica*, 62: 1383–1414.

Barberis, N., 2000, "Investing for the Long Run when Returns Are Predictable", *The Journal of Finance*, 55, 225–264.

Bec, F., A. Rahbek, and N. Shephard, 2008, "The ACR Model: A Multivariate Dynamic Mixture Autoregression", *Oxford Bulletin of Economics and Statistics*, 70, 583–618.

Bickel, P.J., Y. Ritov, and T. Rydén, 1998, "Asymptotic Normality of the Maximum-Likelihood Estimator for General Hidden Markov Models", *Annals of Statistics*, 26, 1614–1635.

Cai, J., 1994, "A Markov Model of Switching-Regime ARCH", *Journal of Business & Economic Statistics*, 12, 309–316.

Carrasco, M., L. Hu, and W. Ploberger, 2014, "Optimal Test for Markov Switching Parameters, *Econometrica*, 82, 765–784.

Cho, J.S. and H. White, 2007, "Testing for Regime Switching", *Econometrica*, 75m 1671–1720.

Davies, R.B., 1987, "Hypothesis Testing when a Nuisance Parameter is Present Only Under the Alternative", *Biometrika*, 74, 33–43.

Douc R., É. Moulines, and T. Rydén, 2004, "Asymptotic Properties of the Maximum Likelihood Estimator in Autoregressive Models with Markov Regime", *Annals of Statistics*, 32, 2254–2304.

Guidolin, M. and A. Timmermann, 2004, "Economic Implications of Bull and Bear Regimes in UK Stock and Bond Returns", *The Economic Journal*, 115, 111–143.

Hamilton, J.D., 1994, *Time Series Analysis*, Princeton University Press.

Hansen, B.E., 1996, "Inference when a Nuisance Parameter is Not Identified Under the Null Hypothesis", *Econometrica*, 64, 413–430.

Kasahara, H. and K. Shimotsu, 2019, "Asymptotic Properties of the Maximum Likelihood Estimator in Regime Switching Econometric Models", *Journal of Econometrics*, 208, 442–467.

Lange T. and A. Rahbek, 2008, "Regime Switching Time Series Models: A Survey", *Handbook of Financial Time Series*.

MacDonald, I.L. and W. Zucchini, 1997, *Hidden Markov and Other Models for Discrete-Valued Time Series*, Taylor & Francis.

Anders Rahbek
Rasmus Søndergaard Pedersen
University of Copenhagen

October 2021

Part VIII
SV: QMLE and the Kalman Filter

In this chapter we consider QML-based estimation of the log-SV model. Due to the fact that the volatility process $\{\sigma_t\}_{t=1}^T$ is unobserved, one cannot write down the log-likelihood function explicitly. Instead, we consider a log-linearization of the model, and use a filtering method, the so-called Kalman Filter, to compute a Gaussian quasi-log-likelihood function.

VIII.1 The log-Normal SV Model

Recall that the log-SV model for log-returns y_t is given by,

$$x_t = \sigma_t z_t, \quad t = 1, ..., T \tag{VIII.1}$$

with $\{z_t\}_{t=1}^T$ *i.i.d.* $\mathrm{N}(0,1)$ and

$$\log \sigma_t = (1-\phi)\alpha + \phi \log \sigma_{t-1} + \kappa_t \tag{VIII.2}$$

with $\{\kappa_t\}_{t=1}^T$ *i.i.d.* $\mathrm{N}(0, \sigma_\kappa^2)$, and independent of $\{z_t\}_{t=1}^T$. Moreover, $\phi \in (-1,1)$ and

$$\sigma_\kappa^2 := \beta^2 \left(1 - \phi^2\right),$$

such that the three parameters to be estimated are given by,

$$\theta = (\alpha, \beta, \phi)' \in \mathbb{R} \times (0, \infty) \times (-1, 1).$$

Recall that, since $|\phi| < 1$, the stationary version of $\log \sigma_t$ is unconditionally normal with mean and variance given, respectively by $E \log \sigma_t = \alpha$, and $V \log \sigma_t = \beta^2$.

VIII.1.1 The Likelihood function

With observations $\{x_t\}_{t=1}^T$ from the log-SV model, the log-likelihood function is

$$\ell_T(\theta) = \log L_T(x_1, ..., x_T \ ; \ \theta) = \log f_\theta(x_1) + \sum_{t=2}^{T} \log f_\theta(x_t | x_{t-1}, ..., x_1),$$

with f_θ denoting a generic density function parametrized by θ. However, even in the simple case of the standard SV model in (VIII.1), $\ell_T(\theta)$ cannot be computed - let alone stated in any closed-form - since $f_\theta(x_1)$ and $f_\theta(x_t|x_{t-1}, ..., x_1)$ are unknown. Likewise, if one used the representation in terms of the unobserved σ_t,

$$\begin{aligned} L_T(x_1, ..., x_T \ ; \ \theta) &= \int L_T(x_1, ..., x_T, \sigma_1, ..., \sigma_T \ ; \ \theta) \, d(\sigma_1, ..., \sigma_T) \\ &= \int [\prod_{t=2}^{T} f_\theta(x_t|\sigma_t) \, f_\theta(\sigma_t|\sigma_{t-1})] f_\theta(x_1|\sigma_1) \, f_\theta(\sigma_1) \, d(\sigma_1, ..., \sigma_T) \end{aligned}$$

the likelihood value cannot be computed, since the state space for (the Markov chain) $\log \sigma_t$ is $\mathbb{R}$, making the integration infeasible.

Note that in the case where σ_t had a finite valued (2 or more) state space, we demonstrated in Part VII that a filtering algorithm could be used to obtain the MLE of θ. With an infinite state space as here, various alternative routes have been followed to overcome the problem of finding the MLE. Some are simulation-based and approximate the likelihood function by various simulation schemes, see below for a quick introduction.

Some compute moments of x_t and apply the so called method of moments to obtain non-MLE estimators and study their properties, see Andersen and Sørensen (1996) and Kim et al. (1998) for discussion of some of the problems with method of moments.

Here we discuss in detail the Quasi-MLE (QMLE) approach where the state space form (SSF) is used to define a quasi-likelihood function which can be recursively computed by the Kalman filter. Thus the QMLE of θ, $\hat{\theta}_T$, is not the MLE as we shall see, due to simplifying assumptions (basically those of Gaussanity). However, in practice often the QMLE is used. This is despite the fact that it can have poor small sample properties.

The discussion will be structured as follows. After a brief introduction to simulation-based estimation, an introduction to the state space model is given. Next, MLE is discussed for linear Gaussian state space models based on the Kalman filter. Finally, the QMLE is discussed for the standard log-SV model.

VIII.1.1.1 The idea of simulated likelihood - a detour

The idea behind this can be – very – briefly described as follows:

Suppose that we wish to compute the density (and hence the likelihood) function for the first observed variable x_1 with σ_0 fixed. Now, by definition,

$$f(x_1|\sigma_0) = \int f(x_1, \sigma_1|\sigma_0)\, d\sigma_1 = \int f(x_1|\sigma_1)\, f(\sigma_1|\sigma_0)\, d\sigma_1,$$

where $f(x_1|\sigma_1)$ is the Gaussian $\mathrm{N}(0, \sigma_1^2)$ density, while $f(\sigma_1|\sigma_0)$ is the density of σ_1 given σ_0. As the distribution of $\log \sigma_1$ conditional on $\log \sigma_0$ is known, so is $f(\sigma_1|\sigma_0)$, such that the density $f(\sigma_1|\sigma_0)$ can be used to simulate draws of σ_1 from (with σ_0 fixed). Let $i = 1, ..., N$ be simulations of σ_1 which are denoted $\sigma_1(i)$, then one can approximate $f(x_1|\sigma_0)$ by,

$$f(x_1|\sigma_0) = \int f(x_1|\sigma_1)\, f(\sigma_1|\sigma_0)\, d\sigma_1 \simeq f^{\mathrm{a}}(x_1|\sigma_0) := \frac{1}{N}\sum_{i=1}^{N} f(x_1|\sigma_1(i)).$$

For the next x_2 conditional on x_1 and σ_0, analogously,

$$f(x_2|x_1, \sigma_0) = \int f(x_2|\sigma_2)\, f(\sigma_2|x_1, \sigma_0)\, d\sigma_2 \simeq f^{\mathrm{a}}(x_2|x_1, \sigma_0) = \frac{1}{N}\sum_{i=1}^{N} f(x_2|\sigma_2(i)).$$

However, here we need to simulate, or sample, $\sigma_2(i)$ conditional on (x_1, σ_0), which is not straightforward (in particular as we do not observe σ_1). Rewriting gives,

$$f(\sigma_2|x_1, \sigma_0) = \frac{f(\sigma_2|\sigma_1)\, f(\sigma_1|\sigma_0)\, f(x_1|\sigma_1)}{f(x_1|\sigma_0)} = \frac{K(\sigma_2, \sigma_1, \sigma_0, x_1)}{f(x_1|\sigma_0)}$$

Thus the problem is how to sample $\sigma_2(i)$ from a density expressed in terms of $K(\cdot)$ and $f(x_1|\sigma_0)$. The three densities in $K(\cdot)$ are known, but the latter we do not know so little can be done.

However, we can compute an approximate value $f^a(x_1|\sigma_0)$ as above for the $f(x_1|\sigma_0)$ term, and next we may use the trick to insert some "sampling density $\phi(\sigma_2)$" of our own choice and write:

$$\begin{aligned} f(x_2|x_1, \sigma_0) &= \int f(x_2|\sigma_2)\, f(\sigma_2|x_1, \sigma_0)\, d\sigma_2 = \int f(x_2|\sigma_2)\, \frac{f(\sigma_2|x_1\sigma_0)}{\phi(\sigma_2)}\phi(\sigma_2)\, d\sigma_2 \\ &\simeq f^a(x_2|x_1, \sigma_0) = \frac{1}{N}\sum f(y_2|\sigma_2(i)) \left(\frac{f(\sigma_2(i)|x_1, \sigma_0)}{\phi(\sigma_2(i))}\right) \end{aligned}$$

where $\sigma_2(i)$ is now drawn from $\phi(\cdot)$, and

$$f(\sigma_2(i)|x_1, \sigma_0) = \frac{f(\sigma_2(i)|\sigma_1(i))\, f(\sigma_1(i)|\sigma_0)\, f(x_1|\sigma_1(i))}{f^a(x_1|\sigma_0)}.$$

Clearly, the choice of $\phi(\cdot)$ is important, and one should aim at choosing it "close" to $f(\sigma_2(i)|x_1, \sigma_0)$.

For $t \geq 2$ the equivalent problems occur, and it should be clear that computing the likelihood this way is non-trivial.

VIII.2 The State Space Model

We consider here the linear Gaussian state space model formulated in terms of two sets of equations. The first set of equations states the dynamics of the observed variable $Z_t \in \mathbb{R}^p$, in terms of the unobserved variable, $X_t \in \mathbb{R}^q$, and an innovation, w_t. The second set of equations states the dynamics of the unobserved (state) variable X_t, $X_t \in \mathbb{R}^q$, in terms of its own past and an innovation v_t. Specifically, for $t = 1, 2, ..., T$,

$$Z_t = AX_t + w_t, \tag{VIII.3}$$
$$X_t = \Phi X_{t-1} + v_t, \tag{VIII.4}$$

with some initial value X_0. The innovation w_t is p-dimensional and the innovation v_t q-dimensional with $(w_t, v_t)'$ *i.i.d.* $\mathrm{N}_{p+q}(0, \Omega)$, with covariance,

$$\Omega = \begin{pmatrix} \Omega_w & 0 \\ 0 & \Omega_v \end{pmatrix}.$$

Finally, A is a $(p \times q)$-dimensional matrix, and Φ $(q \times q)$ dimensional.

Equations (VIII.3) and (VIII.4) constitute the state space form (SSF), where equation (VIII.3) is referred to as the observation equation, and equation (VIII.4) is referred to as the state equation, specifying that the unobserved variable X_t as a (vector) autoregression of order one.

In terms of the initialization of the system, one may keep X_0 fixed, or one may consider X_0 as drawn from the $\mathrm{N}_q(\mu_x, \Omega_x)$ distribution, independently of $\{(w_t, v_t)'\}_{t=1}^T$. This we will return to when discussing the likelihood-based estimation. Let us consider first the log-SV model written in SSF:

Example VIII.1 (The log-SV model) *The log-Normal SV model can not be written in the SSF directly. However, an approximate SSF exists, which can be seen by considering the following transformation of the model. Define the observed variable* Z_t *as*

$$Z_t = \log|x_t|,$$

and set $w_t = \log|z_t| - \mu_z$, $\mu_z = E\log|z_t|$. *With* z_t $N(0,1)$, *then* $\mu_z = E\log|z_t| = -0.63518...$ *and* $\Omega_w = V(\log|z_t|) = \pi^2/8$. *Hence* w_t *is indeed* $i.i.d.(0, \Omega_w)$ *but not Gaussian.*

Moreover, with this notation, and $p = 1$, $q = 2$, one can write the transformed log-Normal SV model as,

$$Z_t = (1,1)\, X_t + w_t$$
$$X_t = \begin{pmatrix} \log \sigma_t - \alpha \\ \alpha + \mu_z \end{pmatrix} = \begin{pmatrix} \phi & 0 \\ 0 & 1 \end{pmatrix} X_{t-1} + \begin{pmatrix} \kappa_t \\ 0 \end{pmatrix},$$

or

$$Z_t = AX_t + w_t$$
$$X_t = \Phi X_{t-1} + v_t$$

where $A = (1,1)$, $\Phi =$diag$(\phi, 1)$ and $v_t = (\kappa_t, 0)'$.

In particular, the SV model is not a Gaussian state space model as w_t is not Gaussian, while v_t indeed is (singular) Gaussian with

$$\Omega_v = \begin{pmatrix} \beta^2 (1 - \phi^2) & 0 \\ 0 & 0 \end{pmatrix}.$$

The fact that w_t is not Gaussian is ignored in the estimation of the parameters $\theta = (\alpha, \phi, \beta^2)'$, which explains the use of the terminology Quasi-MLE (QMLE).

Note that the SSF is not unique in this case, and more generally, often different state space representations exist for the same model(s).

VIII.2.1 State Space Model Likelihood

For the state space model in (VIII.3) and (VIII.4), the log-likelihood in terms of the observations $\{Z_t\}_{t=1}^T$ is given by,

$$l_T(\theta) = \log L_T(Z_1, ..., Z_T\ ;\ \theta) = \log f_\theta(Z_1) + \sum_{t=2}^{T} \log f_\theta(Z_t | Z_{t-1}, ..., Z_1), \tag{VIII.5}$$

where for $f_\theta(Z_1)$ the initial distribution of X_0 is used as by definition $Z_1 = AX_1 + w_1$, and $X_1 = \Phi X_0 + v_1$.

Furthermore, under the assumption of normality of $(w_t, v_t)'$ and X_0 (and hence X_1), we can see that the process $(X_1, Z_1, X_2, Z_2, ..., X_T, Z_T)$ is multivariate Gaussian. In particular, this means that $(Z_1, ..., Z_T)$ is Gaussian and the conditional distribution $Z_t | Z_{t-1}, ..., Z_1$ as well.

Hence for a given value of the parameter θ, the likelihood in (VIII.5) is completely specified by the conditional mean and variance,

$$E(Z_t | Z_{t-1}, ..., Z_1) \quad \text{and } V(Z_t | Z_{t-1}, ..., Z_1), \tag{VIII.6}$$

in addition to the initial distribution of Z_1, as noted above. The thing to emphasize here is that $\sum_{t=2}^{T} \log f_\theta (Z_t|Z_t, ..., Z_1)$ can be calculated from $E(X_t|Z_{t-1}, ..., Z_1)$ and $V(X_t|Z_{t-1}, ..., Z_1)$ due to Gaussianity. Moreover, the conditional mean and variance can be recursively calculated by the Kalman filter, as explained below.

VIII.2.2 The Kalman Filter

The Kalman filter is a recursive algorithm from which the conditional moments in (VIII.6) can be obtained. Define for $0 \leq s \leq t$, the conditional moments:

$$X_{t|s} = E(X_t|Z_s, ..., Z_1) \quad \text{and} \quad \Omega_{t|s} = V(X_t|Z_s, ..., Z_1) \tag{VIII.7}$$

We may set $X_{0|0} = \mu_x$ and $\Omega_{0|0} = \Omega_x$ corresponding to $X_0 = N(\mu_x, \Omega_x)$.

Straightforward calculations as demonstrated below, see also Shumway and Stoffer (2000) and Hamilton (1994), give the following lemma:

Lemma VIII.1 (Kalman) *Consider the Gaussian state space model given by (VIII.3) and (VIII.4). With initial conditions $X_{0|0} = \mu_x$ and $\Omega_{0|0} = \Omega_x$, the conditional moments in (VIII.7) for $t = 1, 2, ..., T$ are given by:*

$$X_{t|t-1} = \Phi X_{t-1|t-1}, \tag{VIII.8}$$
$$\Omega_{t|t-1} = \Phi \Omega_{t-1|t-1} \Phi' + \Omega_v, \tag{VIII.9}$$

with

$$X_{t|t} = X_{t|t-1} + K_t \left(Z_t - A X_{t|t-1}\right), \tag{VIII.10}$$
$$\Omega_{t|t} = (I - K_t A) \Omega_{t|t-1}, \tag{VIII.11}$$

where

$$K_t \equiv \Omega_{t|t-1} A' \left(A \Omega_{t|t-1} A' + \Omega_w\right)^{-1}. \tag{VIII.12}$$

Proof of Lemma VIII.1: To see (VIII.8), set $Z_{1:t-1} = (Z_1, ..., Z_{t-1})$ and observe that

$$E(X_t|Z_{1:t-1}) = E(\Phi X_{t-1} + v_t|Z_{1:t-1}) = \Phi E(X_{t-1}|Z_{1:t-1}) = \Phi X_{t-1|t-1}$$

as $E(v_t|\mathbb{Z}_{t-1}) = 0$ by definition. Likewise, (VIII.9) holds as

$$V(X_t|Z_{1:t-1}) = \Phi \Omega_{t-1|t-1} \Phi' + \Omega_v.$$

Next, using obvious notation, recall from well-known results on the Gaussian distribution, that if

$$\left(\begin{array}{c} X \\ Y \end{array} \middle| Z \right) \stackrel{d}{=} \mathrm{N}(\left(\begin{array}{c} \mu_{x|z} \\ \mu_{y|z} \end{array} \right), \left(\begin{array}{cc} \Omega_{xx|z} & \Omega_{xy|z} \\ \Omega_{yx|z} & \Omega_{yy|z} \end{array} \right)),$$

then

$$\begin{aligned} E\left(X|Y,Z\right) &= \mu_{x|z} + \Omega_{xy|z}\Omega_{yy|z}^{-1}\left(Y - \mu_{y|z}\right) \quad \text{and} \\ V\left(X|Y,Z\right) &= \Omega_{xx|z} - \Omega_{xy|z}\Omega_{yy|z}^{-1}\Omega_{yx|z}. \end{aligned}$$

Set now $X = X_t$, $Y = Z_t$ and $Z = \mathbb{Z}_{t-1} = (Z_{t-1}, ..., Z_1)$, then

$$\begin{aligned} X_{t|t} &= E\left(X_t|Z_t, Z_{1:t-1}\right) \\ &= E\left(X_t|Z_{1:t-1}\right) + \mathrm{Cov}\left(X_t, Z_t|Z_{1:t-1}\right)\mathrm{V}\left(Z_t|Z_{1:t-1}\right)^{-1}\left(Z_t - E\left(Z_t|Z_{1:t-1}\right)\right). \end{aligned} \tag{VIII.13}$$

Next, $E\left(X_t|Z_{1:t-1}\right) = X_{t|t-1}$,

$$\mathrm{Cov}\left(X_t, Z_t|Z_{1:t-1}\right) = \mathrm{Cov}\left(X_t, AX_t + w_t|Z_{1:t-1}\right) = \Omega_{t|t-1}A', \tag{VIII.14}$$

$$\mathrm{V}\left(Z_t|Z_{1:t-1}\right) = \mathrm{V}\left(AX_t + w_t|Z_{1:t-1}\right) = A\Omega_{t|t-1}A' + \Omega_w, \tag{VIII.15}$$

and finally,

$$E\left(Z_t|Z_{1:t-1}\right) = AX_{t|t-1}. \tag{VIII.16}$$

Inserting (VIII.14)-(VIII.16) in (VIII.13) gives directly (VIII.10). Likewise for (VIII.11), where

$$\begin{aligned} \Omega_{t|t} &= \mathrm{V}\left(X_t|Z_t, Z_{1:t-1}\right) \\ &= V\left(X_t|Z_{1:t-1}\right) - \mathrm{Cov}\left(X_t, Z_t|Z_{1:t-1}\right)\mathrm{V}\left(Z_t|Z_{1:t-1}\right)^{-1}\mathrm{Cov}\left(Z_t, X_t|Z_{1:t-1}\right). \end{aligned} \tag{VIII.17}$$

□

VIII.2.3 Computing the likelihood

The likelihood value for a given parameter value θ can be computed by using Lemma VIII.1 and (VIII.16) as well as (VIII.15). First note that, we have directly that,

$$\mathrm{V}\left(Z_t|Z_{1:t-1}\right) = A\Omega_{t|t-1}A' + \Omega_w \qquad \text{and} \qquad E\left(Z_t|Z_{1:t-1}\right) = AX_{t|t-1}.$$

This means that for $t = 1, 2, ..., T$, we can write the likelihood in terms of the Gaussian innovations,

$$\epsilon_t := Z_t - AX_{t|t-1}, \tag{VIII.18}$$

with conditional mean zero, and conditional variance,

$$\Sigma_t = A\Omega_{t|t-1}A' + \Omega_w. \tag{VIII.19}$$

Hence the Gaussian likelihood function equals,

$$l_T(\theta) = \log L_T(Z_1, ..., Z_T\ ;\ \theta) = \log f_\theta(Z_1) + \sum_{t=2}^{T} \log f_\theta(Z_t|Z_t, ..., Z_1) \tag{VIII.20}$$

$$= -\frac{pT}{2}\log(2\pi) - \frac{1}{2}\sum_{t=1}^{T}\left(\log|\Sigma_t| + \epsilon_t'\Sigma_t^{-1}\epsilon_t\right) \tag{VIII.21}$$

As to the initial conditions for $X_{0|0}$ and $\Omega_{0|0}$, we use as mentioned, $X_{0|0} := \mu_x$ and $\Omega_{0|0} := \Omega_x$, corresponding to X_0 chosen as $\mathrm{N}(\mu_x, \Omega_x)$. Hence, the parameters μ_x and Ω_x may be fixed or considered as additional parameters to be found in the QML algorithm maximizing $l_T(\theta)$.

Example VIII.2 *For the log-SV model, since* $|\phi| < 1$, $\log\sigma_t - \alpha$ *has a stationary representation with*

$$E(\log\sigma_t - \alpha) = 0 \text{ and } V(\log\sigma_t - \alpha) = \beta^2.$$

Hence, X_0 *can be given the initial distribution,* $N(\mu_x, \Omega_x)$ *with*

$$\mu_x = (0, \alpha + \mu_z)' \quad \text{and} \quad \Omega_x = diag\left(\beta^2, 0\right).$$

VIII.2.4 Asymptotic theory

Estimating the parameters in the state space model in (VIII.3)-(VIII.4) using the described Kalman-Filter gives the MLE $\hat{\theta}_T$ if $(w_t, v_t)'$ are $i.i.d.$ jointly Gaussian as stated, while $\hat{\theta}_T$ is the QMLE if this assumption does not apply. From Watson (1989), see also Ruiz (1994) and Dunsmuir (1979), it holds that $\sqrt{T}(\hat{\theta}_T - \theta_0)$ is asymptotically Gaussian under the assumption that (the observations) Z_t are stationary and weakly mixing, and provided $(w_t, v_t)'$ are martingale differences with finitie fourth order moments. Note that this in particular requires that Φ has eigenvalues smaller than one in absolute value, see also Shumway and Stoffer (2000).

References

Andersen, T. & B. Sørensen (1996), 'GMM estimation of a stochastic volatility model: a Monte Carlo study', Journal of Business and Economic Statistics 14, 328-52.

Dunsmuir, W. (1979), A Central Limit Theorem for parameter estimation in stationary time series and its applications to models for a signal observed with noise, Annals of Statistics, 7:490-506.

Kim, S., N. Shephard & S. Chib (1998), 'Stochastic volatility: likelihood inference and comparison with ARCH models', Review of Economic Studies 65, 361-393.

Koopman, S.J., N. Shephard & J.A. Doornik (1999), 'Statistical algorithms for models in state space using SsfPack 2.2', Econometrics Journal 2, 113-166.

Ruiz, E. (1994), Quasi Maximum Likelihood Estimation of Stochastic Volatility Models, Journal of Econometrics, 63:289–306

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

Shumway, R.H. & D.S. Stoffer (2000), Time Series Analysis and its Applications, Springer-Verlag New York.

Taylor, S.J. (2005), Asset Prices Dynamics, Volatility, and Prediction, Princeton University Press. 6

Watson, M.W. (1989), Recursive Solution Methods for Dynamic Linear Rational Expectations Models, Journal of Econometrics, 41:65-89

Anders Rahbek and Rasmus Søndergaard Pedersen
University of Copenhagen

November 2022

Part IX
RV: Realized Volatility

This lecture note accompanies Chapter 12-13 of Taylor (2005) on realized volatility. Realized volatility plays a major role in modern financial econometrics and we shall give an introduction to the basic issues of the currently expanding area. We refer to McAleer and Medeiros (2008) and Andersen and Benzoni (2009) for a short introduction to the topic. A recent textbook dedicated to realized volatility (and related topics) is Aït-Sahalia and Jacod (2014).

The realized volatility is simple to compute as can be briefly exemplified (details below): The realized volatility for a given trading day $m \in \{1, ..., M\}$ based on N log-returns (that is, log-returns $r(m, i)$, $i = 1, ..., N$, measured over small time intervals, 10 seconds for example) is defined as:

$$V(m, N) = \sum_{j=1}^{N} r(m, j)^2.$$

What complicates matters is that *(i)* data are expensive and difficult to get (yet), and moreover difficult to 'handle' and *(ii)* underlying the simple computation is a theory for continuous time stochastic processes which motivates and explains the computation. One further thing must be stressed: The computation of $V(\cdot, \cdot)$ is model-free in the sense that the same quantity is computed for various data series independently of underlying econometric models. For each case one tries to make sure that for a rich enough class of underlying (continuous time) models the quantity can be interpreted as (integrated) 'volatility' (sometimes confusingly referred to as realized integrated variance also).

IX.1 A quick introduction to continuous time price processes (and quadratic variation).

Note first that the realized volatility for a given asset is based on the assumption that for any market time (say, within a given trading day) $t \in [0, T]$

there exists a corresponding price, denoted $P(t)$. As usual we will analyze the log-transform of this price, and hence define $p(t) = \log(P(t))$, or simply p_t. Naturally, this continuous time price is only observed whenever there is a transaction in the market, but in liquid markets, as for example the EUR-USD market, transactions are typically only separated by a few (milli-)seconds.

We continue by a quick overview of some key results for continuous time processes. Taylor (2005, chapter 13) provides a more wide-ranging survey, while for example Karatzas and Shreve (1991) gives a thorough treatment of continuous time processes.

Example IX.1 *(The Brownian motion) A key ingredient within continuous time processes is the* ***Brownian motion****. A stochastic process* $(W(t) : t \geq 0)$ *is called a Brownian motion if:*

1. $W(0) = 0$.
2. *W has independent increments, i.e. if* $0 \leq r < s \leq t < u$, *then*

$$W(u) - W(t) \quad \textit{and} \quad W(s) - W(r)$$

   are independent.

3. *The increments are normally distributed, i.e.*

$$W(t) - W(s) \overset{d}{=} N(0, t - s)$$

   for all $0 \leq s \leq t$.

4. *W has continuous trajectories, i.e. it is continuous in t.*

Example IX.2 *A basic continuous time stochastic differential equation (SDE) is given by*

$$dp(t) = \sigma dW(t),$$

where $W(t)$ is a Brownian motion on the time interval $t \in [0, T]$, and σ is constant. This differential equation is identical to, or has solution,

$$p(t) = p(0) + \sigma W(t).$$

Thus, at any time point $p(t)$ is $N(p(0), \sigma^2 t)$ distributed. Moreover, by the definition of a Brownian motion, it has independent and Gaussian distributed increments,

$$p(t) - p(s) = \sigma(W(t) - W(s)) \overset{d}{=} N\left(0, \sigma^2(t - s)\right),$$

where $t > s \geq 0$. *In particular,*

$$\text{Var}\left(p\left(t\right)\right) = \text{Var}\left(p\left(t\right) - p\left(0\right)\right) = \sigma^2 t = \int_0^t \sigma^2 du \qquad \textit{and}$$

$$\text{Var}\left(p\left(t\right) - p\left(s\right)\right) = \sigma^2\left(t - s\right) = \int_s^t \sigma^2 du.$$

In fact, for any $t > s > u$, *we have for example,*

$$\left(p\left(t\right) - p\left(s\right), p\left(s\right) - p\left(u\right)\right)' \overset{d}{=} N\left(0, \Omega_t\right), \qquad \textit{where}$$

$$\Omega_t = \sigma^2 \begin{pmatrix} t - s & 0 \\ 0 & s - u \end{pmatrix}.$$

That is, $p\left(t\right)$ *is a (continuous time) random walk or a Brownian motion initiated in* $p\left(0\right)$.

The above example of a stochastic process $X\left(t\right)$ with $t \in [0, T]$ is obviously not a realistic model for log-returns and many extensions of this exist of which we provide some few key examples (more can be found in Taylor, 2005).

The variance computation for $X\left(t\right) = p\left(t\right)$ above is closely related to the concept of quadratic variation of a stochastic process usually denoted $[X](t)$ (and hence also to the concept of realized volatility as we shall see below). With $X(t)$ univariate this can be defined as the following (in probability) limit: For some fixed $t > 0$, let $\Pi_N = \{t_0, t_1, ..., t_k, .., t_N\}$ be a partition of the interval $[0, t]$ with $0 = t_0 \leq t_1 \leq ... \leq t_k \leq ... \leq t_N = t$. A measure of how fine this partition is, is given by the *mesh* $||\Pi_N|| = \max_{k=1,2,...,N} |t_k - t_{k-1}|$ and

$$[X]\left(t\right) = \underset{||\Pi_N|| \to 0}{\text{plim}} \sum_{k=1}^{N} \left[X(t_k) - X(t_{k-1})\right]^2. \tag{IX.1}$$

Likewise, the d-th power variation is defined as above, but with $\left|X(t_k) - X(t_{k-1})\right|^d$ replacing $\left[X(t_k) - X(t_{k-1})\right]^2$.

Working with continuous time processes can be difficult, and one of the difficulties is the concept of differentiation with respect to time t as for example the Brownian motion $W\left(t\right)$, while continuous, it is nowhere differentiable, while $W\left(t\right)^2$ is. A key tool working with these processes is the Ito-lemma (or Ito's rule). For a quite general class of univariate stochastic processes $X(t)$ defined by some stochastic differential equation (SDE; see Karatzas and Shreve, 1991), the Ito rule explains how functions of X_t evolve, such as for example $f\left(X_t\right) = \log X(t)$ (with $X(t) > 0$) and $f\left(X_t\right) = X_t^2$.

More precisely, with $f(t, X(t))$ some function of time t and of $X(t)$ (which is twice differentiable in $X(t)$), the famous Ito's rule states that,

$$df(t, X(t)) = \dot{f}_t(t, X(t))\, dt + \dot{f}_x(t, X(t))\, dX_t + \frac{1}{2}\ddot{f}_{xx}(t, X(t))\, d[X](t), \quad \text{(IX.2)}$$

where

$$\dot{f}_t(t, X(t)) = \tfrac{\partial}{\partial t} f(t, X(t)), \qquad \dot{f}_x(t, X(t)) = \tfrac{\partial}{\partial X} f(t, X(t)),$$
$$\ddot{f}_{xx}(t, X(t)) = \tfrac{\partial^2}{\partial X^2} f(t, X(t))$$

Example IX.3 *In terms of the previous example of the* ***Brownian motion****, it holds that the quadratic variation is* $[p](t) = \sigma^2[W](t)$*, where*

$$[W](t) = t = \int_0^t ds.$$

Moreover, with $f(t, W(t)) = W^2(t)$ *we have* $X(t) = W(t)$*,* $\dot{f}_t = 0$*,* $\dot{f}_x = 2W_t$ *and* $\ddot{f}_{xx} = 2$*, such that*

$$dW^2(t) = 2W(t)dW(t) + dt.$$

Thus we see that while indeed $W^2(t)$ *is differentiable,* $dW^2(t)/dt \neq 2W(t)$ *as one might expect. Another implication is that* $W^2(t) = 2\int_0^t W(s)dW(s) + t$*, or the stochastic integral known from e.g. unit-root analysis in time series is given by,*

$$\int_0^t W(s)dW(s) = \frac{1}{2}\left(W^2(t) - t\right).$$

Example IX.4 *A large classic class of continuous time stochastic processes is given by the* ***Ito processes****, with* $\mu(t)$ *and* $\sigma(t)$ *(stochastic) functions of time* t*,*

$$dX(t) = \mu(t)\, dt + \sigma(t)dW(t),$$

or, equivalently,

$$X(t) = X(0) + \int_0^t \mu(s)\, ds + \int_0^t \sigma(s)dW(s).$$

Here $\mu(t)$ *and* $\sigma(t)$ *are assumed to be adapted processes, i.e.* $\mu(t)$ *and* $\sigma(t)$ *are known at time* t*.*
It is beyond the scope of this note to discuss all aspects of the class of Ito processes. Importantly, the quadratic variation is

$$[X](s) = \int_0^t \sigma^2(s)ds.$$

Moreover, we emphasize the following important properties about $\int_0^t \sigma(s)dW(s)$*: It holds that*

$$E\left[\int_0^t \sigma(s)dW(s)\right] = 0$$

and

$$E\left[\left(\int_0^t \sigma(s)dW(s)\right)^2\right] = \int_0^t E[\sigma^2(s)]ds.$$

Lastly, if $\sigma(t)$ *is a deterministic function of time,*

$$\int_0^t \sigma(s)dW(s) \stackrel{d}{=} N\left(0, \int_0^t \sigma^2(s)ds\right).$$

Example IX.5 *The* ***Geometric Brownian Motion*** $X(t)$ *known from the Black-Scholes formula for option prices is given as the solution to,*

$$dX(t) = \mu X(t)dt + \sigma X(t)dW(t), \qquad \textit{with } [X](t) = \int_0^t \sigma^2 X^2(s)ds.$$

Applying Ito's rule to $f(t, X(t)) = \log X(t)$*, gives*

$$d\log X(t) = \frac{1}{X(t)}dX(t) - \frac{1}{2X(t)^2}d[X](t) = \mu dt + \sigma dW(t) - \frac{1}{2}\sigma^2 dt,$$

that is,

$$d\log X(t) = \left(\mu - \frac{1}{2}\sigma^2\right)dt + \sigma dW(t). \qquad \textit{with } [\log X](t) = \int_0^t \sigma^2 ds = \sigma^2 t.$$

Note that this motivates why most of the recent literature on realized volatility often starts with the assumption that log*-prices,* $\log X(t)$*, have a general formulation of the form,*

$$d\log X(t) = \tilde{\mu}(t)dt + \tilde{\sigma}(t)dW_t,$$

with time-varying drift term $\tilde{\mu}(t)$ *and volatility* $\tilde{\sigma}(t)$*. In fact, most often* $\tilde{\sigma}(t)$ *is stochastic as in the discrete time SV models.*

We can write the solution for the Geometric Brownian motion as,

$$\log X(t) = \log X(0) + \left(\mu - \frac{1}{2}\sigma^2\right)t + \sigma W(t),$$

from which we can find the explicit solution,

$$X(t) = X(0)\exp\left\{\left(\mu - \frac{1}{2}\sigma^2\right)t + \sigma W(t)\right\}.$$

Hence for $X(0)$ fixed, $X(t)$ is log-normal distributed with,

$$E[X(t)] = X(0)\exp(\mu t) \qquad \text{and} \quad V[X(t)] = X(0)^2\exp(2\mu t)\left(\exp\left(\sigma^2 t\right) - 1\right)$$

Moreover, with $s < t$,

$$\log X(t) = \log X(s) + \left(\mu - \frac{1}{2}\sigma^2\right)(t-s) + \sigma\left(W(t) - X(s)\right).$$

Hence, the increments $\log X(t) - \log X(s)$ are Gaussian, and $X(t)$ conditional on $X(s)$ is log-normal.

Example IX.6 *The **Ornstein-Uhlenbeck** (OU) process is the continuous time equivalent of the discrete time AR(1) process. In its simplest form it is the solution to,*

$$dX(t) = -\kappa X(t)dt + \sigma dW(t).$$

To solve this, apply Ito's rule to $f(t, X(t)) = \exp(\kappa t)X(t)$, to get

$$\begin{aligned} d\left(\exp(\kappa t)X(t)\right) &= \kappa\exp(\kappa t)X(t)dt + \exp(\kappa t)\sigma dX(t) \\ &= \exp(\kappa t)\sigma dW(t) \end{aligned}$$

That is,

$$\exp(\kappa t)X(t) = X(0) + \int_0^t \exp(\kappa s)\sigma dW(s)$$

or

$$X(t) = \exp(-\kappa t)X(0) + \int_0^t \exp\left(-\kappa(t-s)\right)\sigma dW(s)$$

and moreover,

$$[X](t) = \int_0^t \exp\left(-2\kappa(t-s)\right)\sigma^2 ds = \frac{\sigma^2}{2\kappa}\left(1 - \exp(-2\kappa t)\right).$$

We also see that,

$$X(t) = \exp\left(-\kappa(t-s)\right)X(s) + \int_s^t \exp\left(-\kappa(t-u)\right)\sigma dW(s),$$

hence with observing at discrete time points with $t - s = 1$, the OU is an AR(1) process

$$X(t) = \phi X(t-1) + \varepsilon(t),$$

with autoregressive parameter $\phi = \exp(-\kappa)$ and i.i.d. Gaussian innovations $\varepsilon(t)$. The innovations $\varepsilon(t)$ are i.i.d. $N(0, \omega)$ with variance $\omega = \frac{1}{2\kappa}\left(1 - \exp(-2\kappa)\right)$.

IX.2 Realized Volatility

Within a given trading day $m \in \{1, ..., M\}$ let $[m-1, m]$ denote the time interval over which a financial asset with price $p(\cdot)$ is traded. We focus on the case where the price of the asset is observed at $N+1$ equally-spaced discrete points in time, $\tau = t_0, t_1, ..., t_N$, with $t_i = m - 1 + i/N$, $i = 0, 1, ..., N$. Then we may define the ith log-return within trading day m as

$$\begin{aligned} r(m,i) &= p(t_i) - p(t_{i-1}), \quad i = 1, ..., N, \\ &= p(m-1+i/N) - p(m-1+(i-1)/N) \end{aligned}$$

Following McAleer and Medeiros (2008), we define the realized volatility is defined as follows.

Definition IX.1 *The realized volatility for a given day $m \in \{1, ..., M\}$ based on N equally-spaced log-returns is defined as*

$$V(m,N) = \sum_{j=1}^{N} r(m,j)^2.$$

Using the definition of return $r(m,j)$, the realized volatility can be rewritten as

$$\begin{aligned} V(m,N) &= \sum_{j=1}^{N} r(m,j)^2 \\ &= r(m,1)^2 + \cdots + r(m,N)^2 \\ &= [p(t_1) - p(t_0)]^2 + \cdots + [p(t_N) - p(t_{N-1})]^2 \\ &= [p(m-1+1/N) - p(m-1)]^2 + \cdots + [p(m) - p(m-1+(N-1)/N)]^2 \end{aligned}$$

Example IX.7 *The New York Stock Exchange (NYSE) is open Monday - Friday from 9:30 to 16:00 ET except on public holidays. Hence each trading day on the NYSE has 6.5 hours of trading. The table below states half hourly prices for the index from 10/2-09.*

Time	*log-price*	*price*
09:30	*6.7684*	*869.89*
10:00	*6.7506*	*854.60*
10:30	*6.7536*	*857.14*
11:00	*6.7549*	*858.22*
11:30	*6.7505*	*854.48*
12:00	*6.7396*	*845.25*
12:30	*6.7427*	*847.88*
13:00	*6.7342*	*840.71*
13:30	*6.7409*	*846.31*
14:00	*6.7325*	*839.25*
14:30	*6.7333*	*839.89*
15:00	*6.7445*	*849.39*
15:30	*6.7268*	*834.45*
16:00	*6.7180*	*827.16*

This corresponds to $N = 13$*. Hence the realized volatility based on half hourly returns for the 10/2-09 can be computed as*

$$V(\,''10/2\text{-}09'', 13) = (6.7506 - 6.7684)^2 + \cdots + (6.7180 - 6.7268)^2 = 0.0011798$$

If one instead wanted the realized volatility based on hourly returns (counting from 9:30, corresponding to $N = 6$*) it would be computed as*

$$\begin{aligned} V(\,''10/2\text{-}09'', 6) = (6.7536 - 6.7684)^2 + (6.7505 - 6.7536)^2 \\ + \cdots + (6.7268 - 6.7333)^2 = 0.00039274 \end{aligned}$$

Remark IX.1 *The computation of* $V(m, N)$ *is based on the log-price observed at the points* $t_i = m - 1 + i/N$*,* $i = 0, 1, ..., N$*. Here the point* $t_N = m$ *should be understood as the last point where the price is observed during trading day* m*, and not the first observation at trading day* $m + 1$*. These two prices may differ quite a lot due to so-called overnight effects. Likewise, when computing* $V(m+1, N)$ *the point* $t_0 = m$ *corresponds to the first point where the price is observed during trading day* $m + 1$*. In order to deal with this issue, as clarified in the section below on market time, one could consider* $[m - 1, m[$ *as the time interval of trading at trading day* m*. This would lead to introducing some additional notation and conventions which we do not seek to deal with in this note.*

IX.2.1 Quadratic Variation and Realized Volatility

The realized volatility is computed as

$$V(m,N)=\sum_{j=1}^{N}r(m,j)^2=\sum_{j=1}^{N}\left(p(t_j)-p(t_{j-1})\right)^2.$$

A main result is that for a quite general class of SDE's for $X(t)=\log p\,(t)$, on the form

$$dX(t)=d\log p(t)=\mu dt+\sigma(t)dW(t),\qquad [X](t)=\int_0^t\sigma^2(s)ds,$$

(with σ_t possibly stochastic as in the much applied SV models), we have ,

$$V(m,N)\xrightarrow{p}\int_{m-1}^{m}\sigma(s)^2ds=[X](m)-[X](m-1)$$

as $N\to\infty$, i.e. as we increase the sample frequency (the number of sampling points per trading day).

That is, the realized volatility $V\,(m,N)$ converges in probability as $N\to\infty$ to the intra-day quadratic variation, which again is the integrated variance[1] *of p_t.*

To give an understanding why this is the case we demonstrate it for two specific examples.

Example IX.8 *Assume $dp(t)=\sigma dW(t)$ such that $[p](t)=\sigma^2t$, and*

$$p(t)-p(s)=\sigma\left(W(t)-W(s)\right)\overset{d}{=}N\left(0,\sigma^2\left(t-s\right)\right).$$

Hence with $t=m-1+j/N$ and $s=m-1+(j-1)/N$, $t-s=1/N$ and therefore,

$$\begin{aligned}V(m,N)&=\sum_{j=1}^{N}r(m,j)^2=\sum_{j=1}^{N}\left(p(m-1+j/N)-p(m-1+(j-1)/N)\right)^2\\&=\sigma^2\frac{1}{N}\sum_{j=1}^{N}z_j^2,\qquad z_j\ \sim i.i.d.N\left(0,1\right).\end{aligned}$$

[1] Therefore sometimes the term "realized volatility" can be misleading as it is really the "realized integrated volatility".

Now, as $N \to \infty$, the LLN for i.i.d. processes gives directly,

$$V(m,N) = \sigma^2 \frac{1}{N} \sum_{j=1}^{N} z_j^2 \xrightarrow{p} \sigma^2 E z_j^2 = \sigma^2 = [p](m) - [p](m-1).$$

That is, the realized volatility $V(m,N)$ converges in probability to the intra-day integrated volatility, σ^2.

Example IX.9 *Assume next that $dp(t) = \sigma dW(t)$ where $\{\sigma(t)\}_{t\in[0,T]}$ is independent of $\{W(t)\}_{t\in[0,T]}$. Conditional on $\{\sigma(t)\}_{t\in[0,T]}$ we have*

$$[p](t) | \{\sigma(t)\}_{t\in[0,T]} = \int_0^t \sigma^2(u) du,$$

and

$$p(t) - p(s) | \{\sigma(t)\}_{t\in[0,T]} = \left| \int_s^t \sigma(u) dW_u \right| \{\sigma(t)\}_{t\in[0,T]} \overset{d}{=} N\left(0, \int_s^t \sigma^2(u) du\right).$$

We therefore have, conditional on $\{\sigma(t)\}_{t\in[0,T]}$ and with $z_j \sim i.i.d. N(0,1)$,

$$\begin{aligned} V(m,N) &= \sum_{j=1}^{N} r(m,j)^2 = \sum_{j=1}^{N} \left(p(m-1+j/N) - p(m-1+(j-1)/N)\right)^2 \\ &= \sum_{j=1}^{N} \left(\int_{m-1+(j-1)/N}^{m-1+j/N} \sigma^2(u) du \right) z_j^2 \\ &= \sum_{j=1}^{N} \left(\int_{m-1+(j-1)/N}^{m-1+j/N} \sigma^2(u) du \right) + \sum_{j=1}^{N} \left(\int_{m-1+(j-1)/N}^{m-1+j/N} \sigma^2(u) du \right) \left(z_j^2 - 1\right) \\ &= \int_{m-1}^{m} \sigma^2(u) du + \sum_{j=1}^{N} \left(\int_{m-1+(j-1)/N}^{m-1+j/N} \sigma^2(u) du \right) \left(z_j^2 - 1\right) \end{aligned}$$

For the second term, we can use Chebychev's inequality[2], to see that (condi-

[2]For any constant $\eta > 0$, $P(|X| > \eta) \leq E[X^2]/\eta^2$.

tional on $\{\sigma(t)\}_{t\in[0,T]}$*)*

$$
\begin{aligned}
&P\left(\left|\sum_{j=1}^{N}\left(\int_{m-1+(j-1)/N}^{m-1+j/N}\sigma^2(u)du\right)\left(z_j^2-1\right)\right|>\eta\right)\\
&\leq \frac{E\left[\left|\sum_{j=1}^{N}\left(\int_{m-1+(j-1)/N}^{m-1+j/N}\sigma^2(u)du\right)\left(z_j^2-1\right)\right|^2\right]}{\eta^2}\\
&= \frac{E\left[\sum_{j=1}^{N}\left(\int_{m-1+(j-1)/N}^{m-1+j/N}\sigma^2(u)du\right)^2\left(z_j^2-1\right)^2\right]}{\eta^2}\\
&= \frac{2}{\eta^2}\sum_{j=1}^{N}\left(\int_{m-1+(j-1)/N}^{m-1+j/N}\sigma^2(u)du\right)^2
\end{aligned}
$$

since the $z_j's$ *are independent,* $E[z_j^2-1]=0$*, and* $E[(z_j^2-1)^2]=2$*. Next, assuming that* $\max_{u\in[0,T]}\sigma^2(u)<\infty$*,*

$$
\int_{m-1+(j-1)/N}^{m-1+j/N}\sigma^2(u)du\leq\left(\max_u\sigma^2(u)\right)\frac{1}{N},
$$

and hence

$$
\begin{aligned}
\sum_{j=1}^{N}\left(\int_{m-1+(j-1)/N}^{m-1+j/N}\sigma^2(u)du\right)^2 &\leq \left(\max_u\sigma^2(u)\right)^2\frac{1}{N^2}\sum_{j=1}^{N}\\
&= \left(\max_u\sigma^2(u)\right)^2\frac{1}{N}\to 0
\end{aligned}
$$

as $N\to\infty$*. Hence,*

$$
P\left(\left|\sum_{j=1}^{N}\left(\int_{m-1+(j-1)/N}^{m-1+j/N}\sigma^2(u)du\right)\left(z_j^2-1\right)\right|>\eta\right)\to 0,
$$

meaning that the second term tends to zero in probability. We conclude that

$$
V(m,N)\xrightarrow{p}\int_{m-1}^{m}\sigma^2(u)du=[p](m)-[p](m-1)\qquad \text{as } N\to\infty.
$$

Example IX.10 *(Market Microstructure Noise) The previous examples show that the RV is a consistent estimator for QV as the sampling frequency* $N\to\infty$*. However, due to so-called market microstructure noise (MMN), the result is hardly applicable in practice; see e.g. Andersen and Benzoni*

(2009, Section 6) and the references therein for a discussion of potential sources of MMN. To fix ideas, suppose, as in Example IX.8, that the true (efficient) returns are given by

$$r^*(m,j) = \left(p(m-1+j/N) - p(m-1+(j-1)/N)\right), \quad j = 1,\ldots,N,$$

with $dp_t = \sigma dW_t$. Due to MMN, we observe $r^*(m,j)$ subject to noise, that is, we observe

$$r(m,j) = r^*(m,j) + \varepsilon_j,$$

where $\{\varepsilon_j\}_{j=1}^N$ is an i.i.d. process with $E[\varepsilon_t] = 0$, $\mathrm{Var}[\varepsilon_t] = \sigma_\varepsilon^2$ and $E[\varepsilon_t^4] < \infty$, and assume that the processes $\{\varepsilon_j\}_{j=1}^N$ and $\{p(t)\}_{t\in[m-1,m]}$ are independent. Note that the moments of ε_j do not depend on the frequency N. It holds that, with $V(m,N) = \sum_{j=1}^N r(m,j)^2$,

$$E\left[V(m,N)|\{p(t)\}_{t\in[m-1,m]}\right] = \sum_{j=1}^N r^*(m,j)^2 + N\sigma_\varepsilon^2$$

and

$$\mathrm{Var}\left[V(m,N)|\{p(t)\}_{t\in[m-1,m]}\right] = 4\sigma_\varepsilon^2 \sum_{j=1}^N r^*(m,j)^2 + N\,\mathrm{Var}\left(\varepsilon_t^2\right).$$

This suggests that the observed RV, $V(m,N)$, is an unreliable estimator for the RV of the efficient return variation, $\sum_{j=1}^N r^*(m,j)^2$. In particular, when the data are sample at very high-frequency, that is, when N is large, $V(m,N)$ may have little to do with the efficient returns. See, e.g., Zhang et al. (2005) for a rigorous discussion, and Hansen and Lunde (2006) for a setting where $r^*(m,j)$ and ε_j are dependent.

Example IX.11 *(Estimation of the conditional mean) Suppose that*

$$dp(t) = \mu dt + \sigma dW(t),$$

or equivalently,

$$p(t) = p(0) + \mu t + \sigma W(t),$$

with some initial fixed value $p(0)$ and $\mu \in \mathbb{R}$ and $\sigma > 0$. Suppose that we have N returns for a period $[0,T]$, say, one day, given by

$$r(i) = p(t_i) - p(t_{i-1}), \quad \text{with } t_i = \frac{i}{N}T, \quad i = 1,\ldots,N.$$

Hence,

$$r(i) = \mu\frac{T}{N} + \sigma(W(t_i) - W(t_{i-1})),$$

and, using the properties of a Brownian motion, we have that $\{r(i)\}_{i=1,\ldots,N}$ *is an i.i.d. process with*

$$r(i) \stackrel{d}{=} N\left(\mu\frac{T}{N}, \sigma^2\frac{T}{N}\right).$$

For estimating μ *and* σ^2*, we have that the log-likelihood function is given by*

$$L_N(\mu, \sigma^2) = -\sum_{i=1}^{N}\log(\sigma^2) + \frac{\left(r(i) - \mu\frac{T}{N}\right)^2}{\sigma^2\frac{T}{N}},$$

and it follows that the maximum likelihood estimator for μ *is given by*

$$\hat{\mu} = \frac{1}{T}\sum_{i=1}^{N} r(i) = \frac{p(T) - p(0)}{T} = \mu + \frac{\sigma W(T)}{T}.$$

Hence, for a fixed time interval, that is, with T *fixed, the estimator for* μ *does not depend on the sampling frequency* N*, and depends only on the time span of the data,* T*. This is in contrast to the realized volatility, where the infill asymptotics* ($N \to \infty$) *are used for showing convergence to the quadratic variation. It holds that*

$$E[\hat{\mu}] = \mu,$$

and

$$V(\hat{\mu}) = \frac{\sigma^2}{T},$$

so that for any $\epsilon > 0$

$$P(|\hat{\mu} - \mu| > \epsilon) \leq \frac{V(\hat{\mu})}{\epsilon^2} = \frac{\sigma^2}{T\epsilon^2}.$$

Hence, the estimator $\hat{\mu}$ *is unbiased but does only converge (in probability) to* μ *as* $T \to \infty$*, that is, as we increase the time span of the data.*

References

Aït-Sahalia, Y. & J. Jacod (2014), *High-Frequency Financial Econometrics*, Princeton University Press.

Andersen, T.G. & L. Benzoni (2009), "Realized Volatility", In: T. G. Andersen, R. A. Davis, J. P. Kreiss, and T. Mikosch (Eds.), *Handbook of Financial Time Series*. Springer.

Andersen, T.G., T. Bollerslev, F. Diebold & Labys (2001), "The Distribution of Realized Exchange Rate Volatility", *Journal of the American Statistical Association*, Vol. 96, p. 42–55.

Andersen, T.G., T. Bollerslev, F. Diebold & Labys (2003), "Modeling and Forecasting Realized Volatility", *Econometrica*, Vol. 71, p. 579–625.

Hansen, P.R & A. Lunde (2006). "Realized Variance and Market Microstructure Noise", *Journal of Business & Economic Statistics*, Vol. 24, p. 127–161.

Karatzas, I. & S.E. Shreve (1995), *Brownian Motion and Stochastic Calculus*, Springer-Verlag

McAleer, M. & M.C. Medeiros (2008), "Realized Volatility: A review", *Econometric Reviews*, Vol. 27, p. 10–45.

Taylor, S.J. (2005), *Asset Price Dynamics, Volatility and Prediction*, Princeton University Press.

Zhang, L., P.A. Mykland, Y. Aït-Sahalia (2005), "A Tale of Two Time Scales: Determining Integrated Volatility With Noisy High-Frequency Data", *Journal of the American Statistical Association*, Vol. 100, p. 1394–1411.

Appendix: Calendar time and market time

Since the *realized volatility* measure is based on intra-day transactions the usual approach of simply denoting the (daily) returns 1 through T no longer suffices. For empirical applications, we typically need a way of timekeeping, which allows a specific time and date of a transactions as well as market specific conditions such as open-close hours and holidays. If we were to use simple calender time (e.g. 13:35 on the 22$^{\text{nd}}$ of February 2007) we would need to separately keep track of whether the market was open or closed at this time. As a solution we will use market time, which is defined below. For a further discussion see also Andersen, Bollerslev, Diebold & Labys (2001).

Definition IX.2 *Market time is defined specific to a given market. For any time point (where the market is open) it is defined as the fraction*

$$t = \frac{\text{"no. of hours of trading since initial date"}}{\text{"no. of hours of trading per day"}}.$$

By convention market time 0 corresponds to "opening of trade" (that is the time of day where the market opens) at the starting date. Hence all integer valued market times ($t = 0, 1, ..., T-1$), correspond to opening time of the market on day $t+1$. When the considered period holds T trading days market time, t, belongs to the interval $[0, T[$.

To fix ideas consider the following two examples.

Example IX.12 *Consider the New York Stock Exchange (NYSE), which is open Monday - Friday from 9:30 to 16:00 ET except on public holidays. Assume we wish to analyze the following period: Friday the 6/2-09 at 9:30 to Tuesday the 10/2-09 at 16:00. What is the market time (t) of a transaction, which occurred on Tuesday at 11:54:30 and of one that occurred at 9:30 on the same day?*

1) Each trading day on the NYSE has 6.5 hours of trading. Hence from Friday the 6/2-09 at 9:30 to Tuesday at 11:54:30 there has been $6.5 + 6.5 + (11 + 54/60 + 30/60^2) - (9 + 30/60) = 15.40833$ *hours of trading. The market time of Tuesday at 11:54:30 is therefore* $t = 15.40833/6.5 = 2.370512$.

2) By convention 9:30 on Tuesday, which is the opening time of the NYSE on the third day of our period, has the market time $t = 2$.

Example IX.13 *Next consider the EUR-USD foreign exchange (FX) market, which trades 24 hours a day, 7 days a week. However, trading volume is much reduced during the weekends and we will therefore assume that the*

EUR-USD market is open from Monday at 0:00:00 to Friday at 23:59:59 GMT. Again we wish to analyze the period: Friday the 6/2-09 at 0:00:00 to Tuesday the 10/2-09 at 23:59:59. What is the market time (t) of a transaction, which occurred on Tuesday at 11:54:30?

Each trading day on the EUR-USD market has 24 hours of trading. Hence from Friday the 13/2-09 at 0:00 to Tuesday at 11:54:30 there has been $24 + 24 + (11 + 54/60 + 30/60^2) - (0 + 0/60) = 59.90833$ *hours of trading. The market time of Tuesday at 11:54:30 is therefore* $t = 59.90833/24 = 2.496180$.

From the examples and definition it follows that all transactions from trading day no. m, $m \in \{1, ..., T\}$ have market time in the interval $[m-1, m[$. Note that the point m is excluded from the interval, as this point corresponds to the beginning of the next trading day. This insight is important when manipulating databases of high frequency transactions