Anders Rahbek
Rasmus Søndergaard Pedersen
University of Copenhagen

August 2024

Part I
Introduction to Financial Time Series

In this chapter we introduce concepts needed for the probability analysis involved in the econometric analysis of financial time series as in Tsay (2010).

I.1 Time Series

Two of the most classic time series processes in financial econometrics are the autoregressive (AR) process, and the autoregressive conditional heteroskedastic (ARCH) process. These are briefly introduced in the next, before making concepts and ideas precise in the following sections.

I.1.1 AR process

The simplest AR process is of order one (AR(1)), with the AR(1) process given by

$$x_t = \rho x_{t-1} + \varepsilon_t, \tag{I.1}$$

for $t = 1, 2, \ldots$ and with the recursion initiated in $x_0 = x \in \mathbb{R}$. The autoregressive parameter $\rho \in \mathbb{R}$, while the innovations ε_t are independently and identically distributed (i.i.d.), with a normal distribution with mean zero and variance $\sigma^2 > 0$, i.e. ε_t are i.i.d. $\mathrm{N}(0, \sigma^2)$. It follows that $\mathbb{E}[x_t | x_{t-1}] = \rho x_{t-1}$, while $\mathbb{V}(x_t | x_{t-1}) = \sigma^2$, and therefore the time-dependence, or dynamics, is modelled through the conditional mean of x_t given the past (observations).

Clearly the dynamic properties of the AR(1) process depend on the value of ρ. Note in this respect that simple recursion in (I.1) gives

$$x_t = \rho^t x + \sum_{i=0}^{t-1} \rho^i \varepsilon_{t-i}. \tag{I.2}$$

In particular, x_t is Gaussian distributed with unconditional mean $\mu_t = \rho^t x$ and variance

$$v_t = \left(1 + \rho^2 + \rho^4 + ... + \rho^{2(t-1)}\right)\sigma^2. \tag{I.3}$$

Both the mean and variance depends on t, and hence x_t's distribution is varying with t. At the same time, if $|\rho| < 1$, $\mu_t \to 0$, and using the well-known result for power series in Lemma I.3.1 below, we find

$$v_t = \frac{1-\rho^{2t}}{1-\rho^2}\sigma^2 \to \sigma^2 / \left(1-\rho^2\right).$$

Therefore, for $|\rho| < 1$, as $t \to \infty$, x_t will resemble the so-called *linear process*,

$$x_t^* = \sum_{i=0}^{\infty} \rho^i \varepsilon_{t-i}, \tag{I.4}$$

in terms of the sequence $\{\varepsilon_t\}_{t=...-1,0,1,2....}$ of i.i.d. $\mathrm{N}(0, \sigma^2)$ variables. The process $\{x_t^*\}_{t=0,1,2,...}$is Gaussian distributed with $\mathbb{E}[x_t^*] = 0$ and $\mathbb{V}[x_t^*] = \sigma^2 / (1-\rho^2)$. The distribution of x_t^* does not depend on t, and is an example of a *stationary* process as detailed below. Thus if $|\rho| < 1$, x_t is *asymptotically stationary* in the sense that it resembles the stationary process x_t^* for large t. Note that x_t^* solves the recursion in (I.1) as can be seen by simple substitution. Also note that by giving x_0 the initial distribution as given by,

$$x_0 = \sum_{i=0}^{\infty} \rho^i \varepsilon_{0-i}, \tag{I.5}$$

x_t in (I.2) has the same distribution as the stationary solution x_t^*. That is, there exists a stationary solution to the AR(1) recursion, provided $|\rho| < 1$.

Remark I.1.1 *That the representation of x_t^* in (I.4) in terms of the infinite sum is well-defined follows by classic probability results, see e.g. Johansen (1996). While the actual form is appealing, it is not cruical to our theory; rather our focus is on showing that a stationary solution (here to the AR(1) recursion) exists, and that it has certain properties.*

I.1.2 ARCH process

The simplest ARCH process is the ARCH(1) which for $t = 1, 2, \ldots$ is given by

$$x_t = \sigma_t z_t \tag{I.6}$$
$$\sigma_t^2 = \omega + \alpha x_{t-1}^2 \tag{I.7}$$

with initial value $x_0 = x$ and where the innovations z_t are i.i.d. $\mathrm{N}(0, 1)$. Moreover, conditionally on x_{t-1}, x_t is Gaussian distributed with mean zero and (conditional) variance σ_t^2. However, *unlike for the AR(1) process, this does not imply that x_t is Gaussian distributed unconditionally, or marginally.* Instead the marginal 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 Gaussian distributed. The level parameter ω is strictly positive, $\omega > 0$, while the ARCH parameter $\alpha \geq 0$. Note that if $\alpha = 0$ then x_t is simply an i.i.d. N(0,ω) sequence (conditionally and unconditionally). It should be emphasized that the conditional variance σ_t^2 is non-constant and stochastic.

The probabilistic behavior of the ARCH process x_t is complicated as for example the concept *stationarity* and *existence of moments* of x_t demands rather technical analysis. 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 simple restrictions on the parameters (ω, α) for which x_t is well-behaved process in the sense that it is stationary, as well as having other desirable properties.

In line with the recursion for the AR(1), observe that the squared x_t satisfies a simple recursion,

$$\begin{aligned} x_t^2 &= \left(\omega + \alpha x_{t-1}^2\right) z_t^2 = \omega z_t + \alpha \left(\omega + \alpha x_{t-2}^2\right) z_t^2 z_{t-1}^2 \\ &= \omega \sum_{i=0}^{t-1} \alpha^i \prod_{j=0}^{i} z_{t-j}^2 + \omega \alpha^t x^2 \prod_{j=0}^{t-1} z_{t-j}^2. \end{aligned}$$

Hence, if $0 \leq \alpha < 1$, as $t \to \infty$, x_t^2 resembles (as $\alpha^t \to 0$, as $t \to \infty$) the stationary process,

$$(x_t^*)^2 = \omega \sum_{i=0}^{\infty} \alpha^i \prod_{j=0}^{i} z_{t-j}^2. \tag{I.8}$$

However, as we will demonstrate, while $\alpha < 1$ is sufficient for the existence of a stationary solution, it is not a necessary condition. That is, x_t indeed has a stationary solution for a range of values of $\alpha \geq 1$.

I.2 Conditional and unconditional moments

We introduce here some notation and concepts needed when discussing conditional and unconditional moments of the ARCH and AR processes. In particular, conditional expectations play a key role in these considerations and will be defined in terms of densities.

I.2.1 Expectations

Consider two random variables $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ for some $p, q \geq 1$, with well-defined densities $f(x)$ and $f(y)$, joint density $f(x,y)$ and, finally, well-defined conditional density $f(x|y) = f(x,y)/f(y)$. Recall that the expectation of X is given by

$$\mathbb{E}[X] = \int_{\mathbb{R}^p} x f(x)\, dx. \tag{I.9}$$

Likewise, the conditional expectation of X given $Y = y$, is given by

$$\mathbb{E}[X|Y = y] = \int_{\mathbb{R}^p} x f(x|y)\, dx, \tag{I.10}$$

which, by definition, is non-stochastic and depends on the value y. With $g(y) = \mathbb{E}[X|Y = y]$, we define furthermore the random variable, $\mathbb{E}[X|Y]$ as

$$E[X|Y] = g(Y). \tag{I.11}$$

Example I.2.1 *Consider $(X, Y)'$ bivariate $\mathrm{N}(\mu, \Omega)$ distributed, with mean μ and covariance matrix Ω given by,*

$$\mathbb{E}\left[\begin{pmatrix} X \\ Y \end{pmatrix}\right] = \mu = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \ \mathbb{V}\left[\begin{pmatrix} X \\ Y \end{pmatrix}\right] = \Omega = \begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{yx} & \sigma_y^2 \end{pmatrix}.$$

In particular, X is $\mathrm{N}(\mu_x, \sigma_x^2)$ and Y is $\mathrm{N}(\mu_y, \sigma_y^2)$ distributed, while $\mathrm{Cov}[X, Y] = \sigma_{xy}$. Furthermore we have the important result that the conditional expectation of X given $Y = y$, is given by

$$\mathbb{E}[X|Y = y] = \mu_x + \omega(y - \mu_y) = \mu_{x|y}, \quad \textit{where } \omega = \sigma_{xy}/\sigma_y^2. \tag{I.12}$$

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

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

In particular, the conditional density is given by

$$f(x|y) = \frac{1}{\sqrt{2\pi\sigma_{x|y}^2}} \exp\left(-\frac{1}{2\sigma_{x|y}^2}\left(x - \mu_{x|y}\right)^2\right).$$

Moreover, $\mathbb{E}[X|Y] = \mu_x + \omega(Y - \mu_y)$ with

$$\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[\mu_x + \omega(Y - \mu_y)] = \mu_x + \omega(\mathbb{E}[Y] - \mu_y) = \mu_x = \mathbb{E}[X]. \tag{I.13}$$

The fact that $\mathbb{E}\left[\mathbb{E}\left[X|Y\right]\right] = \mathbb{E}\left[X\right]$ in the above example is a general feature of the conditional expectation, often referred to as the law of iterated expectations. We list here some well-known properties of conditional expectations which are simple to verify under the assumed setting of densities. A proof can be found in most probability theory books, see e.g. Durrett (2019, Ch.4.1.2).

Lemma I.2.1 *Consider the random variables X, Y and Z with joint density $f\left(x, y, z\right)$ and finite expectation. For the conditional expectation $\mathbb{E}\left[X|Y\right]$ the law of iterated expectations apply,*

$$\mathbb{E}\left[\mathbb{E}\left[X|Y\right]\right] = \mathbb{E}\left[X\right]. \tag{I.14}$$

If X and Y are independent,

$$\mathbb{E}\left[X|Y\right] = \mathbb{E}\left[X\right]. \tag{I.15}$$

Moreover,

$$\mathbb{E}\left[X|Y\right] = \mathbb{E}\left[\left(\mathbb{E}\left[X|Y,Z\right]|Y\right)\right], \quad \mathbb{E}\left[X|X\right] = X. \tag{I.16}$$

With g and h functions such that $g\left(Y\right)$ and $h\left(Y\right)$ take values in $\mathbb{R}$,

$$\mathbb{E}\left[\left(g\left(Y\right) + h\left(Y\right)X\right)|Y\right] = g\left(Y\right) + h\left(Y\right)\mathbb{E}\left[X|Y\right]. \tag{I.17}$$

Example I.2.2 *By definition of the AR process in (I.1)*

$$\mathbb{E}\left[x_t|x_{t-1}\right] = \mathbb{E}\left[\left(\rho x_{t-1} + \varepsilon_t\right)|x_{t-1}\right] = \rho\mathbb{E}\left[x_{t-1}|x_{t-1}\right] + \mathbb{E}\left[\varepsilon_t\right] = \rho x_{t-1},$$

where we have used (I.17), (I.16) and (I.15).

Example I.2.3 *By definition of the ARCH process in (I.6),*

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

Thus, provided $\mathbb{E}\left[|x_t|\right] < \infty$, the ARCH(1) has mean zero, $\mathbb{E}\left[x_t\right] = \mathbb{E}\left[\mathbb{E}\left[x_t|x_{t-1}\right]\right] = 0$.

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

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

such that,

$$\mathbb{E}[X|X \in A] = \int x f(x|A)\, dx.$$

Example I.2.4 *With* x_t $\mathrm{N}(0, \sigma^2)$, *distributed,* $\mathbb{P}(x_t > 0) = \frac{1}{2}$, *and density of the distribution of* x_t *conditional on* $x_t > 0$,

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

with $\mathbb{I}(\cdot)$ *the indicator function. Next,*

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

Here we have used well-known properties of the Gaussian distribution, including symmetry, and $\mathbb{E}|Z| = \sqrt{2/\pi}$, *with* Z $\mathrm{N}(0, 1)$ *distributed (and* $x_t = \sigma Z$*).*

I.2.1.1 Some further comments on conditioning and moments

It is useful to discuss conditioning not only on x_{t-1} as in Example I.2.3 but also on all past variables $\mathcal{F}_{t-1} = (x_{t-1}, x_{t-2} ..., x_0)$ (or, the σ-algebra, denoted $\sigma(x_{t-1}, x_{t-2} ..., x_0)$). This is often referred to as 'conditioning on all past information' in the literature. For both the AR(1) and the ARCH(1) processes, the distribution of x_t conditional on past information up to time $t - 1$ as given by the lagged variables in $\mathcal{F}_{t-1}$ depends only on x_{t-1}. This is commonly referred to as the *Markov property* and it plays an important role for the analysis of stochastic properties of $\{x_t\}$.

Note also at the same time that the simple specifications of the ARCH(1) and AR(1) are for empirical relevance extended in several possible ways, including adding more lags (x_{t-q}, $q \geq 2$) and non-linear functional forms in the conditional mean (AR) and variance (ARCH).

In terms of the definition of conditional expectations, we give meaning to $\mathbb{E}\left[x_t|\mathcal{F}_{t-1}\right]$ by setting $X = x_t$ and $Y = \mathcal{F}_{t-1}$, and often write it as $\mathbb{E}\left[x_t|\mathcal{F}_{t-1}\right]$,with $\mathcal{F}_{t-1} = (x_{t-1}, x_{t-2} \ldots, x_0)$. Thus as in the example for the ARCH(1) process x_t,

$$\mathbb{E}\left[x_t\right] = \mathbb{E}\left[\mathbb{E}\left[x_t|\mathcal{F}_{t-1}\right]\right] = 0.$$

That the calculations are identical, reflects the already mentioned fact that by the definition of the ARCH(1) process, the distribution of x_t conditional on $\mathcal{F}_{t-1} = (x_{t-1}, x_{t-2}, ..., x_0)$ depends only on x_{t-1}. Continuing with the ARCH(1) process , consider the correlation between x_t and x_{t-1},

$$\mathbb{E}\left[x_{t-1}x_t\right] = \mathbb{E}\left[\mathbb{E}\left[x_t x_{t-1}|\mathcal{F}_{t-1}\right]\right] = \mathbb{E}\left[x_{t-1}\mathbb{E}\left[x_t|\mathcal{F}_{t-1}\right]\right] = 0$$

and likewise,

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

Hence the ARCH(1) process is a mean zero and uncorrelated process. Note that this – unlike for uncorrelated Gaussian variables – *does not imply* that x_t and x_{t-1} are independent as e.g. σ_t^2, and hence x_t, depends on x_{t-1}^2.

Now turn to the second order moment, where, as $\mathbb{E}\left[x_t\right] = 0$,

$$\mathbb{V}\left[x_t\right] = \mathbb{E}\left[x_t^2\right] = \mathbb{E}\left[\mathbb{E}\left[x_t^2|\mathcal{F}_{t-1}\right]\right] = \mathbb{E}\left[\sigma_t^2\mathbb{E}\left[z_t^2\right]\right] = \mathbb{E}\left[\sigma_t^2\right] = \sigma^2 + \alpha\mathbb{E}\left[x_{t-1}^2\right] \tag{I.18}$$

Hence, if $\alpha < 1$ and $\mathbb{E}\left[x_t^2\right]$ is assumed constant such that $\mathbb{E}\left[x_t^2\right] = \mathbb{E}\left[x_{t-1}^2\right]$,

$$\mathbb{V}\left[x_t\right] = \mathbb{E}\left[x_t^2\right] = \frac{\sigma^2}{1-\alpha}. \tag{I.19}$$

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

$$\mathbb{E}\left[x_t^{2k+1}\right] = 0$$

i.e. all odd moments are zero. For the 4th order moment it follows that, if $\alpha < 1/\sqrt{3}$,

$$\mathbb{E}\left[x_t^4\right] = 3(\tfrac{1-\alpha^2}{1-3\alpha^2})\left(\mathbb{E}\left[x_t^2\right]\right)^2 > 3\left(\mathbb{E}(x_t^2)\right)^2. \tag{I.20}$$

As $\mathbb{E}\left[x_t^4\right] / \left(\mathbb{E}\left[x_t^2\right]\right)^2 = 3$ for the Gaussian case, we conclude that the ARCH(1) process has *excess kurtosis* since $1 - \alpha^2 > 1 - 3\alpha^2$.

The above reflects that the existence of finite moments of a process x_t in particular for the case of non-linear time series depends on the parameter values. For the linear AR(1) process in (I.1), we find

$$\mathbb{E}\left[x_t\right] = \mathbb{E}\left[\mathbb{E}\left[x_t | \mathcal{F}_{t-1}\right]\right] = \rho \mathbb{E}\left[x_{t-1}\right],$$

hence if $\mathbb{E}\left[x_t\right]$ is constant, $\mathbb{E}\left[x_t\right] = 0$. Likewise,

$$\mathbb{V}\left[x_t\right] = \mathbb{E}\left[x_t^2\right] = \sigma^2 + \rho^2 \mathbb{E}\left[x_{t-1}^2\right] = \sigma^2 / \left(1 - \rho^2\right)$$

provided $\mathbb{E}\left[x_t^2\right]$ is constant and $|\rho| < 1$. In fact, all moments $\mathbb{E}\left[\left|x_t^k\right|\right] < \infty$, $k \geq 1$, with $|\rho| < 1$.

The "constant", or rather, non time-varying, moment assumption used repeatedly above is closely related to the concept of stationarity discussed next.

I.3 Stationarity and dependence

Above we considered moments $\mathbb{E}\left[x_t^k\right]$ for the ARCH(1) and AR(1) process assuming that they were identical for all time points. This leads to the concept of stationarity.

Definition I.3.1 *The process $\{X_t\}_{t=0,1,2,\ldots}$ is said to be stationary, or simply X_t is stationary, if for all t and h with $t, h \geq 0$, 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 $\mathbb{E}\left[X_t\right]$ and variance $\mathbb{V}\left[X_t\right]$ are constant, while the covariance between X_t and X_{t+h}, $\text{Cov}\left(X_t, X_{t+h}\right)$ depends only on h, and not on t.

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

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

with $\Omega_h = \sigma^2 I_{h+1}$ where I_{h+1} is the $(h+1)$-dimensional identity matrix. 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, $\mathbb{E}[X_t]$ is constant and $\mathrm{Cov}[X_t, X_{t+h}] = v(h)$ 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. $\mathrm{N}(0, \sigma^2)$. In particular x_t is a stationary process with $\mathbb{E}[x_t] = 0$, $\mathbb{V}[x_t] = (1 + \theta^2)\sigma^2$, $\mathrm{Cov}(x_t, x_{t+1}) = \theta\sigma^2$ and

$$\mathrm{Cov}[x_t, x_{t+h}] = 0 \quad \textit{for } h > 1.$$

Hence the MA(1) process is stationary as the Gaussian distribution is characterized fully by the first and second order moments.

In the next example we use that for power series:

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

$$\sum_{i=0}^{\infty} \phi^i = 1/(1 - \phi). \tag{I.21}$$

Example I.3.3 *Consider the solution x_t^* to the AR(1) process in (I.4). It follows that $x_t^* = \sum_{i=0}^{\infty} \rho^i \varepsilon_{t-i}, |\rho| < 1$, is Gaussian with $\mathbb{E}[x_t^*] = 0$, using (I.21) with $\phi = \rho^2 < 1$,*

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

Next, as $\mathrm{Cov}\left[\varepsilon_t, \varepsilon_{t+h}\right] = 0$ *for* $h \geq 1$*,*

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

Hence x_t^* *is indeed 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 Markov chain theory below will show that it is in fact not needed to introduce the infinite sums and the explicit stationary solution* x_t^* *to discuss stationarity and dependence structure of the AR process.*

I.3.1 Dependence vanishing over time

The definition of stationarity addresses 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. If the stationary process $\{X_t\}_{t\in\mathbb{Z}}$ is dependent over time, with $\mathbb{E}\left[|X_t|^2\right] < \infty$, a key indicator often used for detection of dependence is the so-called autocorrelation. For a stationary process $X_t \in \mathbb{R}$, the so-called autocovariance function is given by

$$\mathrm{v}(h) = \mathrm{Cov}\left[X_t, X_{t+h}\right], \tag{I.22}$$

and the autocorrelation function (ACF) is defined by,

$$\mathrm{ACF}(h) = \mathrm{Corr}\left[X_t, X_{t+h}\right] = \frac{\mathrm{Cov}\left[X_t, X_{t+h}\right]}{\sqrt{\mathbb{V}\left[X_t\right]\mathbb{V}\left[X_{t+h}\right]}} = \frac{v(h)}{v(0)}, \tag{I.23}$$

where the last equality holds by stationarity. The functions $\rho\left(h\right)$ and $v\left(h\right)$ for various h describe the correlatedness, and hence indicates possible dependence over periods of time.

More generally, various kinds of dependence over time, often referred to as "mixing", or asymptotic independence, is used to describe vanishing dependence between X_t and X_{t+h} as h increases. This idea is crucial for time series and replaces the concept of independence. The idea is that a stationary process $(X_t)_{t=0,1,\ldots}$ is said to be mixing (or, ergodic), if for all t, h and sets

A, B,

$$\begin{aligned} & P((X_0, \dots, X_t) \in A, (X_h, \dots, X_{t+h}) \in B) \to \\ & P((X_0, \dots, X_t) \in A) P((X_0, \dots, X_t) \in B) \quad h \to \infty. \end{aligned} \tag{I.24}$$

Here "mixing" is used loosely and the literature includes many different forms of "mixing" (including α and β-mixing) definitions; but they all share the idea that dependence between X_t and X_{t+h} vanishes as h increases. Or stated differently, events removed far in time from one another are independent. Moreover, they imply various types of law of large numbers (LLNs) apply.

Below we discuss the so-called drift criterion from Markov chain theory which provides a useful tool to establish conditions under which LLNs and central limit theorems (CLTs) hold for time series.

Before turning to the drift criterion, we note the following result which relates mixing and correlations

Lemma I.3.2 *If the stationary process $\{X_t\}_{t=0,1,2,..}$ satisfies (I.24) and has finite variance, then the covariance $v(h) = \text{Cov}[X_t, X_{t+h}]$ tends to zero as $h \to \infty$. On the other hand, if $\{X_t\}$ is a stationary Gaussian process for which $v(h) \to 0$ as $h \to \infty$, then X_t satisfies (I.24).*

Example I.3.4 *It follows by Example I.3.2 that the MA(1) process satisfies (I.24), and likewise, if $|\rho| < 1$, the AR(1) process x_t^* does.*

I.4 Drift criterion from Markov chain theory

If a Markov chain $\{X_t\}_{t=0,1,2,\dots}$ satisfies the drift criterion, a first important implication is that the initial value, X_0, can be given a distribution such that X_t is stationary. This resembles the considerations made for the AR(1) process x_t where as shown one can explicitly choose an initial distribution of x_0 such that $x_t \overset{D}{=} x_t^*$ and hence stationary. Second, the drift criterion implies finiteness of certain moments for the stationary version. Moreover, variants of the law of large numbers (LLN) and the central limit theorem (CLT) apply.

In short, the drift criterion is very helpful in many ways and a powerful tool for time series analysis. The introduction here is based on Meyn and Tweedie (1993) and Tjøstheim (1990).

I.4.0.1 Assumptions

A common key feature of the AR(1) and ARCH(1) processes is that with X_t denoting either of the two, then the distribution of,

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

depends only on X_{t-1}. More precisely, in the AR(1) case x_t conditionally on x_{t-1}, is $\mathrm{N}(\rho x_{t-1}, \sigma^2)$ distributed, while for the ARCH(1) x_t conditionally on x_{t-1}, is $\mathrm{N}(0, \sigma_t^2)$ 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 attractive features.

We make the following assumption:

Assumption I.4.1 *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-n}, for some $n \geq 1$, has a positive (n−step) conditional density $f(y|x) > 0$, which is continuous in both arguments.*

Note that (i) implies that $\{X_t\}_{t=0,1,2,...}$is a Markov chain on $\mathbb{R}^p$, sometimes referred to as a Markov chain on a general state space. Also note that the condition (ii) of continuity is often simple to validate, in particular for $k = 1$.

Example I.4.1 *For the AR(1) process in (I.1), x_t conditional on x_{t-1} has density*

$$f(y|x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\tfrac{(y-\rho)^2}{2\sigma^2}\right),$$

which is positive and continuous in y and x. Often we simply write the conditional density in terms of x_t,

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

Example I.4.2 *For the ARCH process in (I.6), x_t conditional on x_{t-1} has the Gaussian density,*

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

which is positive and continuous.

For the next example recall initially the following well-known result from probability analysis:

Lemma I.4.1 *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 $\mathbb{V}[X] < \infty$, then $\mathbb{E}[Y] = c\mathbb{E}[X]$ and $\mathbb{V}[Y] = c^2\mathbb{V}[X]$.*

Example I.4.3 *In the ARCH process x_t in (I.6) the assumption of z_t i.i.d.$N(0,1)$ is sometimes replaced by the assumption that z_t is i.i.d. with $\mathbb{E}[z_t] = 0$, $\mathbb{V}[z_t] = 1$ and 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.4.1.*

To see this note first that if X is t_v-distributed with $v > 2$, then X has $\mathbb{E}[X] = 0$ and $\mathbb{V}[X] = v/(v-2)$. 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 $\gamma(v) = \Gamma\left(\frac{v+1}{2}\right)/\Gamma\left(\frac{v}{2}\right)$, with $\Gamma(\cdot)$ the so-called Gamma function. As $\mathbb{V}[X] = v/(v-2)$, then using Lemma I.4.1, $z_t = \left(\sqrt{\frac{v-2}{v}}\right) X$ satisfies $\mathbb{V}[z_t] = 1$ and $\mathbb{E}(z_t) = 0$. By simple insertion 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 using Lemma I.4.1 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.4.0.2 Drift function

Next, we define *a drift function* for a process X_t satisfying Assumption I.4.1. With X_t a time series, a drift function for X_t is some function $\delta(X_t)$, 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.4.4 *A much used drift function in the analysis of univariate AR and ARCH processes with $X_t \in \mathbb{R}$, is*

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

while, if $X_t = (X_{1t}, ..., X_{pt})' \in \mathbb{R}^p$,

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

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 of X_t itself. considering the conditional expectation of $\delta(X_t)$ given X_{t-1} or some more "distant" past value of X_t, say X_{t-m}. That is we are interested in studying the dynamics of

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

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

Example I.4.5 *For the AR(1) process x_t with $\delta(x_t) = 1 + x_t^2$,*

$$\begin{aligned}
\mathbb{E}[\delta(x_t)|x_{t-1}] &= \mathbb{E}\left[1 + (\rho x_{t-1} + \varepsilon_t)^2 |x_{t-1}\right] \\
&= 1 + \rho^2 \mathbb{E}\left[x_{t-1}^2 | x_{t-1}\right] + 2\rho x_{t-1} \mathbb{E}[\varepsilon_t | x_{t-1}] + \mathbb{E}\left[\varepsilon_t^2 | x_{t-1}\right] \\
&= 1 + \rho^2 x_{t-1}^2 + 2\rho x_{t-1} \mathbb{E}[\varepsilon_t] + \mathbb{E}\left[\varepsilon_t^2\right] \\
&= 1 + \sigma^2 + \rho^2 x_{t-1}^2 \\
&= \rho^2 \delta(x_{t-1}) + c, \qquad \text{(I.25)}
\end{aligned}$$

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 $\eta_t = (\delta(x_t) - \mathbb{E}(\delta(x_t)|x_{t-1}))$, such that by definition $\mathbb{E}[\eta_t] = 0$. Thus if $\rho^2 < 1$, $\delta(x_t)$ resembles a stationary AR(1) process.

Example I.4.6 *With ARCH(1) process in (I.6) and $\delta(x_t) = 1 + x_t^2$,*

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

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.

For the dynamics of the drift function we make the following assumption:

Assumption I.4.2 *Assume that $\{X_t\}_{t=0,1,2,..}$, with $X_t \in \mathbb{R}^p$, satisfies Assumption I.4.1. With drift function δ, $\delta(X_t) \geq 1$, assume that there are positive constants M, C and ϕ, $\phi < 1$, such that for some $m \geq 1$,*

$$\begin{array}{lll} (i) & \mathbb{E}\left[\delta\left(X_{t+m}\right)|X_t = X\right] \leq \phi\delta\left(X\right) & \text{for } X'X > M, \\ (ii) & \mathbb{E}\left[\delta\left(X_{t+m}\right)|X_t = X\right] \leq C < \infty & \text{for } X'X \leq M. \end{array}$$

If $\{X_t\}$ satisfies Assumption I.4.2 then we say that X_t satisfies the drift criterion with drift function $\delta(\cdot)$. Note also that a simple way to verify *(i)*, is to show that if $\mathbb{E}\left[\delta\left(X_{t+m}\right)|X_t = X\right] \leq g(X)$, say, then *(i)* holds if

$$g(X)/\delta(X) \to \phi \text{ as } X'X \to \infty.$$

Example I.4.7 *The AR process x_t in satisfies the drift criterion if $\rho^2 < 1$ with $\delta(x_t) = 1 + x_t^2$ with x_{t-1}^2 chosen large. Recall from Example I.4.4,*

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

hence (i) holds with $\rho^2 < \phi < 1$ for $x_{t-1}^2 > M$, M large enough. For $x_{t-1}^2 \leq M$,

$$\mathbb{E}\left(\delta\left(x_t\right)|x_{t-1}\right) = \rho^2\delta\left(x_{t-1}\right) + c \leq \rho^2\delta\left(M\right) + c = C.$$

Example I.4.8 *The ARCH process x_t satisfies the drift criterion if $\alpha < 1$ with $\delta(x_{t-1}) = 1 + x_{t-1}^2$. By Example I.4.6,*

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

and we see that the considerations in Example I.4.7 can be applied here with $\alpha = \rho^2$.

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

Theorem I.4.1 *Assume that $\{X_t\}_{t\geq 0}$ satisfies Assumption I.4.2 with drift function δ. Then X_0 can be given an initial distribution such that X_t initiated in X_0 is stationary. With X_t^* denoting the stationary version, $\mathbb{E}\left[\delta\left(X_t^*\right)\right] < \infty$. Moreover, X_t is mixing in the sense that, for any initial value X_0, the LLN in Theorem I.4.2 below holds.*

Restate

More precisely, X_t satisfying Theorem I.4.1 is referred to as being "geometrically ergodic" in the Markov chain literature, see e.g. Tjøstheim (1990) and Francq and Zakoian (2019) for details. The idea is that, with $f^{(n)}(y|x)$ denoting the (n-step) density of X_{t+n} conditional on X_t, and $f^*(\cdot)$ the density of the stationary solution X_t^*, then

$$f^{(n)}(y|x) \to f^*(y)$$

exponentially fast as $n \to \infty$. And, importantly, the exponential speed implies that a LLN (as well as a CLT) applies.

Thus, if X_t satisfies the drift criterion, not only can the process be considered stationary (by giving X_0 the correct distribution) and geometrically ergodic, but also the LLN applies, independently of the initial value. Moreover, as $\mathbb{E}(\delta(X_t^*)) < \infty$, any moments of X_t which are bounded by the drift function δ are finite.

Example I.4.9 *For the AR(1) process we may conclude from Example I.4.5 that $\mathbb{E}[x_t^{*2}] < \infty$, and that the law of large numbers apply to x_t by Theorem I.4.1. 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 x_t^*, and since it is Gaussian, in fact, $\mathbb{E}\left[(x_t^*)^{2k}\right] < \infty$ for any k. To give an understanding of the role of initial value x_0, use that simple recursion gives*

$$x_{t+n} = \rho^n x_t + \sum_{i=0}^{n-1} \rho^i \varepsilon_{t+n-i},$$

Hence the n-step transition density is given by,

$$f^{(n)}(y|x) = \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left(-\frac{(y-\rho^n x)^2}{2\sigma_n^2}\right), \qquad \sigma_n^2 = \sigma^2 \frac{1-\rho^{2n}}{1-\rho^2}.$$

Clearly, $f^{(n)}(y|x) \to f^*(y)$, corresponding to the stationary version,

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

Observe in particular that $\rho^m \to 0$ exponentially fast.

Example I.4.10 *For the ARCH(1) 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 conclude that if $0 \leq \alpha < 1$ then x_t has a stationary solution with $\mathbb{E}\left[x_t^{*2}\right] < \infty$. Hence any moments of order lower than 2 are finite, for example $\mathbb{E}\left[\left|x_t^*\right|\right] < \infty$ since $|x| \leq \delta(x) = 1 + x^2$. However, we do not know if for example x_t has finite fourth order moments, $\mathbb{E}\left[x_t^4\right] < \infty$. To find out under which conditions this holds we need to consider a drift function from which we can conclude this. An example is $\delta(x_t) = 1 + x_t^4$, where using $\mathbb{E}\left[z_t^4\right] = 3$, we find,

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

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 $3\alpha^2 < 1$. Hence the conclusion is that while a stationary x_t exists for $\alpha < 1$ and $\mathbb{E}\left[x_t^{*2}\right] < \infty$ 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 $\mathbb{E}\left[(x_t^*)^4\right] < \infty$.

The considerations for the ARCH(1) illustrated that the value of α in the conditional variance $\sigma_t^2 = \sigma^2 + \alpha x_{t-1}^2$ was crucial for stationarity of x_t and also for the existence of 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.

A key implication of Theorem I.4.1 is that X_t is geometrically ergodic and the following LLN applies from Jensen and Rahbek (2007).

Theorem I.4.2 *Assume that with $X_t \in \mathbb{R}^p$, $\{X_t\}_{t=0,1,2,3,...}$ is a geometrically ergodic Markov chain with stationary solution $\{X_t^*\}$. Assume furthermore that the function $g : \mathbb{R}^{p(m+1)} \to \mathbb{R}$, $m \geq 0$, satisfies $\mathbb{E}\left[\left|g\left(X_t^*, X_{t-1}^*, ..., X_{t-m}^*\right)\right|\right] < \infty$, then as $T \to \infty$,*

$$\frac{1}{T}\sum_{t=1}^{T} g\left(X_t, X_{t-1}, ..., X_{t-m}\right) \overset{P}{\to} \mathbb{E}\left[g\left(X_t^*, X_{t-1}^*, ..., X_{t-m}^*\right)\right]. \tag{I.26}$$

We emphasize that the LLN holds for any initial value X_0, which is important for the later statistical analysis.

Note in that respect that $\mathbb{E}[\cdot]$ in (I.26) means the expectation under the stationary measure (X_t^* stationary), while in the average

$$\frac{1}{T}\sum_{t=1}^{T} g\left(X_t, X_{t-1}, ..., X_{t-m}\right),$$

X_0 does not have to be equipped with the stationary distribution and X_t therefore not stationary. Henceforth, unless important to make the distinction between X_t and X_t^*, we may sometimes just write

$$\frac{1}{T}\sum_{t=1}^{T} g\left(X_t, X_{t-1}, ..., X_{t-m}\right) \xrightarrow{P} \mathbb{E}\left[g\left(X_t, X_{t-1}, ..., X_{t-m}\right)\right].$$

Example I.4.11 *With X_t is univariate, and geometrically ergodic with finite second order moments, then*

$$\frac{1}{T}\sum_{t=1}^{T} X_t^2 \xrightarrow{P} \mathbb{E}\left[X_t^2\right] \quad \textit{and} \quad \frac{1}{T}\sum_{t=1}^{T} X_t X_{t-1} \xrightarrow{P} \mathbb{E}\left[X_t X_{t-1}\right],$$

by applying Theorem I.4.2 with

$$g\left(X_t\right) = X_t^2 \ \textit{and}\ g\left(X_t, X_{t-1}\right) = X_t X_{t-1}.$$

Example I.4.12 *If $\mathbb{E}\left[X_t\right] = 0$, such that $\mathbb{V}\left[X_t\right] = \mathbb{E}\left[X_t^2\right]$ and $\operatorname{Cov}\left[X_t, X_{t+h}\right] = \mathbb{E}\left[X_t X_{t+h}\right]$, it follows that the empirical autocorrelation function, see (I.23), as defined by,*

$$\widehat{\rho}(h) \equiv \frac{\frac{1}{T}\sum_{t=1}^{T-h} X_t X_{t+h}}{\sqrt{\frac{1}{T}\sum_{t=1}^{T} X_t^2 \frac{1}{T}\sum_{t=1}^{T-h} X_{t+h}^2}}, \tag{I.27}$$

will converge in probability to (the theoretical) $\rho\left(h\right)$ by Theorem I.4.2 motivating that most software for time series allows one to compute these directly.

As emphasized for the AR(1) process x_t, the autoregressive coefficient ρ is the key parameter to an understanding of the dynamics. We know that if $|\rho| < 1$, x_t is stationary (or, a stationary solution exists) and with ε_t Gaussian all moments are finite. At the same time we know that the restriction of $|\rho| < 1$ is crucial in the sense that if $\rho = 1$, as often found empirically, then x_t is a so-called “unit-root” process which is non-stationary. Surprisingly this is not so for the ARCH(1) process, where $\alpha = 1$ still allows x_t to be stationary. However, with $\alpha = 1$, then only moments up to order one are finite, $\mathbb{E}\left[|x_t|\right] < \infty$. That is, $\alpha = 1$ implies stationarity but for a process with infinite 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.6):	
$x_t = \sigma_t z_t, \;\; \sigma_t = \sigma^2 + \alpha x_{t-1}^2$ and z_t $i.i.d.$N$(0,1)$.	
Stationary for $0 \leq \alpha < 3.56$	
Finite moments:	
$\frac{\pi}{2} \leq \alpha < 3.56$	$\mathbb{E}\left[\lvert x_t \rvert^{\delta}\right] < \infty$, some $\delta \in (0,1)$
$1 \leq \alpha < \frac{\pi}{2}$	$\mathbb{E}\left[\lvert x_t \rvert\right] < \infty$
$\frac{1}{\sqrt{3}} \leq \alpha < 1$	$\mathbb{E}\left[x_t^2\right] < \infty$
$0 \leq \alpha < \frac{1}{\sqrt{3}}$	$\mathbb{E}\left[x_t^4\right] < \infty$

Actually the number 3.56 is an approximation of $\exp(-\mathbb{E}\left[\log\left(z^2\right)\right])$, with z N$(0,1)$ distributed. In fact, if z_t has another distribution such as the t_v distribution mentioned in Example I.4.3 the intervals above would change.

I.4.1 Central limit theorem

As noted, a powerful implication of Theorem I.4.1 for X_t is that independently of the initial value, X_0 the LLN in Theorem I.4.2 applies. 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 for Y_t a continuous function of X_t and, possibly, its lagged values, that is, $Y_t = f\left(X_t, X_{t-1}, ..., X_{t-m}\right)$:

Theorem I.4.3 (Meyn and Tweedie, 1993, Theorem 17.0.1) *Assume that Theorem I.4.1 applies to* $\{X_t\}_{t\geq 0}$, $Y_t = f\left(X_t, X_{t-1}, ..., X_{t-m}\right)$. *With* $Y_t^* = f\left(X_t^*, X_{t-1}^*, ..., X_{t-m}^*\right)$, *suppose that* $\mathbb{E}\left[Y_t^*\right] = 0$, $\mathbb{E}\left[\left(Y_t^*\right)^2\right] < \infty$ *and*

$$\gamma = \lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\left(\sum_{t=1}^{T} Y_t^*\right)^2\right] > 0.$$

Then as $T \to \infty$,

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Y_t \xrightarrow{D} N(0,\gamma).$$

A different version of the CLT is a corrollary from Brown (1971) and we will often apply this (or a variant thereof) when discussing asymptotic normality later in the statistical analysis. As above, consider the mapping $Y_t = f\left(X_t, X_{t-1}, ..., X_{t-m}\right)$ with $f\left(\cdot\right)$ continuous, then Y_t is an example of a Martingale difference (MGD) sequence wrt. $\mathcal{F}_t = \left(X_t, ..., X_0\right)$, if $\mathbb{E}\left[Y_t | \mathcal{F}_{t-1}\right] = 0$.

Theorem I.4.4 (Corollary to Brown, 1971) *For a given sequence $\{X_t\}_{t\geq 0}$, consider $Y_t = f(X_t, X_{t-1}, ..., X_{t-m})$, $f(\cdot)$ continuous, with $\mathbb{E}[Y_t|\mathcal{F}_{t-1}] = 0$, where $\mathcal{F}_t = (X_t, ..., X_0)$. If, as $T \to \infty$,*

$$\textit{(i):}\ T^{-1}\sum_{t=1}^{T}\mathbb{E}\left[Y_t^2|\mathcal{F}_{t-1}\right] \xrightarrow{P} \sigma^2 > 0$$

and either (ii) or (ii') hold,

$$\textit{(ii):}\ T^{-1}\sum_{t=1}^{T}\mathbb{E}\left[Y_t^2\mathbb{I}\left(|Y_t| > \delta T^{1/2}\right)\right] \to 0,$$
$$\textit{(ii'):}\ T^{-1}\sum_{t=1}^{T}\mathbb{E}\left[Y_t^2\mathbb{I}\left(|Y_t| > \delta T^{1/2}\right)|\mathcal{F}_{t-1}\right] \xrightarrow{P} 0,$$

for any $\delta > 0$, then $\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Y_t \xrightarrow{D} N(0, \sigma^2)$.

Note that if $\{X_t\}_{t\geq 0}$ satisfies Theorem I.4.1, and $\mathbb{E}[Y_t^{*2}] < \infty$, then (i) in Theorem I.4.4 holds by the LLN in Theorem I.4.2. As to the so-called "Lindeberg"-type conditions in (ii) and (ii'), note that even though (ii) is "classic", also (ii') is stated here in as it often is useful in the context of time series. Also one may note that (ii) holds under stationarity (and dominated convergence), since if Y_t is stationary with $\mathbb{E}[|Y_t|^{2+\eta}] < \infty$, for some $\eta > 0$, we have

$$T^{-1}\sum_{t=1}^{T}\mathbb{E}\left[Y_t^2\mathbb{I}\left(|Y_t| > \delta T^{1/2}\right)\right] \leq \frac{1}{\left(\delta T^{1/2}\right)^{\eta}}\mathbb{E}\left[|Y_t|^{2+\eta}\right] \to 0.$$

Various different versions of the CLT for MGDs exist, including more general definitions of the MGD properties, and well as different versions of the Lindeberg conditions (ii) and (ii'), see e.g. Brown (1971).

As an example of the application of the CLTs, consider:

Example I.4.13 *The empirical autocovariance function of order one for the ARCH(1) process x_t, is given by,*

$$\frac{1}{T}\sum_{t=1}^{T} x_t x_{t-1}. \tag{I.28}$$

If $\alpha < 1$, such that $\mathbb{E}[x_t^2] < \infty$, the LLN indeed implies the obvious result that as T tends to ∞, then (I.28) will converge in probability to $\mathbb{E}[x_t x_{t-1}] = 0$

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.4.4 would apply to (I.28). 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* $\mathbb{E}[Y_t | x_{t-1}] = 0$ *as desired. Moreover, if* $\mathbb{E}[x_t^4] < \infty$, *or* $\alpha < 1/\sqrt{3}$,

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

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

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

with $\mathbb{E}\left[x_{t-1}^2\right]$ *and* $\mathbb{E}\left[x_{t-1}^4\right]$ *given in (I.19) and (I.20) respectively.*

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

$$\frac{1}{T}\sum_{t=1}^{T} x_t x_{t-1} \xrightarrow{P} \mathbb{E}[x_t x_{t-1}] = 0,$$

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

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

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

$$\frac{1}{T}\sum_{t=1}^{T}\left(x_t^2 - \left(\frac{1}{T}\sum_{t=1}^{T} x_t^2\right)\right) x_{t-1}^2. \tag{I.29}$$

Considerations as above lead to the requirement that $\mathbb{E}[x_t^4] < \infty$ *for convergence in probability, while by using Theorem I.4.3 the requirement is* $\mathbb{E}[x_t^8] < \infty$ *for convergence in distribution. These are quite strong restrictions, and therefore to avoid such, often the autocorrelation function is given for* $|x_t|$ *instead.*

I.5 Extending the AR(1)

A key process of interest is the vector version of the AR(1), the VAR(1), which for $X_t \in \mathbb{R}^p$, is given by

$$X_t = AX_{t-1} + \varepsilon_t, \quad t = 1, 2, ..., T \tag{I.30}$$

with initial value X_0 and ε_t i.i.d. $\mathrm{N}(0, \Omega)$, Ω symmetric and positive definite, $\Omega > 0$. Moreover, the autoregressive matrix $A = (A_{ij})_{i,j=1,2,\ldots,p} \in \mathbb{R}^{p \times p}$, such that for $p = 2$, we have with $X_t = (X_{1t}, X_{2t})'$ and $\varepsilon_t = (\varepsilon_{1t}, \varepsilon_{2t})'$,

$$\begin{pmatrix} X_{1t} \\ X_{2t} \end{pmatrix} = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} X_{1t-1} \\ X_{2t-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{pmatrix} \tag{I.31}$$

or

$$\begin{aligned} X_{1t} &= A_{11} X_{1t-1} + A_{12} X_{2t-1} + \varepsilon_{1t} \\ X_{2t} &= A_{21} X_{1t-1} + A_{22} X_{2t-1} + \varepsilon_{2t} \end{aligned} \tag{I.32}$$

By definition of the multivariate Gaussian distribution, the transition density of $X_t \in \mathbb{R}^p$ conditional on X_{t-1} is given by

$$f(X_t | X_{t-1}) = \det(\Omega)^{-1/2} \left(1/\sqrt{2\pi}\right)^p \exp\left(-\frac{1}{2}(X_t - AX_{t-1})' \Omega^{-1} (X_t - AX_{t-1})\right)$$

which thus satisfies Assumption I.4.1. Moreover, as for the univariate case we have

$$X_t = \sum_{i=0}^{t-1} A^i \varepsilon_{t-i} + A^t X_0,$$

which, if $A^t \to 0$ as $t \to \infty$, resembles the stationary solution,

$$X_t^* = \sum_{i=0}^{\infty} A^i \varepsilon_{t-i}. \tag{I.33}$$

To see that a stationary solution exists, we may use the drift criterion with $\delta(X) = 1 + X'X = 1 + \|X\|^2$. Note first that,

$$X_{t+m} = \sum_{i=0}^{m-1} A^i \varepsilon_{t+m-i} + A^m X_t. \tag{I.34}$$

Recall that with $A \in \mathbb{R}^{p \times p}$, then $\mathrm{tr}(A) = \sum_{i=1}^p A_{ii}$, and hence by the properties of $\mathrm{tr}(\cdot)$, we have $\|X\|^2 = \mathrm{tr}(X'X) = \mathrm{tr}(XX')$, see e.g. Magnus and Neudecker (2007) for further properties and general results for matrix calculus. Using this and (I.34), we find

$$\begin{aligned} \mathbb{E}[\delta(X_{t+m}) | X_t = X] &= 1 + \mathbb{E}\left[\mathrm{tr}\left(\sum_{i=0}^{m-1} A^i \varepsilon_{t+m-i} \varepsilon'_{t+m-i} \left(A^i\right)'\right)\right] + \|A^m X\|^2 \\ &= 1 + \underbrace{\mathrm{tr}\left(\sum_{i=0}^{m-1} A^i \Omega \left(A^i\right)'\right)}_{D_1} + \underbrace{\|A^m X\|^2}_{D_2} \end{aligned}$$

By Lemma A.1, in terms of the (sup-)matrix norm $\|\cdot\|$, we have

$$D_2 = \|A^m X\|^2 \leq \|A^m\|^2 \|X\|^2 .$$

With $\rho(A)$ denoting the spectral radius, that is, the largest (in absolute value) of the eigenvalues of A, we have $\|A^m\| \to 0$ as $m \to \infty$, provided $\rho(A) < 1$, see Lemma A.3. In particular this means that for $m > m^*$, say, if $\rho(A) < 1$,

$$\|A^m\| < \phi, \text{ with } \phi < 1.$$

Next, by Lemma A.4, as $\Omega > 0$ and $\rho(A) < 1$, with c a generic constant,

$$D_1 \leq \left\| \sum_{i=0}^{m-1} A^i \Omega \left(A^i\right)' \right\| \leq c \|\Omega\| \sum_{i=0}^{m-1} [\rho(A)]^i \to c \|\Omega\| (1 - \rho(A))^{-1} < \infty.$$

Hence we conclude that the drift criterion is satisfied with $\delta(X) = 1 + \|X\|^2$, as for $m > m^*$, and $\|X\|^2 > M$,

$$\mathbb{E}[\delta(X_{t+m}) | X_t = X] \leq g(X), \quad \text{with } g(X) = c + \phi \|X\|^2 ,$$

while $\mathbb{E}[\delta(X_{t+m}) | X_t = X] \leq c$ for $\|X\|^2 \leq M$, using Lemma A.4(iii) .

More generally, if $\rho(A) < 1$, the (unique) stationary solution is given by (I.33), and from properties of the Gaussian distribution, X_t^* have all moments finite.

Remark I.5.1 *Often in the time series literature, the condition $\rho(A) < 1$ is stated in terms of the so-called characteristic polynomial, $A(z) = I_p - Az$, $z \in \mathbb{C}$. Here the condition that $\det(A(z)) = 0$ implies $|z| > 1$, is equivalent to $\rho(A) < 1$. The condition is referred to as "the roots of the characteristic polynomial are larger than one in absolute value".*

I.5.1 Further extensions

Often one is interested in adding further lags, and consider therefore the VAR(k) process as given by

$$X_t = A_1 X_{t-1} + ... + A_k X_{t-k} + \varepsilon_t$$

with initial values $X_0, X_{-1}, ..., X_{-k}$ and ε_t i.i.d. $\mathrm{N}(0, \Omega)$. Obviously, in this case X_t does not satisfy Assumption I.4.1 as X_t by definition depends on $X_{t-1}, ..., X_{t-k}$. However, a way to circumvent this is to consider the so-called companion form in terms of

$$Y_t = \left(X_t', ..., X_{t-k-1}'\right)' \in \mathbb{R}^{pk}$$

which satisfies,

$$Y_t = AY_{t-1} + \epsilon_t,$$

where

$$A = \left[\begin{array}{cccc} A_1 & \cdots & A_{k-1} & A_k \\ I_p & & & \\ & \ddots & & \\ & & I_p & \end{array} \right]$$

and $\epsilon_t = \left(\varepsilon_t', 0_p', ..., 0_p' \right)'$. Hence $Y_t \in \mathbb{R}^{pk}$ is indeed a Markov chain, has k-step transition density given by

$$\begin{aligned} f\left(Y_{t+k} | Y_t\right) &= f\left(\left(X_{t+k}, ..., X_{t+1}\right) | \left(X_t, ..., X_{t-k}\right)\right) \\ &= \frac{f\left(X_{t+k}, ..., X_{t-k}\right)}{f\left(X_t, ..., X_{t-k}\right)} \\ &= \prod_{m=1}^{k} f\left(X_{t+m} | X_{t+m-1}, ..., X_{t+m-k}\right), \end{aligned}$$

with $f\left(X_{t+m} | X_{t+m-1}, ..., X_{t+m-k}\right)$ positive and continuous. Hence, as desired, Assumption I.4.1 holds for the Markov chain Y_t.

Application of the drift criterion gives, as for the VAR(1), that a sufficient condition for stationarity is indeed $\rho(A) < 1$. As for the VAR(1), this may alternatively be stated in terms of the characteristic polynomial which here is given by, $A(z) = I - \sum_{i=1}^{k} A_i z^i$, and the condition is as for the VAR(1), $\det(A(z)) = 0$, implies $|z| > 1$.

References

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

Durrett, R. (2019) Probability Theory and Examples, 5th Edition, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press.

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

Horn, R.A. and Johnson, C.R. (2013) Matrix Analysis, Second Edition, Cambridge University Press.

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

Johansen, S. (1996) Likelihood-Based Inference in Cointegrated Vector Autoregressive Models, Oxford University Press.

Magnus, J. and Neudecker, H. (2007) Matrix Differential Calculus with Applications in Statistics and Econometrics, Third Edition, Wiley.

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

Tjøstheim, D. (1990) Non-Linear Time Series and Markov Chains, *Advances in Applied Probability*, 22:587–611.

Tsay, R.S. (2010) Analysis of Financial Time Series, Wiley Series in Probability and Statistics, Third Edition, Wiley.

A Matrix results

Results on matrices and matrix norms can be found several places. The results listed here are from Horn and Johnson (2013).

Let M, N be real $p \times p$ matrices, then a matrix norm $\|\cdot\| : \mathbb{R}^{p\times p} \to \mathbb{R}$ satisfies

(i) $\|M\| \geq 0$

(ii) $\|M\| = 0$ if and only if $M = 0$

(iii) $\|aM\| = |a| \, \|M\|$ for $a \in \mathbb{R}$

(iv) $\|M + N\| \leq \|M\| + \|N\|$

(v) $\|MN\| \leq \|M\| \, \|N\|$.

Key examples of norms of matrices include the Euclidean norm given by,

$$\|M\|^2 = \mathrm{tr}\,(MM') = \sum_{i,j} |m_{ij}|^2,$$

and the "sup"-norm as given by,

$$\|M\| = \sup\left\{\|MX\| \;\mid X \in \mathbb{R}^p \text{ and } \;\|X\| \leq 1\right\}.$$

Here $\|X\|^2 = X'X$, $X \in \mathbb{R}^p$, and unless otherwise specified this (the Euclidean) norm for vectors will be used.

Lemma A.1 *With M a real $p \times p$ matrix, $\|\cdot\|$ the sup norm, and $X \in \mathbb{R}^p$,*

$$\|MX\| \leq \|M\| \, \|X\| \tag{I.35}$$

Lemma A.1 is stated in Theorem 5.6.2 in Horn and Johnson (2013) as a a simple consequence of the definition of the sup-norm.

For any $p \times p$ matrix M, whether symmetric or not, the eigenvalues λ_i, $i = 1, ..., p$, solve

$$\det\left(\lambda I_p - M\right) = 0$$

and the spectral radius $\rho(M)$ of M is given by

$$\rho(M) \equiv \max_{i=1,\ldots,p} \{|\lambda_i|\} \tag{I.36}$$

This is an alternative measure of the 'size' of M and is related to any matrix norm by the following (Theorem 5.6.9 of Horn and Johnson, 2013) as follows:

Lemma A.2 *For any matrix norm* $\|\cdot\|$,

$$\rho(M) \leq \|M\|. \tag{I.37}$$

This reflects that M can have an arbitraily large norm, while the eigenvalues of M are small such as for

$$M = \begin{pmatrix} 0 & m_{12} \\ 0 & 0 \end{pmatrix}$$

with $\rho(M) = 0$, while $\|M\|$ can be made arbitraily large through m_{12}.

In the analysis of dynamic systems it is of interest to study iterations of the form $M^n X$ and the following results provide useful (Horn and Johnson, 2013, Theorem 5.6.12):

Lemma A.3 *With M a real $p \times p$ matrix, and any matrix norm* $\|\cdot\|$,

$$\lim_{n\to\infty} (M^n) = 0 \quad \textit{if and only if } \rho(M) < 1 \tag{I.38}$$

$$\rho(M) = \lim_{n\to\infty} \|M^n\|^{1/n} \tag{I.39}$$

Used frequently are the results addressing convergence of sums of M^n (Horn and Johnson, 2013, Lemma 5.6.10, see also Johansen, 1996, Corollary A.2):

Lemma A.4 *Assume that M a real $p \times p$ matrix, with eigenvalues $(\lambda_i)_{i=1,\ldots,p}$ and $\rho(M) < 1$. It then holds that:*

(i) *There exists a $c > 0$ such that for any $n \geq 0$, $\|M^n\| \leq c \cdot \rho(M)^{n/2}$*

(ii) $\lim_{n\to\infty} \left(\sum_{i=0}^{n} M^i\right) = \sum_{i=0}^{\infty} M^i = (I_p - M)^{-1}$

(iii) *With $\Omega > 0$, $\lim_{n\to\infty} \left(\sum_{i=0}^{n} M^i \Omega M^{i\prime}\right) = \sum_{i=0}^{\infty} M^i \Omega M^{i\prime} > 0$,*

with $\|\sum_{i=0}^{\infty} M^i \Omega M^{i\prime}\| < \infty$.

Proof: The results in *(ii)* and *(iii)* follow by Lemma A.3. To see this note that,

$$\sum_{i=0}^{n} M^i = \left(I_p - M^{n+1}\right)(I_p - M)^{-1}$$

since as $\rho(M) < 1$, all $|\lambda_i| < 1$. Hence *(ii)* holds by $M^{n+1} \to 0$. Likewise, for *(ii),*

$$\left\|\sum_{i=0}^{n} M^i \Omega M^{i\prime}\right\| \leq \|\Omega\| \sum_{i=0}^{n} \left\|M^i\right\|^2 \leq c^2 \|\Omega\| \sum_{i=0}^{n} \rho(M)^{2i} \to c^2 \|\Omega\| \left(1 - \rho(M)^2\right)^{-1}.$$

The proof of *(i)* is more involved. Exploiting the Jordan form of M , cf. Lemma A.1 in Johansen (1996), gives that

$$\|M^n\| \leq \max_{i=1,\ldots,p} |\lambda_i|^n P(n) \tag{I.40}$$

where $P(\cdot)$ is a polynomial of finite order (less than p). Hence, by

$$\|M^n\| \leq \rho(M)^n P(n) \leq \rho(M)^{n/2} \sup_n \rho(M)^{n/2} P(n) \leq c\rho(M)^{n/2} \tag{I.41}$$

the result in *(i)* follows. □

Anders Rahbek
Rasmus Søndergaard Pedersen
University of Copenhagen

September 2024

Part II
Autoregressive Models

II.1 Introduction

In this part likelihood inference is discussed for autoregressive models with focus on maximum likelihood (ML). The LLN for geometric ergodic processes from Part I is applied repeatedly, as is the CLT for martingale differences.

Initially the AR(1) model is considered and estimation as well as asymptotic inference is treated. And a general linear regression model is discussed which is useful for the analysis of AR(k), VAR(k) models, and variants thereof.

II.2 AR(1) Model

The autoregressive model of order one, the AR(1) model, is given by

$$x_t = \rho x_{t-1} + \varepsilon_t, \quad t = 1, ..., T \tag{II.1}$$

with x_0 fixed, $\rho \in \mathbb{R}$ and ε_t i.i.d. $\mathrm{N}(0, \sigma^2)$. By definition, the density of x_t conditional on x_{t-1}, $f(x_t|x_{t-1})$, is the Gaussian density with mean ρx_{t-1} and variance σ^2. Moreover the joint density of $\{x_t\}_{t=1,...,T}$, with the initial value x_0 fixed, factorizes as follows

$$f(x_T, x_{T-1}, ..., x_1|x_0) = \prod_{t=1}^{T} f(x_t|x_{t-1}). \tag{II.2}$$

Denote the likelihood function by $L(\rho, \sigma^2)$, then by the factorization in (II.2) of the joint density of $x_1, ..., x_T$ given x_0, the log-likelihood function is given by

$$\log L\left(\rho,\sigma^2\right) = \log\left(\prod_{t=1}^{T} f\left(x_t|x_{t-1}\right)\right) \tag{II.3}$$
$$= -\frac{T}{2}\log\left(2\pi\right) - \frac{T}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{t=1}^{T}\left(x_t - \rho x_{t-1}\right)^2.$$

The maximum likelihood estimators (MLEs) are denoted $\hat{\rho}$ and $\hat{\sigma}^2$ respectively. Note in this respect that often term $\log(2\pi)$ in (II.3) is omitted as it is not a function of the parameters ρ and σ^2, and therefore plays no role when discussing maximization of the log-likehood function.

Theorem II.2.1 *The MLEs of ρ and σ for the AR(1) model are given by*

$$\hat{\rho} = S_{yz}S_{zz}^{-1} \tag{II.4}$$
$$\hat{\sigma}^2 = \frac{1}{T}\sum_{t=1}^{T}\left(x_t - \hat{\rho}x_{t-1}\right)^2 = S_{yy\cdot z} = S_{yy} - S_{yz}S_{zz}^{-1}S_{yz}$$

where, with $y_t = x_t$ and $z_t = x_{t-1}$, the product moments are given by $S_{yz} = \frac{1}{T}\sum_{t=1}^{T} i_t j_t$, with $i, j = y, z$. The maximized likelihood function is (apart from a constant factor) given by:

$$L_{\max}\left(\hat{\rho},\hat{\sigma}^2\right) = \left(\hat{\sigma}^2\right)^{-T/2}. \tag{II.5}$$

Proof: Simple diffentiation of (II.3) with respect to ρ and σ^2 gives the first order conditions:

$$\sum_{t=1}^{T}\left(x_t - \rho x_{t-1}\right)x_{t-1} = 0, \qquad \frac{1}{T}\sum_{t=1}^{T}\left(x_t - \rho x_{t-1}\right)^2 = \sigma^2.$$

The first equality leads to, $\hat{\rho} = S_{yz}S_{zz}^{-1}$, and substitution in the second shows that $\hat{\sigma}^2$ is given by the residual sum of squares,

$$\hat{\sigma}^2 = \frac{1}{T}\sum_{t=1}^{T}\left(x_t - \hat{\rho}x_{t-1}\right)^2 = S_{yy\cdot z}$$

The second order derivatives evaluated at $(\hat{\rho},\hat{\sigma}^2)$ equal,

$$\left.\frac{\partial^2}{\partial\rho^2}\log L\right|_{(\hat{\rho},\hat{\sigma}^2)} = -\frac{T}{\hat{\sigma}^2}S_{zz}, \qquad \left.\frac{\partial^2}{\partial\left(\sigma^2\right)^2}\log L\right|_{(\hat{\rho},\hat{\sigma}^2)} = -\frac{T}{2\hat{\sigma}^4} \quad \text{and}$$
$$\left.\frac{\partial^2}{\partial\sigma^2\partial\rho}\log L\right|_{(\hat{\rho},\hat{\sigma}^2)} = 0,$$

and $L(\rho, \sigma^2)$ has a maximum $(\hat{\rho}, \hat{\sigma}^2)$ given by (II.5). □

Next consider the asymptotic properties of the MLEs.

We let ρ_0 and σ_0^2 denote the so-called "true-values", i.e. the values of the parameters ρ and σ^2 under which the probabilisitic arguments are made.

Theorem II.2.2 *For $|\rho_0| < 1$, the ML estimators of the AR(1) model in (II.1) are consistent, $\hat{\rho} \xrightarrow{P} \rho_0$ and $\hat{\sigma}^2 \xrightarrow{P} \sigma_0^2$, as $T \to \infty$. Moreover,*

$$\sqrt{T}(\hat{\rho} - \rho_0) \xrightarrow{D} \mathrm{N}\left(0, 1 - \rho_0^2\right) \tag{II.6}$$

and a consistent estimator of the asymptotic variance is given by $\hat{\sigma}^2 S_{zz}^{-1}$, such that

$$\sqrt{T}(\hat{\rho} - \rho_0)\sqrt{S_{zz}/\hat{\sigma}^2} \xrightarrow{D} \mathrm{N}(0, 1),$$

as $T \to \infty$.

Note the requirement that $|\rho_0| < 1$ which is crucial. If, say $\rho_0 = 1$, as often met in the analysis of stock market prices, a different asymptotic distribution applies. This is discussed separately when dealing with so-called unit roots and cointegration.

How well the asymptotic distribution applies for finite T may be investigated by simulation studies where for different known ρ_0, the empirical distribution of $\hat{\rho}$ is studied. Loosely speaking the approximation by a Gaussian distribution as stated works well for even small samples provided $|\rho_0|$ is not "close to one", i.e. for the cases where there are no unit roots.

Proof: Recall that as $|\rho_0| < 1$, then x_t is mixing, or geometrically ergodic, and the LLN in Theorem I.4.2 applies. Consider $\hat{\rho}$ as given by,

$$\hat{\rho} = S_{yz} S_{zz}^{-1} = \left(\frac{1}{T}\sum_{t=1}^{T} x_t x_{t-1}\right)\left(\frac{1}{T}\sum_{t=1}^{T} x_{t-1}^2\right)^{-1}. \tag{II.7}$$

By the LLN for applied to $x_t x_{t-1}$ and x_{t-1}^2, and as $\mathbb{E}\left[x_t^{*2}\right] < \infty$, it follows directly that

$$\hat{\rho} \xrightarrow{P} \mathrm{Cov}\left[x_t^*, x_{t-1}^*\right] / \mathbb{V}\left[x_t^*\right] = \rho_0.$$

Likewise

$$\begin{aligned}\hat{\sigma}^2 &\xrightarrow{P} \mathbb{V}\left[x_t^*\right] - \left(\mathrm{Cov}\left[x_t^*, x_{t-1}^*\right]\right)^2 / \mathbb{V}\left[x_t^*\right] \\ &= \frac{\sigma_0^2}{1 - \rho_0^2} - \rho_0^2 \frac{\sigma_0^2}{1 - \rho_0^2} = \sigma_0^2\end{aligned}$$

as claimed, and also the proposed variance estimator is consistent, $\hat{\sigma}^2 S_{zz}^{-1} \xrightarrow{p} 1-\rho_0^2$.

For the asymptotic distribution note that,

$$\sqrt{T}\left(\hat{\rho}-\rho_0\right) = \frac{\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\varepsilon_t x_{t-1}}{\frac{1}{T}\sum_{t=1}^{T}x_{t-1}^2}$$

where $\frac{1}{T}\sum_{t=1}^{T}x_{t-1}^2 \xrightarrow{p} \frac{\sigma_0^2}{1-\rho_0^2}$ by the LLN. With $Y_t \equiv \varepsilon_t x_{t-1}$, Y_t is a martingale difference sequence with respect to $\mathcal{F}_t = (x_t, x_{t-1}, ..., x_0)$, as $\varepsilon_t = x_t - \rho_0 x_{t-1}$, and the CLT (Theorem I.4.4) can be applied. Observe first that,

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[Y_t^2|\mathcal{F}_{t-1}\right] = \sigma_0^2\left(\frac{1}{T}\sum_{t=1}^{T}x_{t-1}^2\right) \xrightarrow{p} \frac{\sigma_0^4}{1-\rho_0^2}.$$

Next, use that for a r.v. X with $\mathbb{E}\left[X^4\right] < \infty$, $\mathbb{E}\left[X^2\mathbb{I}\left(|X| > \delta\sqrt{T}\right)\right] \leq \mathbb{E}\left[X^4\right]/T^2\delta$, such that

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[Y_t^2\mathbb{I}\left(|Y_t| > \delta\sqrt{T}\right)|\mathcal{F}_{t-1}\right] \leq \frac{1}{T^2\delta^2}\sum_{t=1}^{T}\mathbb{E}\left[Y_t^4|\mathcal{F}_{t-1}\right] \tag{II.8}$$

$$= \frac{1}{T\delta^2}\left(\frac{1}{T}\sum_{t=1}^{T}x_{t-1}^4\right)\mathbb{E}\left[\varepsilon_t^4\right] \xrightarrow{p} 0.$$

We conclude that,

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}Y_t = \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\varepsilon_t x_{t-1} \xrightarrow{D} N\left(0, \sigma_0^4/\left(1-\rho_0^2\right)\right)$$

and the result for $\hat{\rho}$ follows by simple insertion. □

The reason for providing also a consistent estimator for the variance of the asymptotic distribution of $\hat{\rho}$ is that one may then apply the results directly in empirical analyses to report $\hat{\rho}$ as well as its empirical standard deviation, $\sqrt{\frac{1}{T}\hat{\sigma}^2 S_{zz}^{-1}}$. This implies that for example the hypothesis that $\rho = 0$, can be investigated by computing the classic t-ratio from regression analysis,

$$\hat{\rho}/\sqrt{\frac{1}{T}\hat{\sigma}^2 S_{zz}^{-1}} = \hat{\rho}\sqrt{\frac{\sum_{t=1}^{T}x_{t-1}^2}{\hat{\sigma}^2}} \tag{II.9}$$

which is asymptotically $\mathrm{N}(0,1)$ distributed. More generally, one can consider the simple hypothesis,

$$H : \rho = \rho_0$$

where ρ_0 is some known value and consider the likelihood ratio statistic, $\mathrm{LR}(\rho = \rho_0)$:

Theorem II.2.3 *The LR test statistic of the hypothesis $H : \rho = \rho_0$ in the AR(1) model in (II.1) is given by*

$$LR\left(\rho = \rho_0\right) = T \log\left(1 + W_T\right), \qquad W_T = \left(\hat{\rho} - \rho_0\right)^2 \frac{S_{zz}}{\hat{\sigma}^2}. \tag{II.10}$$

For $|\rho_0| < 1$, the LR statistic is asymptotically χ^2 distributed with one degree of freedom,

$$\mathrm{LR}\left(\rho = \rho_0\right) \xrightarrow{D} \chi_1^2 \quad as\ T \to \infty. \tag{II.11}$$

Note that the term

$$TW_T \equiv W \tag{II.12}$$

is known as the Wald statistic, which here is the t-ratio squared for the hypothesis that $\rho = \rho_0$.

Note also that in the W_T term the residual variance is estimated under the alternative by $\hat{\sigma}^2$. With $\tilde{\sigma}^2$ denoting the variance estimator under the hypothesis, that is,

$$\tilde{\sigma}^2 = \frac{1}{T} \sum_{t=1}^{T} \left(x_t - \rho_0 x_{t-1}\right)^2, \tag{II.13}$$

the $\mathrm{LR}(\rho = \rho_0)$ statistic may alternatively be written as

$$\mathrm{LR}\left(\rho = \rho_0\right) = -T \log\left(1 - \tilde{W}_T\right) \tag{II.14}$$

with $\tilde{W}_T \equiv \left(\hat{\rho} - \rho_0\right)^2 \frac{S_{zz}}{\tilde{\sigma}^2}$. This will be clear from the proof of Theorem II.2.3 where a fundamental decomposition is used:

Proof: Maximizing $L\left(\rho, \sigma^2\right)$ when $\rho = \rho_0$, see (II.3), immediately gives $\tilde{\sigma}^2$ in (II.13) and $L_{\max}\left(\rho_0, \tilde{\sigma}^2\right) = \left(\frac{1}{\tilde{\sigma}^2}\right)^{T/2}$. Hence the likelihood ratio statistic $\mathrm{LR}(\rho = \rho_0) = -2 \log Q$, where

$$Q^{-2/T} = \frac{\tilde{\sigma}^2}{\hat{\sigma}^2}.$$

This can be simplified by using the fundamental decomposition from regression analysis:

$$x_t - \rho_0 x_{t-1} = \left(x_t - \hat{\rho} x_{t-1}\right) + (\hat{\rho} - \rho_0) x_{t-1},$$

which implies that

$$\begin{aligned} T\tilde{\sigma}^2 &= \sum_{t=1}^{T}(x_t - \rho_0 x_{t-1})^2 = \sum_{t=1}^{T}(x_t - \hat{\rho}x_{t-1})^2 + (\hat{\rho}-\rho_0)^2\sum_{t=1}^{T}x_{t-1}^2 \\ &= T\hat{\sigma}^2 + T(\hat{\rho}-\rho_0)^2 S_{zz} \end{aligned}$$

since the first order condition for $\hat{\rho}$ implies that the double product vanishes. This shows that

$$Q^{-2/T} = 1 + \frac{S_{zz}}{\hat{\sigma}^2}(\hat{\rho}-\rho_0)^2 = 1 + W_T, \tag{II.15}$$

as claimed since $\text{LR}(\rho = \rho_0) = -2\log Q = T\log(1+W_T)$.

For the asymptotic distribution note first that by the results in Theorem II.2.2, when $|\rho_0| < 1$,

$$W_T \overset{P}{\to} 0 \quad \text{and } TW_T \overset{D}{\to} \chi_1^2.$$

Next, a Taylor expansion of $f(w) = \log(1+w)$ for $w \to 0$, gives $f(w) = w + o(w)$, and hence

$$\text{LR}(\rho = \rho_0) = TW_T + o_P(1), \tag{II.16}$$

where the term $o_P(1)$ converges to zero in probability, while the first term converges in distribution to a χ_1^2 distribution. And the result follows. The validity of the stochastic Taylor expansion in (II.16) is derived in the appendix. □

II.3 Extending the AR(1) Model[*Can be skipped*]

To allow for a non-zero level in x_t, as is often needed in empirical applications, consider therefore the model given by

$$x_t = \rho x_{t-1} + \mu + \varepsilon_t \quad \text{for } t = 1, 2, ..., T \tag{II.17}$$

where $\rho, \mu \in \mathbb{R}$, x_0 is fixed and ε_t is i.i.d.N$(0,\sigma^2)$. With $|\rho| < 1$ straightforward calculations give,

$$x_t = \rho^t\left(x_0 - \frac{\mu}{1-\rho}\right) + \frac{\mu}{1-\rho} + \sum_{i=0}^{t-1}\rho^i\varepsilon_{t-i},$$

with stationary solution,

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

Note that $\mathbb{E}\left[x_t^*\right] = \frac{\mu}{1-\rho}$, while the variance and covariances are identical to the AR(1) process without μ. Thus by including the level parameter μ, or a constant regressor, in the model a non-zero level of x_t is allowed for, while preserving the correlation structure.

Statistical analysis, in particular estimation of the parameters in (II.17), is most easily addressed by rewriting it as the linear regression model given by

$$Y_t = \beta' Z_t + \varepsilon_t$$

with $Y_t = x_t$, $Z_t = (x_{t-1}, 1)'$ and $\beta' = (\rho, \mu)$. This way it is a special case of the general linear regression model considered in the next section from which the following results can be derived directly:

Theorem II.3.1 *The ML estimators of the AR(1) model in (II.17) are given by*

$$\begin{aligned}(\hat{\rho}, \hat{\mu}) &= S_{yz} S_{zz}^{-1} \\ \hat{\sigma}^2 &= \frac{1}{T}\sum_{t=1}^{T}\left(x_t - (\hat{\rho}, \hat{\mu})\, Z_t\right)^2 \\ &= S_{yy\cdot z} = S_{yy} - S_{yz} S_{zz}^{-1} S_{yz}\end{aligned} \tag{II.18}$$

where with $Y_t = x_t$ and $Z_t = (x_{t-1}, 1)'$, the product moments are given by $S_{ij} = \frac{1}{T}\sum_{t=1}^{T} i_t j_t'$, with $i, j = Y, Z$. The maximized likelihood function is apart from a constant factor given by:

$$L_{\max}\left(\hat{\rho}, \hat{\sigma}^2\right) = \left(\hat{\sigma}^2\right)^{-T/2}.$$

Note that simple linear algebra shows that (II.18) reduces to

$$\hat{\rho} = \frac{\sum_{t=1}^{T}(x_{t-1} - \bar{x}_{-1})x_t}{\sum_{t=1}^{T}(x_{t-1} - \bar{x}_{-1})^2}, \qquad \hat{\mu} = x - \hat{\rho} x_{-1}$$

where $\bar{x} = \frac{1}{T}\sum_{t=1}^{T} x_t$ and $\bar{x}_{-1} = \frac{1}{T}\sum_{t=1}^{T} x_{t-1}$. That is, by the inclusion of the constant term, x_t is corrected for its empirical mean x in the statistical calculations, thereby reducing the impact of the initial value x_0 in the estimation of the parameters.

Again using the results for the general regression model in the next section, the asymptotic properties can be stated as:

Theorem II.3.2 *If $|\rho| < 1$ then $\hat{\rho}, \hat{\mu}$ and $\hat{\sigma}^2$ are consistent. Also as $T \to \infty$,*

$$\sqrt{T}(\hat{\rho} - \rho_0, \hat{\mu} - \mu_0)' \xrightarrow{D} N_2(0, \begin{pmatrix} 1 - \rho^2 & -\mu\,(1+\rho) \\ -\mu\,(1+\rho) & \sigma^2 + \mu^2 \frac{1+\rho}{1-\rho} \end{pmatrix}).$$

Moreover, the likelihood ratio statistic for the hypothesis $\rho = \rho_0$ $(|\rho_0| < 1)$, $\mu = \mu_0$ *or* $\sigma^2 = \sigma_0^2$ *is asymptotically distributed as* χ^2 *with 1 degree of freedom,* χ_1^2.

The AR(1) model may of course be extended in all possible directions by inclusion of various deterministic terms d_t. With d_t n-dimensional, the models may in general be formulated as:

$$x_t = \rho x_{t-1} + \theta' d_t + \varepsilon_t, \quad t = 1, ..., T \tag{II.19}$$

with $\rho \in \mathbb{R}$, $\theta \in \mathbb{R}^n$, x_0 fixed and ε_t i.i.d$N(0, \sigma^2)$. The statistical analysis can be dealt with as in the case before with $d_t = 1$ and $\theta = \mu$ by reformulating it as a special case of the linear regression model in the next section. But it is important to stress that the properties of x_t as a stochastic process depend on the form of $d_t's$ included in the model. This influences the interpretation of the model and also verification of assumptions for asymptotic inference.

Key examples of d_t include a linear trend, $d_t = t$, and deterministic breaks, where for example, $d_t = 1(t \geq T_0)$, with $T_0 < T$ some known point in the sample. The latter is often combined with the constant in order to allow for a shift in the mean by setting

$$\theta' d_t = \mu + \mu_s 1(t \geq T_0) = (\mu, \mu_s)(1, 1(t \geq T_0))'.$$

Deterministic seasonal variation is often modelled by the inclusion of so called seasonal dummies. With quarterly data for example it is reasonable to assume that the mean for each quarter is at a different level. This can be formulated by means of $d_t = (d_{1t}, ..., d_{4t})'$, where

$$d_{1t} = \begin{cases} 1 & \text{for } t = 1, 5, 9, .. \\ 0 & \text{otherwise} \end{cases}, \quad d_{2t} = \begin{cases} 1 & \text{for } t = 2, 6, 10, .. \\ 0 & \text{otherwise} \end{cases}$$

and so forth. As $d_{1t} + ... + d_{4t} = 1$, the dummies are linearly dependent of the constant function and it is convenient to introduce only three dummies and the constant in order to avoid a singular matrix in the estimation of the coefficients. The model is thus modified to

$$x_t = \rho x_{t-1} + \mu + \mu_1 d_{1t} + \mu_2 d_{2t} + \mu_3 d_{3t} + \varepsilon_t.$$

In line with the note above on stochastic properties of x_t, note that the values $\mu + \mu_i$ have the interpretation as the expected value of a 'surprise' in the i'th quarter, but the mean of the process is something entirely different. It follows from the representation of the solution,

$$x_t^* = \sum_{i=0}^{\infty} \rho^i (\mu + \mu_1 d_{1t-i} + \mu_2 d_{2t-i} + \mu_3 d_{3t-i} + \varepsilon_{t-i}),$$

that the mean of the process is determined by,

$$E(x_t^*) = \frac{\mu}{1-\rho} + \sum_{i=0}^{\infty} \rho^i (\mu_1 d_{1t-i} + \mu_2 d_{2t-i} + \mu_3 d_{3t-i}).$$

This shows that the mean is periodic with period 4 and given by combinations of all parameters. Thus strictly speaking the process is no longer stationary. However, the asymptotic inference remains the same as for the case of geometrically ergodic processes.

II.4 A Linear Regression Model

As discussed in the previous section extensions of the AR(1) model can be discussed by means of a linear regression model discussed here. The form and presentation means that the theory discussed here can also be applied directly to the general class of VAR(k) models.

Consider the p-dimensional VAR(k) model as given by

$$X_t = A_1 X_{t-1} + ... + A_k X_{t-k} + \varepsilon_t, \qquad t = 1, 2, ..., T \tag{II.20}$$

with $A_i \in \mathbb{R}^{p\times p}$, initial values $X_0, ..., X_{-k+1}$ fixed and ε_t i.i.d. $\mathrm{N}(0, \Omega)$.

Defining $Y_t = X_t$, $Z_t = (X'_{t-1}, ..., X'_{t-k})'$, $q = pk$, and $\beta' = (A_1, ..., A_k)$ the VAR(k) model is a special case of the linear regression model with stochastic regressor Z_t, as given by:

$$Y_t = \beta' Z_t + \varepsilon_t, \quad t = 1, 2, ..., T \tag{II.21}$$

with Y_t p-dimensional, Z_t q-dimensional, Z_1 fixed, $\beta \in \mathbb{R}^{q\times p}$ and ε_t i.i.d. $\mathrm{N}(0, \Omega)$.

The model in (II.21) is only partial in the sense that there is no model for the stochastic regressor Z_t (unless Z_t is the lagged values of the endogenous X_t as for the VAR model). Instead alone the conditional distribution of Y_t given Z_t is specified as Gaussian with density,

$$f(Y_t|Z_t) = \left(\sqrt{2\pi}\right)^{-p} [\det(\Omega)]^{-1/2} \exp\left(-\frac{1}{2}(Y_t - \beta' Z_t)' \Omega^{-1} (Y_t - \beta' Z_t)\right).$$

This leads to a (partial) log-likelihood function given by,

$$\log L(\beta, \Omega) = \log\left(\prod_{t=1}^{T} f(Y_t|Z_t)\right) \tag{II.22}$$

$$= -\frac{Tp}{2}\log(2\pi) - \frac{T}{2}\log\det(\Omega) - \frac{1}{2}\sum_{t=1}^{T}(Y_t - \beta' Z_t)' \Omega^{-1} (Y_t - \beta' Z_t)$$

corresponding to the successive conditioning of Y_t on Z_t. Clearly information in terms of the non-modelled joint distribution of $(Y_t)_{t=1,2,....,T}$ and $(Z_t)_{t=1,2,....,T}$ is neglected this way. How much, and in what sense precisely, is discussed in the literature on "partial systems" and "weak exogeneity".

In other words, the linear regression model in (II.21) is in fact nothing but a convenient way of specifying the likelihood function in (II.22), which for autoregressive models is the full likelihood. The properties of the estimators $\hat{\beta}$ and $\hat{\Omega}$ which maximize the likelihood will be studied in a general way in order to emphasize sufficient conditions for consistency and asymptotic normality, and χ^2 distributed test statistics. This will in particular allow for departures from normality of the ε_t's as can be relevant in applications. Conditions stated below on Z_t are shown to hold for autoregressive models, but in general these can only be verified by also including a model for Z_t – hence the insistence on the terminology 'partial'.

II.4.1 Estimation

Theorem II.4.1 *Consider the linear regression model in (II.21). The estimators $\hat{\beta}$ and $\hat{\Omega}$ which maximize the partial likelihood in (II.22) are given by,*

$$\hat{\beta}' = S_{yz}S_{zz}^{-1} \tag{II.23}$$

$$\begin{aligned}\hat{\Omega} &= \frac{1}{T}\sum_{t=1}^{T}\left(Y_t - \hat{\beta}'Z_t\right)\left(Y_t - \hat{\beta}'Z_t\right)' \\ &= S_{yy\cdot z} = S_{yy} - S_{yz}S_{zz}^{-1}S_{yz}\end{aligned} \tag{II.24}$$

where the product moment matrices are given by $S_{ij} = \frac{1}{T}\sum_{t=1}^{T} i_t j_t'$, with $i, j = Y, Z$. The maximized likelihood function is given by,

$$L_{\max}\left(\hat{\beta}, \hat{\Omega}\right) = c\left[\det(\hat{\Omega})\right]^{-T/2}, \tag{II.25}$$

where the constanct factor $c = (2\pi e)^{-Tp/2}$.

The estimators $\hat{\beta}$ and $\hat{\Omega}$ are commonly referred to as OLS estimators, where OLS stands for ordinary least squares.

Proof of Theorem II.4.1:

In terms of differentials, it follows that in the direction β and Ω respectively, with $\varepsilon_t(\beta) \equiv Y_t - \beta' Z_t$,

$$d\log L(\beta, \Omega; d\beta) = \operatorname{tr}\left\{\Omega^{-1}\sum_{t=1}^{T}\varepsilon_t(\beta) Z_t' d\beta\right\},$$

$$d\log L(\beta, \Omega; d\Omega) = -\frac{T}{2}\operatorname{tr}\left\{\Omega^{-1}d\Omega\right\} + \frac{1}{2}\operatorname{tr}\left\{\Omega^{-1}\sum_{t=1}^{T}\varepsilon_t(\beta)\varepsilon_t(\beta)'\Omega^{-1}d\Omega\right\}.$$

Here it has been used that $\operatorname{tr}\{AB\} = \operatorname{tr}\{BA\}$, and with $f(B) = \operatorname{tr}\{AB\}$, $g(B) = \log\det(B)$, and $h(B) = \operatorname{tr}\{AB^{-1}\}$, then the differentials are given by, $df(B; dB) = \operatorname{tr}\{AdB\}$, $dg(B; dB) = \operatorname{tr}\{B^{-1}dB\}$, and finally $dh(B; dB) = -\operatorname{tr}\{AB^{-1}(dB)B^{-1}\}$, where A and B are appropriate matrices; see the Appendix for details and references.

The first order condition for β is given by $d\log L(\beta, \Omega; d\beta) = 0$, for all $d\beta$, or

$$\sum_{t=1}^{T}\varepsilon_t(\beta) Z_t' = \sum_{t=1}^{T}(Y_t - \beta' Z_t) Z_t' = 0.$$

This gives $\hat{\beta}' = S_{yz}S_{zz}^{-1}$. Likewise, the first order condition for Ω is given by

$$\Omega^{-1} = \Omega^{-1}\frac{1}{T}\sum_{t=1}^{T}\varepsilon_t(\beta)\varepsilon_t(\beta)'\Omega^{-1},$$

and hence $\hat{\Omega} = \hat{\Omega}(\hat{\beta})$, where $\hat{\Omega}(\beta) = \frac{1}{T}\sum_{t=1}^{T}\varepsilon_t(\beta)\varepsilon_t(\beta)'$. Computation of the second order differentials shows that $L(\beta, \Omega)$ has a maximum in $\left(\hat{\beta}, \hat{\Omega}\right)$. □

II.4.2 Asymptotics

To derive the asympotic properties of $\hat{\beta}$ and $\hat{\Omega}$ some assumptions are needed specifically for Z_t. In order to allow for the possibility that ε_t are not i.i.d. Gaussian this condition is relaxed in the assumptions. If ε_t are not i.i.d. Gaussian the estimators are usually referred to as quasi maximum likelihood (QML) estimators as the likelihood function in (II.22) used for maximization is then not correctly specified, even as a partial likelihood.

Before stating the assumptions note the so-called Cramer-Wold device by which univariate CLTs can be applied to vectors. Thus, with $X, X_T \in \mathbb{R}^p$, then the Cramer-Wold device states that $X_T \overset{D}{\to} X$, as $T \to \infty$, if and only if, for any $\lambda \in \mathbb{R}^p$, $\lambda \neq 0$,

$$\lambda' X_T \overset{D}{\to} \lambda' X.$$

Below the same device is applied to state a CLT result for sequences of matrices rather than sequences of vectors. To do so we introduce here some results for operations with matrices, see Magnus and Neudecker (2007) for further details as well as the appendix here.

II.4.3 Matrix calculus | A short list of useful identities

With $M = (M_{ij})_{i=1,...,p,j=1,...,q} \in \mathbb{R}^{p\times q}$, let

$$\text{vec}\,(M) = (M_{11}, M_{21}, ..., M_{p1}, ..., M_{pq})'$$

that is, vec (M) is the vector obtained by stacking the columns of the matrix M. In terms of the vec $(\cdot)$ and tr $(\cdot)$, we have for $N \in \mathbb{R}^{p\times q}$, the identity

$$\text{vec}\,(N)'\,\text{vec}\,(M) = \text{tr}\,(N'M)\,. \tag{II.26}$$

Moreover, with $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{p\times q}$, introduce the $\otimes$-product (Kronecker product) with $A \otimes B \in \mathbb{R}^{mp\times nq}$, where

$$A \otimes B = \begin{pmatrix} A_{11}B & & A_{1n}B \\ & \ddots & \\ A_{m1}B & & A_{mn}B \end{pmatrix},$$

and $(A \otimes B)(C \otimes D) = (AC \otimes BD)$.

We mention here two key identities

$$\text{vec}\,(ABC) = (C' \otimes A)\,\text{vec}\,(B) \tag{II.27}$$

$$\text{tr}\,(ABCD)) = \text{vec}\,(D')'\,(C' \otimes A)\,\text{vec}\,(B) \tag{II.28}$$

In terms of matrices, the matrix $M \in \mathbb{R}^{p\times q}$ is Gaussian distruted with mean zero and $(qp \times qp)$-dimensional covariance matrix $\Omega, \Omega > 0$, if vec (M) is (vector) $\text{N}(0, \Omega)$ distributed. Often the covariance Ω has a so-called Kronecker-product structure of the form

$$\Omega = (\Sigma_{qq} \otimes \Sigma_{pp})\,, \tag{II.29}$$

where $\Sigma_{qq} > 0$ is $(q \times q)$-dimensional, and $\Sigma_{pp} > 0$ is $(p \times p)$-dimensional. With Ω given by (II.29), it follows that e.g. $M\Sigma_{qq}^{-1}$ is $\text{N}\left(0, \left(\Sigma_{qq}^{-1} \otimes \Sigma_{pp}\right)\right)$ distributed. Thus by applying (II.27)

$$\begin{aligned} \text{vec}\left(M\Sigma_{qq}^{-1}\right) &= \left(\Sigma_{qq}^{-1} \otimes I_p\right)\text{vec}\,(M) \\ &= \left(\Sigma_{qq}^{-1} \otimes I_p\right)\text{N}(0, (\Sigma_{qq} \otimes \Sigma_{pp})) \\ &= \text{N}\big(0, \left(\Sigma_{qq}^{-1} \otimes \Sigma_{pp}\right)\big) \end{aligned}$$

by standard properties of the Gaussian distribution since

$$\left(\Sigma_{qq}^{-1} \otimes I_p\right)\left(\Sigma_{qq} \otimes \Sigma_{pp}\right)\left(\Sigma_{qq}^{-1} \otimes I_p\right) = \left(\Sigma_{qq}^{-1} \otimes \Sigma_{pp}\right).$$

We are now in position to state the following assumption.

Assumption II.4.1 *Consider the OLS estimators in Theorem II.4.1 specified as functions of the variables Z_t and ε_t, where $\varepsilon_t = Y_t - \beta' Z_t$. Make the following assumptions:*

(OLS.1) As $T \to \infty$, a LLN applies to the product moment matrix $S_{vv} = \frac{1}{T}\sum_{t=1}^{T} v_t v_t'$ of $v_t \equiv (\varepsilon_t', Z_t')' \in \mathbb{R}^{p+q}$,

$$S_{vv} = \begin{pmatrix} S_{\varepsilon\varepsilon} & S_{\varepsilon z} \\ S_{z\varepsilon} & S_{zz} \end{pmatrix} \overset{P}{\to} \begin{pmatrix} \Sigma_{\varepsilon\varepsilon} & \Sigma_{\varepsilon z} \\ \Sigma_{z\varepsilon} & \Sigma_{zz} \end{pmatrix} > 0. \tag{II.30}$$

(OLS.2) With $\varepsilon_t Z_t' \in \mathbb{R}^{p \times q}$, the process $\{\varepsilon_t Z_t'\}_{t=1,2,\ldots,T}$ is a martingale difference sequence with respect to the filtration $\mathcal{F}_t = (\varepsilon_t, Z_{t+1}, \varepsilon_{t-1}, Z_t, \ldots)$, such that

$$\mathbb{E}\left[\varepsilon_t | \mathcal{F}_{t-1}\right] = 0, \tag{II.31}$$

and $\Sigma_{\varepsilon z} = 0$. Moreover, for any $V \in \mathbb{R}^{p \times q}$, $V \neq 0$, then

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\left(\operatorname{tr}(V'\varepsilon_t Z_t')\right)^2 | \mathcal{F}_{t-1}\right] \overset{P}{\to} \tag{II.32}$$

$$\operatorname{tr}\left(\Sigma_{\varepsilon\varepsilon} V \Sigma_{zz} V'\right) = \operatorname{vec}(V)'\left(\Sigma_{zz} \otimes \Sigma_{\varepsilon\varepsilon}\right)\operatorname{vec}(V) > 0.$$

And for any $\delta > 0$, as $T \to \infty$,

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\left(\operatorname{tr}(V'\varepsilon_t Z_t')\right)^2 \mathbb{I}\left(|\operatorname{tr}(V'\varepsilon_t Z_t')| < \delta\sqrt{T}\right) | \mathcal{F}_{t-1}\right] \overset{P}{\to} 0 \tag{II.33}$$

Assumption II.4.1 states that a CLT and LLN apply as was applied in the proof of Theorem II.2.2. Specifically in the AR(1) case with $Y_t = x_t$, $Z_t = x_{t-1}$, $\varepsilon_t = x_t - \rho x_{t-1}$, it was used that x_t was geometrically ergodic. This implied in particular that the LLN applied and (OLS.1-2) followed with $\Sigma_{\varepsilon\varepsilon} = \sigma^2$ and $\Sigma_{zz} = \sigma^2/(1-\rho^2)$ in the AR(1) case.

In terms of the introduced matrix notation, (OLS.1-2) imply that

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} \operatorname{vec}\left(\varepsilon_t Z_t'\right) \overset{D}{\to} \mathrm{N}(0, \Sigma_{zz} \otimes \Sigma_{\varepsilon\varepsilon}).$$

To see this note that the results imply by the CLT, that

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\text{vec}\,(V)'\,\text{vec}\,(\varepsilon_t Z_t') = \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\text{tr}\,(V'\varepsilon_t Z_t) \xrightarrow{D} \text{N}\,(0, \text{tr}\,(\Sigma_{\varepsilon\varepsilon} V \Sigma_{zz} V'))\,.$$

Next, by (II.28),

$$\text{tr}\,(\Sigma_{\varepsilon\varepsilon} V \Sigma_{zz} V') = \text{vec}\,(V)'\,(\Sigma_{zz}\otimes\Sigma_{\varepsilon\varepsilon})\,\text{vec}\,(V)$$

and we have

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\text{vec}\,(V)'\,\text{vec}\,(\varepsilon_t Z_t') \xrightarrow{D} \text{N}\left(0, \text{vec}\,(V)'\,(\Sigma_{zz}\otimes\Sigma_{\varepsilon\varepsilon})\,\text{vec}\,(V)\right).$$

And by the Cramer-Wold device, we get

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\text{vec}\,(\varepsilon_t Z_t') \xrightarrow{D} \text{N}\,(0, (\Sigma_{zz}\otimes\Sigma_{\varepsilon\varepsilon}))\,,$$

that is, $\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\varepsilon_t Z_t'$ is asymptotically (matrix) Gaussian distributed with covariance $\Omega = (\Sigma_{zz}\otimes\Sigma_{\varepsilon\varepsilon})$.

Assumption II.4.1 has the immediate consequence:

Theorem II.4.2 *Under Assumption II.4.1, then $\hat{\beta} \xrightarrow{P} \beta$ and $\hat{\Omega} \xrightarrow{P} \Sigma_{\varepsilon\varepsilon}$ where the OLS estimators are defined in (II.23) and (II.24) respectively in Theorem II.4.1. Moreover,*

$$\sqrt{T}\left(\hat{\beta}-\beta\right)' \xrightarrow{D} \text{N}\left(0, \Sigma_{zz}^{-1}\otimes\Sigma_{\varepsilon\varepsilon}\right) \tag{II.34}$$

with Σ_{zz} consistently estimated by S_{zz}.

Proof of Theorem II.4.2:

Turn first to consistency. By definition of $\hat{\beta}$ in (II.23), $\left(\hat{\beta}-\beta\right)' = S_{\varepsilon z}S_{zz}^{-1}$, where by (OLS.1), $S_{zz} \xrightarrow{P} \Sigma_{zz}$ and, $S_{\varepsilon z} \xrightarrow{P} 0$, by the martingale difference assumption in (OLS.2). Likewise,

$$\hat{\Omega} = \frac{1}{T}\sum_{t=1}^{T}\left(Y_t - \hat{\beta}'Z_t\right)\left(Y_t - \hat{\beta}'Z_t\right)' = S_{\varepsilon\varepsilon} + \left(\hat{\beta}-\beta\right)' S_{zz}\left(\hat{\beta}-\beta\right) \xrightarrow{P} \Sigma_{\varepsilon\varepsilon}.$$

Next for the asymptotic distribution, (OLS.2) implies by direct application of the CLT for martingale differences that, as $T \to \infty$,

$$\sqrt{T} S_{\varepsilon z} = \frac{1}{T} \sum_{t=1}^{T} \varepsilon_t Z_t' \xrightarrow{D} \mathrm{N}(0, \Sigma_{zz} \otimes \Sigma_{\varepsilon\varepsilon}) .$$

And hence, using the considerations above

$$\sqrt{T} \left(\hat{\beta} - \beta \right)' \xrightarrow{D} \mathrm{N}(0, \Sigma_{\varepsilon\varepsilon} \otimes \Sigma_{zz}) \, \Sigma_{zz}^{-1} \overset{D}{=} \mathrm{N}\big(0, \Sigma_{zz}^{-1} \otimes \Sigma_{\varepsilon\varepsilon}\big)$$

as desired. □

Many versions of the kind of assumptions in Assumption II.4.1 exist in the literature, some more general than others. But basically they all, as in (OLS.1), imply that a LLN apply to the sample product moment there. This is in particular implied if the LLN for mixing or for asymptotically stable processes apply as can be used for the VAR(k) processes.

It is worthwhile to comment a little further on the conditions:

If (OLS.1) applies, (OLS.2) holds in particular if ε_t is i.i.d.$(0, \Sigma_{\varepsilon\varepsilon})$ and independent of the regressor Z_t and the past variables in $\mathcal{F}_{t-1}$ as in the classical regression set-up, and as in the formulation of the autoregressive models.

On the other hand, the martingale difference assumption in (OLS.2) rules out that ε_t is correlated with the regressor Z_t. Consider for example the case where a lagged Z_t was omitted in the OLS estimation, that is $\varepsilon_t = \theta Z_{t-1} + \eta_t$, with η_t i.i.d.$(0, \Sigma_{\eta\eta})$ and η_t a martingale difference satisfying (OLS.2). Then, $E(\varepsilon_t Z_t') = \theta E(Z_{t-1} Z_t') \neq 0$, and $\hat{\beta}$ would be inconsistent. In general this would also be the case if ε_t was an MA or AR type process, and in empirical work much attention is therefore devoted to make sure that no autocorrelation appears in the residuals ε_t as measured by the empirical residuals,

$$\varepsilon_t \equiv Y_t - \hat{\beta}' Z_t.$$

The variance specification of the limiting Gaussian distribution of $\hat{\beta}$ in (II.34), comes directly from the condition (II.32) in (OLS.2). More general structures can be allowed for by requiring that the average,

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E} \left[\left(\mathrm{tr}(V' \varepsilon_t Z_t') \right)^2 | \mathcal{F}_{t-1} \right] ,$$

converges to a term which does not factorize as in (II.32). Consider for the univariate ($p = q = 1$) case, the example where ε_t is an autoregressive conditional heteroscedastic (ARCH) process as given by,

$$\varepsilon_t = \left(\sqrt{1+\alpha\varepsilon_{t-1}^2}\right)\eta_t, \qquad \eta_t \text{ i.i.d}N(0,1) \tag{II.35}$$

It follows that ε_t is a martingale difference, and ε_t is uncorrelated with Z_t but not independent as $E(\varepsilon_t^2|\mathcal{F}_{t-1}) = 1+\alpha\varepsilon_{t-1}^2$. Hence consistency holds for $\hat{\beta}$, but as,

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\varepsilon_t^2 Z_t^2|\mathcal{F}_{t-1}\right] = \frac{1}{T}\sum_{t=1}^{T}Z_t^2 + \alpha\frac{1}{T}\sum_{t=1}^{T}\varepsilon_{t-1}^2 Z_t^2.$$

then provided ε_t and Z_t have suitable (fourth order) moments, this will converge in probability to $\Sigma_{zz}+\theta_z$, $\theta_z\neq 0$. Hence a different variance would appear in (II.34).

To summarize: The partial likelihood function in (II.22) was defined in order to see how the OLS estimators in general can be found by optimization. The conditions in Assumption II.4.1 point at what conditions are sufficient for consistency and asymptotic normality. Hence if the likelihood function is changed, other estimators appear, and if Assumption II.4.1 is replaced by another version, other asymptotic results may hold. Which must be derived for each case, and the preceeding discussion and presentation demonstrate the type of considerations needed.

II.4.4 Hypothesis Testing:

Consider the general linear hypothesis on $\beta\in\mathbb{R}^{q\times p}$ as given by:

$$H_{\text{lin}} : \beta = H\varphi \tag{II.36}$$

where H is some known $q\times s$ dimensional matrix, $s\leq q$, and $\varphi\in\mathbb{R}^{s\times p}$ the freely varying parameters under H_{lin}. Note that H_{lin} may equivalently be written as

$$H_{\text{lin}} : R'\beta = 0, \tag{II.37}$$

where R is $q\times r$, where $r=q-s$ is 'the number of restrictions in each equation', and $R'H=0$ such that the matrix (H,R) has full rank, that is $R=H_\perp$.

Central examples of the linear hypothesis include omission of variables and the hypothesis that only a few linear combinations of Z_t effect Y_t. For example, with $p=2$ and $q=3$ such that $Y_t=(Y_{1t},Y_{2t})'$ and $Z_t=$

$(Z_{1t}, Z_{2t}, Z_{3t})'$, consider first the hypothesis that Z_{3t} can be omitted. This can be written as

$$\beta = H\varphi = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \varphi_{11} & \varphi_{12} \\ \varphi_{21} & \varphi_{22} \end{pmatrix}$$

or simply, $R'\beta = (0,0,1)\,\beta = 0$. Likewise the hypothesis that only the 'spread' between Z_{1t} and Z_{2t}, $Z_{1t} - Z_{2t}$, appears as regressor can be stated as

$$\beta = H\varphi = \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix} \begin{pmatrix} \varphi_{11} & \varphi_{12} \end{pmatrix}.$$

In terms of the likelihood in (II.22), the likelihood ratio statistic of the hypothesis H_{lin} , $\text{LR}(H_{\text{lin}})$ and its asymptotic distribution is stated in the next theorem which generalizes Theorem II.2.3:

Theorem II.4.3 *Consider the linear regression model in (II.21) under the hypothesis $H_{lin} : \beta = H\varphi$, where H is $(q \times s)$-dimensional and known. The estimators which maximize the partial likelihood in (II.22) are given by $\tilde{\beta} = H\tilde{\varphi}$ and $\tilde{\Omega}$ with*

$$\tilde{\varphi}' = S_{yz}H\left(H'S_{zz}H\right)^{-1} \tag{II.38}$$

$$\tilde{\Omega} = \frac{1}{T}\sum_{t=1}^{T}\left(Y_t - \tilde{\beta}'Z_t\right)\left(Y_t - \tilde{\beta}'Z_t\right)' \tag{II.39}$$

$$= S_{yy} - S_{yz}H\left(H'S_{zz}H\right)^{-1}H'S_{zy}$$

where the product moment matrices are given by $S_{ij} = \frac{1}{T}\sum_{t=1}^{T} i_t j_t'$, with $i, j = Y, Z$. The maximized likelihood function is apart from a constant factor given by, $L_{\max}(\tilde{\beta}, \tilde{\Omega}) = |\tilde{\Omega}|^{-T/2}$, and hence the LR statistic of H_{lin} is given by

$$LR\left(H_{lin}\right) = T\log\det\left(I_p + W_T\right), \qquad W_T = \hat{\Omega}^{-1}\left(\hat{\beta} - \tilde{\beta}\right)' S_{zz}\left(\hat{\beta} - \tilde{\beta}\right). \tag{II.40}$$

Under Assumption II.4.1, the LR statistic is asymptotically χ^2 distributed with rp degrees of freedom, χ^2_{rp}, where $r = q - s$.

Similar to the remarks made in connection to Theorem II.2.3, note that the term

$$\text{tr}\left\{TW_T\right\} \equiv W \tag{II.41}$$

is known as the Wald statistic W for the hypothesis that $\beta = H\varphi$ which, by the proof of Theorem II.4.3, is also asymptotically χ^2_{rp}.

For the univariate case where $p = 1$, the Wald statistic $W = TW_T$ and the classic F statistic well-known from regression analysis are related by

$$F = \frac{(T-q)}{(q-s)} W_T. \tag{II.42}$$

Of course the F statistic is not $F\,(q-s, T-q)$ distributed as it would be for the case of fixed deterministic regressors, instead $rF = (q-s)\,F$ is asymptotically χ^2_{rp}. This observation is useful when interpreting empirical output where F statistics are often reported also for time series even though the F distribution is not usually adequate.

Sometimes, the form

$$W = T\hat{\Omega}^{-1}\hat{\beta}' R(R' S_{zz}^{-1} R)^{-1} R'\hat{\beta} \tag{II.43}$$

is preferred for the Wald statistic W (and similarly for the F statistic). That is, the form emphasizing the restrictions as given by R. To see the equivalence, recall the identities,

$$I_q = R\,(R'R)^{-1}\,R' + H\,(H'H)^{-1}\,H' \tag{II.44}$$

$$I_q = R\left(R'\Sigma^{-1}R\right)^{-1} R'\Sigma^{-1} + \Sigma H\,(H'\Sigma H)^{-1}\,H', \tag{II.45}$$

corresponding to orthogonal and skew projections respectively, here in terms of R, H and any $q \times q$ dimensional positive definite matrix $\Sigma > 0$.

Note likewise that in the W_T term the residual covariance is estimated under the alternative by $\hat{\Omega}$, but alternatively the LR statistic may be stated as,

$$\mathrm{LR}\,(H_{\mathrm{lin}}) = -T \log\det\left(I_p - \tilde{W}_T\right)$$

with $\tilde{W}_T \equiv \tilde{\Omega}^{-1}\left(\hat{\beta} - \tilde{\beta}\right)' S_{zz}\left(\hat{\beta} - \tilde{\beta}\right)$.

Thus there are many equivalent ways of representing the test statistic for the linear hypothesis.

Proof of Theorem II.4.3: The expressions for $\tilde{\beta}, \tilde{\varphi}$ and $\tilde{\Omega}$ in (II.38) and (II.39) follow immediately by Theorem II.4.1 by using that under H_{lin} the regresssion model can be written as:

$$Y_t = \varphi'\,(H'Z_t) + \varepsilon_t.$$

And the likelihood ratio statistic $\mathrm{LR}(H_{\mathrm{lin}}) = -2\log Q$, where

$$Q^{-2/T} = \det(\tilde{\Omega})(\det(\hat{\Omega}))^{-1} = \det(\hat{\Omega}^{-1}\tilde{\Omega}).$$

Using the decomposition from regression analysis,

$$Y_t - \tilde{\beta}'Z_t = (Y_t - \hat{\beta}'Z) + (\hat{\beta} - \tilde{\beta})'Z_t,$$

this leads to the key identity,

$$T\tilde{\Omega} = T\hat{\Omega} + T(\hat{\beta} - \tilde{\beta})'S_{zz}(\hat{\beta} - \tilde{\beta}).$$

In particular, this establishes (II.40) as,

$$Q^{-2/T} = \det(\hat{\Omega}^{-1}\tilde{\Omega}) = \det(I_p + \hat{\Omega}^{-1}(\hat{\beta} - \tilde{\beta})'S_{zz}(\hat{\beta} - \tilde{\beta})) = \det(I_p + W_T). \tag{II.46}$$

For the asymptotic distribution note first that by Theorem II.4.2, $\tilde{\beta}$, $\hat{\beta}$ and $\hat{\Omega}$ are consistent, and hence $W_T \overset{P}{\to} 0$. Moreover, as will be argued below,

$$\operatorname{tr}\{TW_T\} \overset{D}{\to} \chi^2_{(q-s)p}. \tag{II.47}$$

A Taylor expansion, see appendix, of $f(W) = \log\det(I_p + W)$ for $W \to 0$, with $W \in \mathbb{R}^{p\times p}$, gives $f(W) = \operatorname{tr}\{W\} + o(\|W\|)$, and hence

$$\operatorname{LR}(H_{\text{lin}}) = \operatorname{tr}\{TW_T\} + o_P(1), \tag{II.48}$$

where the term $o_P(1)$ converges to zero in probability, while the first term converges in distribution to the $\chi^2_{(q-s)p}$ distribution. And the result follows.

To see (II.47), note first that by definition

$$\left(\hat{\beta} - \tilde{\beta}\right)' = \left(\hat{\beta} - \beta\right)' - \left(\tilde{\beta} - \beta\right)' = S_{\varepsilon z}S_{zz}^{-1} - S_{\varepsilon z}H\left(H'S_{zz}H\right)^{-1}H'$$

As in the proof of Theorem II.4.2 $\sqrt{T}S_{\varepsilon z} \overset{D}{\to} N_{p\times q}\left(0, \Sigma_{zz} \otimes \Sigma_{\varepsilon\varepsilon}\right)$, while

$$S_{zz}^{-1} - H\left(H'S_{zz}H\right)^{-1} \overset{P}{\to} \Sigma_{zz}^{-1} - H\left(H'\Sigma_{zz}H\right)^{-1}H = \Sigma_{zz}^{-1}R\left(R'\Sigma_{zz}^{-1}R\right)^{-1}R'\Sigma_{zz}^{-1},$$

where the last equality holds by using the skew-projection in (II.45).

Note that, $\operatorname{tr}\{TW_T\} = \operatorname{tr}\{V_TV_T'\}$, where

$$V_T = \hat{\Omega}^{-1/2}\sqrt{T}(\hat{\beta} - \tilde{\beta})'S_{zz}^{1/2} \overset{D}{\to} VM,$$

with $V = N_{p\times q}\left(0, I_q \otimes I_p\right)$ and $M = \Sigma_{zz}^{-1/2}R\left(R'\Sigma_{zz}^{-1}R\right)^{-1}R'\Sigma_{zz}^{-1/2}$. As $U \equiv V\Sigma_{zz}^{-1/2}R\left(R'\Sigma_{zz}^{-1}R\right)^{-1/2}$ is $N_{p\times r}\left(0, I_r \otimes I_p\right)$ distributed,

$$\operatorname{tr}\{TW_T\} \overset{D}{\to} \operatorname{tr}\{VMV'\} = \operatorname{tr}\{UU'\} \overset{D}{=} \chi^2_{pr}.$$

and the result follows. □

II.5 The VAR(k) model: Estimation and Asymptotic theory

Recall that the the p-dimensional VAR(k) model is given by

$$X_t = A_1 X_{t-1} + ... + A_k X_{t-k} + \varepsilon_t, \qquad t = 1, 2, ..., T \tag{II.49}$$

with $A_i \in \mathbb{R}^{p \times p}$, initial values $X_0, ..., X_{-k+1}$ fixed and ε_t i.i.d. $\mathrm{N}(0, \Omega)$. And the corresponding characteristic polynomial is given by $A(z) = I_p - A_1 z - ... - A_1 z^k$, $z \in \mathbb{C}$.

An immediate application of Theorems II.4.1, II.4.2, and II.4.3 gives:

Theorem II.5.1 *Consider the VAR(k) model as defined by (II.49) and set $\beta' = (A_1, ..., A_k)$. With $Y_t \equiv x_t$ and $Z_t \equiv \left(x'_{t-1}, ..., x'_{t-k}\right)'$, the maximum likelihood estimators of β and Ω are given by the OLS estimators in Theorem II.4.1. In particular, these maximize the likelihood function for $x_1, ..., x_T$ conditional on Z_1. Moreover, the likelihood ratio statistic of a linear hypothesis on β, $\beta = H\varphi$, with H $q \times s$, is given by (II.40) in Theorem II.4.3.*

If furthermore, $\det(A(z)) = 0$ *implies $|z| > 1$, then the asymptotic distributions in Theorem II.4.2 apply and the likelihood ratio statistic of the linear hypothesis $\beta = H\varphi$ is asymptotically $\chi^2_{p(q-s)}$ distributed.*

Note the distinction between estimation and asymptotic inference: The estimators and the LR test statistics still apply even if x_t does not have the properties needed for the asymptotic distributions. If the assumption regarding the roots of $A(z)$ does not apply the limiting distributions of the estimators and test statistic will however be different as will be explored later.

Note also that by setting $p = 1$, $A_i = \rho_i$ for $i = 1, 2, ..., k$ the theorem also applies to the AR(k) model.

Proof of Theorem II.5.1:

As already noted the partial likelihood in (II.22) is the full likelihood for the VAR(k) model conditional on the initial value, Z_1, and the result on estimation and test statistic hold immediately by Theorems II.4.1 and II.4.3.

That the limit theory results from Theorems II.4.2 and II.4.3 hold, follow by noting that the assumption about the roots of $A(z)$ implies that $Z_t = \left(x'_{t-1}, ..., x'_{t-k}\right)'$ is geometrically ergodic. In particular Assumption II.4.1 holds: (OLS.1) and (OLS.2) hold by the LLN for geometrically ergodic processes as used in the simple case in the proof of Theorem II.6. □

References

[1] Magnus, J. and Neudecker, H. (2007) Matrix Differential Calculus with Applications in Statistics and Econometrics, Third Edition, Wiley.

[2] Mann and Wald (1943), On Stochastic Limit and Order Relationships, *Annals of mathematical Statistics*, 14, 390-402.

[3] Brockwell and Davis (1995), *Time Series: Theory and Methods*, Springer.

Appendix

A Stochastic orders

To *ease* various asymptotic derivations introduce the following notation:

- $O_P(\cdot)$, "big O in probability"
- $o_p(\cdot)$, "small o in probability"

which are the stochastic equivalents of "O" and "o" from ordinary analysis. Excellent references are Mann and Wald, "On Stochastic Limit and Order Relationships", *Annals of mathematical Statistics*, 14, 390-402, and also Brockwell and Davis (1995, Time Series: Theory and Methods, Springer, Ch.6)

A.1 Definition and results

Recall that if x_T is a deterministic sequence then $x_T = o(1)$ if $x_T \to 0$ as $T \to \infty$, and likewise $x_T = O(1)$ if the sequence is bounded. Likewise the concepts of "small o in probability" and "boundedness in probability" are defined by:

Definition A.1

$o_P(\cdot)$: $x_T = o_P(T^\delta)$ *for some* $\delta > 0$, *if* $T^{-\delta} x_T \xrightarrow{P} 0$ *as* $T \to \infty$.

$O_P(\cdot)$: $x_T = O_P(T^\delta)$ *for some* $\delta > 0$, *if for all* $\varepsilon > 0$ *there exist* $c = c(\varepsilon) > 0$ *and* $T^* > 0$ *such that for* $T > T^*$

$$P\left(T^{-\delta}||x_n|| > c\right) < \varepsilon.$$

If $x_T = O_P(1)$ the x_T sequence is often referred to as "tight". Note also that $x_T = o_P(1)$ is identical to $x_T \xrightarrow{P} 0$.

Next some results for how to use these.

Proposition A.1

1. $x_T \xrightarrow{P} c \Rightarrow x_T = O_P(1)$
2. $x_T \xrightarrow{D} x \Rightarrow x_T = O_P(1)$
3. $x_T = o_P(T^\delta) \Rightarrow x_T = O_P(T^\delta)$

4. $x_n = O_p(T^{\delta}) \Rightarrow x_T = o_P(T^{\mu})$ *if* $\mu > \delta$*.*

5. If $x_T = O_P(T^{\delta})$ *and* $Y_T = O_P(T^{\mu})$ *then*

$$x_T + Y_T = O_P(T^{\max(\delta,\mu)}) \text{ and } x_T Y_T = O_P(T^{\delta+\mu})$$

The same results hold for $o_P(\cdot)$*.*

6. If $x_T = O_P(T^{\delta})$ *and* $Y_T = o_P(T^{\mu})$ *then*

$$x_T Y_T = o_P(T^{\delta+\mu})$$

7. $x_T = O_P((E||x_T||^r)^{1/r})$ *for* $r > 0$ *and* $E||x_T||^r < \infty$*.*

B Matrix differentitation

Some notation is needed in order to handle derivatives of functions of matrices, see Magnus and Neudecker (2007) for a general introduction to matrix differential calculus.

Recall that the function $f : \mathbb{R} \to \mathbb{R}$, given by $f(x) = x^2$ is differentiable with differential $df(x; dx) = 2xdx$ as,

$$\begin{aligned} f(x+h) &= f(x) + df(x;h) + o(|x|) \\ &= x^2 + 2xh + o(|h|). \end{aligned}$$

This is simple to show directly, as $(x+h)^2 = x^2 + 2xh + h^2$, and by definition the term h^2 is $o(|h|)$ as $h \to 0$. Often the derivative $f'(x) = 2x$ is simply reported rather than the differential for obvious reasons. It is the opposite for matrix valued and real valued functions of matrices, where it is more convenient to work in terms of differentials.

Consider the matrix valued function f,

$$f : \mathbb{R}^{k \times l} \to \mathbb{R}^{m \times n}$$

where k, l, m and n are integers. Then f is differentiable of order one in $x \in \mathbb{R}^{k \times l}$ with differential df if

$$f(x+H) = f(x) + df(x;H) + o(\|H\|), \tag{II.50}$$

for $H \in \mathbb{R}^{k+l}$ and as $\|H\| \to 0$ with $\|\cdot\|$ some matrix norm. Here, $df(x;H)$ is the differential of f evaluated at x with increment H and is linear in H –

just like in the univariate case above. Similarly, one may define higher orders of differentiability, say order 3, by

$$f(x+H) = f(x) + df(x;H) + d^2f(x;H,H) + d^3f(x;H,H,H) + o\left(\|H\|^3\right)$$

as $\|H\| \to 0$.

The idea behind the notation in (II.50) is as noted to emphasize the differential rather than the derivatives or Jacobian as usually applied in calculus. For example, with $f : \mathbb{R} \to \mathbb{R}$, $f(x) = \log x$, the classic Taylor expansion around $x = 1$,

$$f(x) = \log x = x - 1 + o(\|x-1\|),$$

can be stated as, or derived from,

$$f(x+h) = f(x) + df(x;h) + o(\|h\|),$$

using that the differential in standard notation is given by $df(x;dx) = \frac{1}{x}dx$.

With $f : \mathbb{R}^{k\times k} \to \mathbb{R}$, $f(x) = \log|x|$, the differential will be generalized below to $df(x;dx) = \text{tr}\{x^{-1}dx\}$, where dx a small $k \times k$ matrix, and hence, similar to the univariate case,

$$\log|x| = \text{tr}\{x - I\} + o(\|x-I\|).$$

Below some important differentials are given which provide the necessary results for the matrix calculus.

Note that although it shall *not be* used here, alternatively one can use the *vec*-operator to define differentiability of matrix valued functions of matrices. With x a $k \times l$ matrix and x_j its j'th column the *vec*-operator is defined by

$$\text{vec}(x) = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix}$$

Differentiability may, as often done in econometrics, then be defined by 'treating matrices as vectors' – The Jacobian, well-known from calculus, $\frac{\partial}{\partial\,\text{vec}(x)}\,\text{vec}(f(x))$, and the differential are connected by the identity,

$$\text{vec}(df(x,H)) = \left[\frac{\partial\,\text{vec}(f(x))}{\partial(\text{vec}(x))'}\right]'\text{vec}(H)$$

Likewise for the second order derivative or Hessian.

B.1 Some differentials

From the brief introduction above it will be sufficient to report some differentials. The list here is sufficient for the study of multivariate models herein as well as the later multivariate e.g. cointegration models:

Lemma B.1 *Assume that the matrices x, A are of appropriate dimensions such that the functions are well-defined:*

1. *With $f(x) = \mathrm{tr}\{Ax\}$, then $df(x; dx) = \mathrm{tr}\{Adx\}$.*
2. *With $f(x) = \mathrm{tr}\{x'x\}$, then $df(x; dx) = \mathrm{tr}(dx'x) + \mathrm{tr}(x'dx) = 2\,\mathrm{tr}(x'dx)$.*
3. *With x a square matrix with $|x| > 0$, then:*

$$\begin{aligned}&\text{With } f(x) = |x|, \text{ then } df(x; dx) = |x|\,\mathrm{tr}(x^{-1}dx).\\ &\text{With } f(x) = \log|x|, \text{ then } df(x; dx) = \mathrm{tr}(x^{-1}dx).\\ &\text{With } f(x) = x^{-1}, \text{ then } df(x; dx) = -x^{-1}dxx^{-1}.\end{aligned}$$

Next, these differentials are applied to simple examples in order to demonstrate the notation:

Example B.1 *For the simple function $f(x) = \mathrm{tr}\{x\}$, then by 1. above:,*

$$\mathrm{tr}(x + A) - \mathrm{tr}(x) = \mathrm{tr}(A) + o(||A||)$$

or alternatively, a Taylor expansion around x_0,

$$\mathrm{tr}(x) = \mathrm{tr}(x_0) + \mathrm{tr}(x - x_0) + o(||x - x_0||).$$

Note that is this case $0 = ||x - x_0|| = ||A||$ as $f(x) = \mathrm{tr}(x)$ is linear in x. This is in fact a proof of the result 1 above.

Example B.2 *With $f(x) = \log|x|$,*

$$\log|x + H| = \log|x| + \mathrm{tr}(x^{-1}H) + o(||H||)$$

or equivalently, a Taylor expansion around $x_0 = I$,

$$\log|x| = \mathrm{tr}(x - I) + o(||x - I||)$$

which generalizes the well-known univariate result, $\log(x) = x - 1 + o(|x - 1|)$.

Example B.3 *With $f(x) = \log|I + x|$, $df(x; dx) = \mathrm{tr}\left\{(I + x)^{-1}dx\right\}$, and*

$$\log|I + x + H| = \log|I + x| + \mathrm{tr}\left\{(I + x)^{-1}H\right\} + o(||H||)$$

Equivalently, a Taylor expansion around $x_0 = 0$,

$$\log|I + x| = \mathrm{tr}\{x\} + o(||x||)$$

which is used in the proof of Theorem II.4.3.

C Stochastic Taylor expansions

Let $f : \mathbb{R}^k \mapsto \mathbb{R}$ be continuous and differentiable of suitable order. What is of interest is a Taylor expansion of f with stochastic arguments.

Theorem C.1 *Let x_T be a sequence of stochastic variables in $\mathbb{R}^k$ with*

$$x_T = c + O_P(T^\delta),$$

where $c \in \mathbb{R}^k$ and $\delta < 0$, such that $T^\delta \to 0$ as $T \to \infty$. Then if f is continuously differentiable in c,

$$f(x_T) = f\left(c\right) + df\left(c; x_T - c\right) + o_P\left(T^\delta\right) \tag{II.51}$$

Proof: By the classic Taylor's expansion as $x \to c$,

$$f(x) = f(c) + df(c; x - c) + o(||x - c||)$$

Define $\tilde{f}(x) = [f(x) - f(c) - df\left(c; x - c\right)]/||x - c||$ for $x \neq c$ and $\tilde{f}(c) = 0$. As f is continously differentiable in c, $\tilde{f}$ is continuous in c, and hence $\tilde{f}(x_T) \xrightarrow{P} \tilde{f}(c) = 0$. That is $\tilde{f}(x_T) = o_P(1)$ which implies that

$$\tilde{f}(x_T)||x_T - c|| = o_P(1)O_P(T^\delta) = o_P(T^\delta)$$

as desired. □

An immediate corollary is the following:

Corollary C.1 *Consider a univariate sequence W_T for which $TW_T \xrightarrow{D} W$ as $T \to \infty$. Then*

$$T \log\left(1 + W_T\right) \xrightarrow{D} W$$

Proof: $f\left(w\right) = \log\left(1 + w\right)$, and $df\left(w; dw\right) = \frac{1}{1+w}dw$. Hence Theorem C.1 gives the result with $c = 0$ as $W_T = O_P\left(T^{-1}\right)$,

$$Tf\left(W_T\right) = T \log\left(1 + W_T\right) = TW_T + T \cdot o_P\left(T^{-1}\right)$$

□

If the Taylor expansion is derived with s'th derivatives ($s = 1, 2, 3, \ldots$) then identical results hold with the remainder term of order $o_P(T^{\delta_s})$. The result also holds for $f : \mathbb{R}^k \mapsto \mathbb{R}^m$ and for functions of matrices, where the expansion is best stated in terms of the differential as discussed before.

Example C.1 *Using the notation from Theorem II.4.3, consider now*

$$x_T = I + W_T$$

where $W_T = O_P(T^{-1})$ *as* TW_T *converges in distribution. Hence, by Example B.3 and the stochastic Taylor expansion in (II.51),*

$$-2\log Q = T\log|x_T| = T\left(\operatorname{tr}\{W_T\} + o_P(T^{-1})\right) = \operatorname{tr}\{TW_T\} + o_P(1)$$

Thus it is the asymptotic distribution of $\operatorname{tr}\{TW_T\}$ *which defines the asymptotic distribution of the likelihood ratio statistic.*

Anders Rahbek
Rasmus Søndergaard Pedersen
University of Copenhagen

September 2024

Part III
Autoregressive modelling and unit roots

III.1 Introduction

This is an introduction to the analysis of autoregressive (AR) models in the case where the characteristic polynomial is allowed to have a root at one, a so-called unit root, rather than as has been the case until now, all roots outside the unit circle. The focus here is the univariate case, whereas the multivariate analysis, or in other words, cointegration analysis, is treated separately. Apart from reparametrizations of the univariate AR models, estimation in the presence of unit roots is based on linear regression analysis. What is changed is the asymptotic inference which is based on non classic limit distributions.

The explicit underlying assumption of the previous analyses of autoregressive models has been that of geometric ergodicity, which in terms of the simple AR(1) model with x_0 fixed and ε_t i.i.d.N$(0,\sigma^2)$,

$$x_t = \rho x_{t-1} + \varepsilon_t,\ t = 1, ..., T \tag{III.1}$$

can be stated as the parameter restriction $|\rho| < 1$. When $\rho = 1$,

$$x_t = x_{t-1} + \varepsilon_t = \sum_{t=1}^{t} \varepsilon_i + x_0, \tag{III.2}$$

that is, x_t is the sum of a random walk $\sum_{t=1}^{t} \varepsilon_i$ and the initial value x_0. Clearly when $\rho = 1$, x_t is not stationary, not even asymptotically, as for example the variance conditional on x_0, $\mathbb{V}[x_t] = t\sigma^2$, which is increasing in t. This is the kind of non-stationarity that is commonly referred to when discussing ‘non-stationary’ in the context of unit root analysis. It is easy to see that,

$$\Delta x_t = x_t - x_{t-1} = \varepsilon_t$$

i.e. the differenced process is stationary.

In short, for $\rho = 1$, x_t is a simple example of a so called I(1) process, in the sense that x_t is a non-stationary process with a random walk component, while Δx_t is stationary and also geometrically ergodic. Unit root analysis provides a framework to discriminate these two situations: the geometrically ergodic (or, simply stationary) case, and the non-stationary random walk type behavior. As is common in the literature, this is sometimes referred to as the hypothesis of stationarity and non-stationarity respectively, which should not cause any confusion.

The assumption of $|\rho| < 1$ implies in particular that inference on the parameters is based on well-known asymptotic Gaussian and χ^2 distributions. However, many if not most economic time series do not show stationary behavior and inference with such variables is not standard. To avoid this, the data series are often transformed by for example differencing (Δx_t) and taking the logarithm, or a combination thereof, to obtain approximately stationary series, which may then be analyzed using autoregressive models under the assumption of geometric ergodicity.

Univariate unit root analysis is a first step towards cointegration analysis where relations between non-stationary key economic variables can be analyzed.

In most analyses of economic time series it is not easy to distinguish if the series analyzed behave as stationary processes with possibly a linear trend or, as process with a random walk component with or without a linear trend. Unit root analysis in AR(k) models with deterministic terms help to do so and will be discussed.

As a final remark before turning to the analysis of unit roots it is important to stress that many other kinds of non-stationarity appears in the literature. For example, fractionally integrated processes where $\Delta^d x_t = (1 - L)^d x_t$ is stationary for some $0 < d < 1$, and processes with breaks in, say, the mean. The former kind is not treated here and with regard to the latter these can be analyzed by inclusion of dummy variables as in standard AR models.

III.2 The AR(1) model

In order to introduce the kind of new inference due to non-stationary processes consider again the simple AR(1) model in (III.1),

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

Now with $\rho \in \mathbb{R}$ the maximum likelihood estimator is given by,

$$(\hat{\rho} - \rho_0) = S_{\varepsilon z} S_{zz}^{-1} = \frac{\frac{1}{T}\sum_{t=1}^{T} x_{t-1}\varepsilon_t}{\frac{1}{T}\sum_{t=1}^{T} x_{t-1}^2} \tag{III.3}$$

From previous analysis it follows that if $|\rho_0| < 1$ then $x_{t-1}\varepsilon_t$ is a martingale difference sequence satisfying the assumptions of the Central Limit Theorem (CLT) for martingale differences. As a result the estimator is consistent, $\hat{\rho} \xrightarrow{P} \rho_0$ and asymptotically normally distributed,

$$\sqrt{T}(\hat{\rho} - \rho_0) = \sqrt{T} S_{\varepsilon z} S_{zz}^{-1} \xrightarrow{D} N(0, 1 - \rho_0^2).$$

In the case of $\rho_0 = 1$ and with $x_0 = 0$,

$$(\hat{\rho} - 1) = S_{\varepsilon z} S_{zz}^{-1} = \frac{\frac{1}{T}\sum_{t=1}^{T} \varepsilon_t x_{t-1}}{\frac{1}{T}\sum_{t=1}^{T} x_{t-1}^2} = \frac{\frac{1}{T}\sum_{t=1}^{T} \varepsilon_t (\sum_{i=1}^{t-1} \varepsilon_i)}{\frac{1}{T}\sum_{t=1}^{T} (\sum_{i=1}^{t-1} \varepsilon_i)^2}$$

Now first of all $x_t = \sum_{i=1}^{t} \varepsilon_i$ is not (geometrically) ergodic, let alone i.i.d., so that the usual class of laws of large numbers (LLN) do not apply. Second, while $\varepsilon_t x_{t-1}$ is a martingale difference (MGD), it does not satisfy the assumptions of a CLT. In particular with $\mathcal{F}_t = (x_t, x_{t-1}, ..)$ (or, $\sigma(x_t, ...)$) we find

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\varepsilon_t^2 x_{t-1}^2 | \mathcal{F}_{t-1}\right] = \left(\frac{1}{T}\sum_{t=1}^{T} x_{t-1}^2\right)\sigma_0^2, \quad x_t = \sum_{i=1}^{t} \varepsilon_i.$$

This does not converge in probability to a constant as required for the CLT to hold. In fact it holds that $T^{1/2}(\hat{\rho} - 1) \xrightarrow{P} 0$, $\hat{\rho}$ is said to be 'super-consistent', and furthermore, as we will see

$$T(\hat{\rho} - 1) \xrightarrow{D} \frac{\int_0^1 \mathcal{W}_u d\mathcal{W}_u}{\int_0^1 \mathcal{W}_u^2 du} = \frac{\int_0^1 \mathcal{W} d\mathcal{W}}{\int_0^1 \mathcal{W}_u^2 du}$$

where $\mathcal{W}$ is a standard Brownian motion – which is introduced in the next section – on the interval $u \in [0, 1]$. The distribution is a non-standard asymmetric distribution and was tabulated originally by Dickey and Fuller (1979). This implies that reporting the usual $t-$statistic for $\hat{\rho}$ has no meaning if x_t is indeed a random walk.

Recall that the likelihood ratio test statistic of the hypothesis $H : \rho = \rho_0$ is given by

$$\text{LR}(\rho = \rho_0) = T\log(1 + W_T), \text{ where } W_T = (\hat{\rho} - \rho_0)^2 S_{zz}/\hat{\sigma}^2. \tag{III.4}$$

When $|\rho_0| < 1$,

$$W = TW_T \overset{D}{\to} \chi^2, \tag{III.5}$$

while for the case of a unit root, $\rho_0 = 1$,

$$W = TW_T \overset{D}{\to} \frac{(\int_0^1 \mathcal{W} d\mathcal{W})^2}{\int_0^1 \mathcal{W}_u^2 du} \neq \chi^2, \tag{III.6}$$

which is the (square) of the so-called Dickey-Fuller distribution. It has broader tails than the χ^2 distribution. For example the 95% quantile is approximately 4.2 which should be compared with 95% quantile of the χ^2 distribution, 3.84.

These results are explained in detail in the next sections and it is briefly commented upon that the results are similar for AR processes with more lags.

III.3 Brownian motion

In this section it is discussed in what sense LLN and CLTs hold for functions of the random walk.

III.3.1 Brownian motion

Consider again the random walk,

$$x_0 = 0 \tag{III.7}$$

$$x_t = \sum_{i=1}^{t} \varepsilon_i \text{ for } t = 1, \ldots, T \tag{III.8}$$

The main features of the random walk defined this way are that x_t is $\mathrm{N}(0, t\sigma^2)$ distributed and that x_t has independent increments, i.e. Δx_t and Δx_{t+k} are independent for $k \neq 0$. That $x_0 = 0$ is merely a convenient convention. In the analyses of AR models the initial value x_0 is fixed, but can be ignored in the asymptotic analysis.

The continuous time equivalent of the random walk relevant here is the Brownian motion, $\mathcal{B}$ defined on the unit-interval $[0, 1]$. Now $\mathcal{B}$ is a function of time $u \in [0, 1]$, it is a stochastic process, and a realization of $\mathcal{B}$ is a continuous function on $[0, 1]$ with the following properties:

Definition III.3.1 *The Brownian motion $\mathcal{B}$ with variance σ^2 defined on $[0, 1]$ is a stochastic process with the properties:*

1. $\mathcal{B}_0 = 0$
2. $\mathcal{B}_u$ *is* $\mathrm{N}(0, u\sigma^2)$ *distributed for all* $u \in [0,1]$
3. *For any* $0 \leq u_1 < \ldots < u_k \leq 1$ *the increments* $(\mathcal{B}(u_2) - \mathcal{B}(u_1)), \ldots, (\mathcal{B}(u_k) - \mathcal{B}(u_{k-1}))$ *are independent*
4. $\mathcal{B}_u$ *is continuous as a function of* u.

If $\sigma^2 = 1$, $\mathcal{B} \equiv \mathcal{W}$ *is a standard Brownian motion.*

The main difference between the random walk and the Brownian motion is the continuity. The class of functions on the unit interval which are continuous will be referred to as $C(0,1)$.

In order to change the time scaling of the random walk to be on the unit interval define for $u \in [0,1]$ the process,

$$x_0^T = 0 \tag{III.9}$$

$$x_u^T = \frac{1}{\sqrt{T}} \sum_{i=1}^{[Tu]} \varepsilon_i \tag{III.10}$$

where $[Tu]$ is the integer value of Tu, $u \in [0,1]$. For $u = 0, 1/T, 2/T, .., 1 = T/T$ clearly x_u^T is just the random walk divided by $\sqrt{T}$. In between these points x_u^T is constant, see figure 1.

Thus x_u^T is an example of a process on $[0,1]$, which is right-continuous and has limits from the left, a so-called càdlàg process. I.e. it is on the right scale but not continuous. It is no problem to connect the points of the random walk to make it a $C(0,1)$ function, but this makes the derivations complicated. In fact, that x_u^T is not continuous can be ignored in the following, since the limit is indeed continuous as demonstrated in the outline of the proof of Theorem III.3.2 below. The main theorem (the invariance principle) states that

$$x_\cdot^T = \frac{1}{\sqrt{T}} \sum_{i=1}^{[T\cdot]} \varepsilon_i \xrightarrow{D} \mathcal{B}. \tag{III.11}$$

uniformly on $[0,1]$ as $T \to \infty$, where $\mathcal{B}$ is a Brownian motion with variance σ^2. This is in accordance with Figure 1.

III.3.2 Invariance principle

The limit theorems so far regarding convergence of stationary processes have treated convergence in distribution and probability of random variables defined on $\mathbb{R}$ or $\mathbb{R}^p$. The statement in (III.11) is a statement about convergence in distribution of a random variable on $C(0,1)$ instead, i.e. a functional limit theorem.

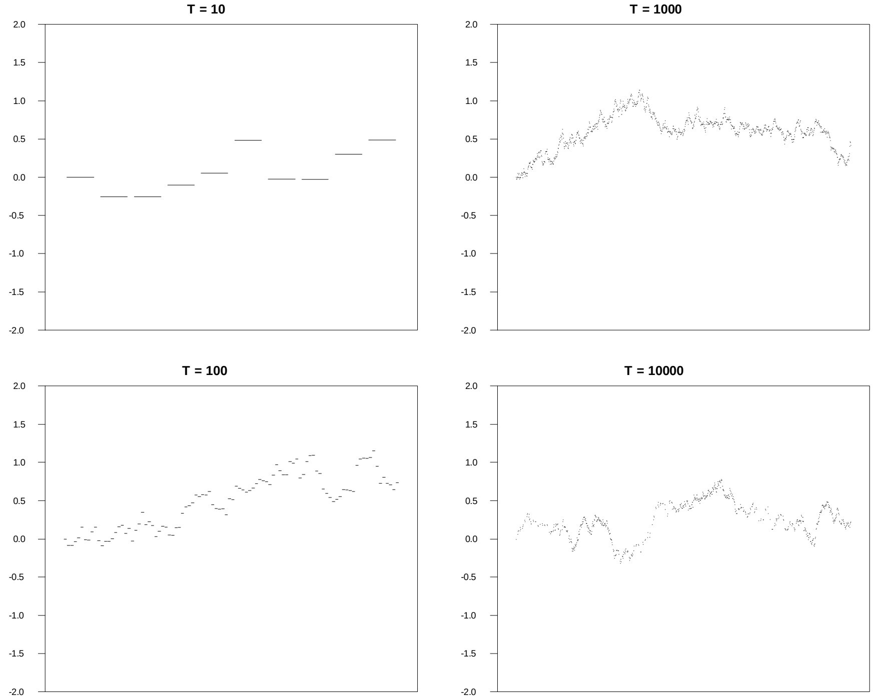


Figure 1: Simulations of $X_{[Tu]}$ based on $\epsilon_t's$ drawn from the $N(0,1)$ distribution.

In order to understand the convergence, introduce the metric on $C(0,1)$ which is given by the supremum, i.e. with x in $C(0,1)$,

$$x^T \to x \text{ if } \sup_u |x_u^T - x_u| \to 0,\, T \to \infty \tag{III.12}$$

Thus for example if x_u^T, or simply x^T, is a sequence of random variables on $C(0,1)$ then,

$$x_u^T \xrightarrow{P} 0, \tag{III.13}$$

where P is defined on $C(0,1)$, means that $\sup_{u\in[0,1]} |x_u^T| \xrightarrow{P} 0$ on $\mathbb{R}$.

To prove convergence in distribution on $C(0,1)$, as in (III.11), two things are needed. First the finite-dimensional distributions of x_u^T need to converge, i.e. for any $0 \leq u_1 < u_2 < \ldots < u_k \leq 1$,

$$(x_{u_1}^T, \ldots, x_{u_k}^T) \xrightarrow{D} (\mathcal{B}_{u_1}, \ldots, \mathcal{B}_{u_k}) \tag{III.14}$$

Thus e.g. for any fixed u, $x_u^T \xrightarrow{D} \mathcal{B}_u$ which is simply the $N(0, \sigma^2 u)$ distribution. For the case here this is trivial since any CLT gives for i.i.d. processes gives,

$$x_u^T = \frac{1}{\sqrt{T}} \sum_{i=1}^{[Tu]} \varepsilon_i = \sqrt{\frac{[Tu]}{T}} \frac{1}{\sqrt{[Tu]}} \sum_{i=1}^{[Tu]} \varepsilon_i \xrightarrow{D} \sqrt{u} N(0, \sigma^2) = N(0, \sigma^2 u). \tag{III.15}$$

What is needed ,in addition to prove convergence on $C(0,1)$, is 'tightness'. When discussing convergence on $\mathbb{R}$ for fixed u this is equivalent to $x_u^T = O_P(1)$. On $C(0,1)$ the equivalent definition of tightness is that there exists a compact set $K \subset C(0,1)$, such that

$$P(x^T \in K) \geq 1 - \delta,\ \delta > 0 \ \text{ for all } T. \tag{III.16}$$

Thus the probability mass cannot 'escape to infinity'. Tightness is not discussed further. Instead an excellent reference is Billingsley (1968), where also a detailed proof of Theorem III.3.1 below is given.

Theorem III.3.1 *(Donsker's Theorem or Invariance Principle)*
Let $\varepsilon_t, t = 1, .., T$ be i.i.d. $\mathrm{N}(0, \sigma^2)$ *then*

$$\frac{1}{\sqrt{T}} \sum_{i=1}^{[T\cdot]} \varepsilon_i \xrightarrow{D} \mathcal{B}. \quad , \tag{III.17}$$

on $C(0,1)$, with $\mathcal{B}$ a Brownian motion on $[0,1]$ with variance σ^2.

III.3.3 Invariance principle for martingale differences

In fact, Theorem III.3.1 is a corollary to the general invariance principle for martingales, the functional central limit theorem (FCLT). This we stated for $u = 1$ as the CLT in Theorem I.4.4 (stated as a corollary to Brown, 1971).

Theorem III.3.2 *Let $(Y_t)_{t=1,2,...}$, with $\mathbb{E}[Y_t^2] < \infty$, be a martingale difference sequence with respect to the increasing sequence $\mathcal{F}_t$. Assume further that, (i) and (ii), or (i) and (ii') hold for some $\delta > 0$ and as $T \to \infty$,*

$$\textit{(i)}: \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[Y_t^2 | \mathcal{F}_{t-1}\right] \xrightarrow{P} \sigma_y^2 > 0 \quad \textit{and} \tag{III.18}$$

$$\textit{(ii)}: \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[Y_t^2 \mathbb{I}(|Y_t| > \delta\sqrt{T})\right] \to 0, \textit{ or} \tag{III.19}$$

$$\textit{(ii)'}: \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[Y_t^2 \mathbb{I}(|Y_t| > \delta\sqrt{T}) | \mathcal{F}_{t-1}\right] \xrightarrow{P} 0. \tag{III.20}$$

Then, as $T \to \infty$*,*

$$\frac{1}{\sqrt{T}} \sum_{i=1}^{[T\cdot]} Y_i \overset{D}{\to} \mathcal{B}.$$

where $\mathcal{B}$ *is a Brownian motion with variance* σ^2*.*

III.3.4 Convergence to integrals

Now turning to the estimator $\hat{\rho}$ and the likelihood ratio test statistic of the previous section for the hypothesis that $\rho = 1$ it follows that what is interesting is the asymptotic behavior of terms such as,

$$\sum_{t=1}^{T} x_{t-1}^2 = \sum_{t=1}^{T} (\sum_{i=1}^{t-1} \varepsilon_i)^2 \text{ and } \sum_{t=1}^{T} x_{t-1}\varepsilon_t = \sum_{t=1}^{T} (\sum_{i=1}^{t-1} \varepsilon_i)\varepsilon_t.$$

The latter involves the definition of the stochastic integral $\int_0^1 \mathcal{B}_u d\mathcal{B}_u$ (or, simply $\int \mathcal{B} d\mathcal{B}$) whereas the former involves $\int_0^1 \mathcal{B}_u^2 du$.

Recall that $x_T \overset{D}{\to} x$ on $\mathbb{R}$ means that $f(x_T) \overset{D}{\to} f(x)$ for any continuous function $f : \mathbb{R} \mapsto \mathbb{R}$ This part of the definition of convergence in distribution is important, and is commonly referred to as the 'continuous mapping theorem'. By definition of convergence in distribution, this holds for any metric space, in particular $C(0,1)$. That is, for continuos functions, or mappings, as for example, $f : C(0,1) \to \mathbb{R}$ or $f : C(0,1) \to C(0,1)$.

Rewrite next, $\sum_{t=1}^{T} x_{t-1}^2$ for $\rho = 1$ as,

$$T^{-2} \sum_{t=1}^{T} (\sum_{i=1}^{t} \varepsilon_i)^2 = T^{-1} \sum_{u=1/T}^{1} (\frac{1}{\sqrt{T}} \sum_{i=1}^{[Tu]} \varepsilon_i)^2 = \int_0^1 (\frac{1}{\sqrt{T}} \sum_{i=1}^{[Tu]} \varepsilon_i)^2 du, \tag{III.21}$$

using the piecewise constancy of $\sum_{i=1}^{[Tu]} \varepsilon_i$. Now the mapping $x \mapsto f(x) = \int_0^1 x_u du$ from $C(0,1)$ into $\mathbb{R}$ is continuous. To see this let $x^T \to x$ on $C(0,1)$ and evaluate

$$|\int_0^1 (x_u^T - x_u) du| \leq \int_0^1 |x_u^T - x_u| du \leq \sup_{u \in [0,1]} |x_u^T - x_u| \tag{III.22}$$

which tends to zero by definition of convergence on $C(0,1)$. Similarly, $f(x) = \int_0^1 x_u^2 du$ is continuous and hence by applying Donsker's theorem and the definition of convergence in distribution,

$$T^{-2} \sum_{t=1}^{T} (\sum_{i=1}^{t} \varepsilon_i)^2 \overset{D}{\to} \int_0^1 \mathcal{B}_u^2 du \tag{III.23}$$

Unfortunately, the continuous mapping argument cannot be applied for the convergence of $\sum_{t=1}^{T}(\sum_{i=1}^{t-1}\varepsilon_i)\varepsilon_t$ and instead a heuristic argument is given. A proof of the result is found in Chan & Wei (1988). Rewrite the term as follows

$$\begin{aligned} T^{-1}\sum_{t=1}^{T}(\sum_{i=1}^{t-1}\varepsilon_i)\varepsilon_t &= \sum_{u=1/T}^{1}(\frac{1}{\sqrt{T}}\sum_{i=1}^{[Tu]-1}\varepsilon_i)\Delta(\frac{1}{\sqrt{T}}\sum_{i=1}^{[Tu]}\varepsilon_i) \\ &\simeq \sum_{u=1/T}^{1}\mathcal{B}_u d\mathcal{B}_u \overset{"D"}{\to} \int_0^1 \mathcal{B}_u d\mathcal{B}_u \text{ as T} \to \infty. \qquad \text{(III.24)} \end{aligned}$$

In both cases the important thing is not the exact form of the limit, but that in fact the terms do converge in distribution.

What it means is that for example the stochastic variable $\int_0^1 \mathcal{B}d\mathcal{B}$ can be simulated by $T^{-1}\sum_{t=1}^{T}(\sum_{i=1}^{t-1}\varepsilon_i)\varepsilon_t$ for large T. In fact all the quoted limit distributions have been tabulated that way.

Collecting the results gives:

Theorem III.3.3 *Under the assumption that $\varepsilon_t, t = 1, .., T$, are iid$N(0, \sigma^2)$ then*

$$(\frac{1}{\sqrt{T}}\sum_{i=1}^{[Tu]}\varepsilon_t, T^{-1}\sum_{t=1}^{T}(\sum_{i=1}^{t-1}\varepsilon_i)\varepsilon_t, T^{-2}\sum_{t=1}^{T}(\sum_{i=1}^{t}\varepsilon_i)^2) \xrightarrow{D} (\mathcal{B}_u, \int_0^1 \mathcal{B}_u d\mathcal{B}_u, \int_0^1 \mathcal{B}_u^2 du)$$

where $\mathcal{B}$ is a Brownian motion with variance σ^2.

As for the FCLT the result can be further generalized to martingale difference sequences by Hansen (1992, Theorem 2.1):

Theorem III.3.4 *Let Y_t be a martingale sequence with respect to $\mathcal{F}_t$, for which as $T \to \infty$,*

$$\text{(i)}: \quad \frac{1}{\sqrt{T}}\sum_{t=1}^{[Tu]} Y_t \xrightarrow{D} \mathcal{B}_u \text{ with variance } \sigma^2 \text{ and}$$

$$\text{(ii)}: \quad \sup_T \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[Y_t^2\right] < \infty$$

Then, as $T \to \infty$,

$$(\frac{1}{\sqrt{T}}\sum_{i=1}^{[Tu]} Y_t, T^{-1}\sum_{t=1}^{T}(\sum_{i=1}^{t-1} Y_i)Y_t, T^{-2}\sum_{t=1}^{T}(\sum_{i=1}^{t} Y_i)^2) \xrightarrow{D} (\mathcal{B}_u, \int_0^1 \mathcal{B}d\mathcal{B}, \int_0^1 \mathcal{B}_u^2 du).$$

III.3.5 The AR(1) model reconsidered

Given the above presented theory, the results for the AR(1) model in (III.1) can be collected in the following theorem:

Theorem III.3.5 *Consider the AR(1) model in (III.1) with ML estimators given by,*

$$\hat{\rho} = S_{yz}S_{zz}^{-1} \text{ and } \hat{\sigma}^2 = \frac{1}{T}\sum_{t=1}^{T}(x_t - \hat{\rho}x_{t-1})^2 = S_{yy} - S_{yz}S_{zz}^{-1}S_{zy}, \tag{III.25}$$

where y, z in the product moments refer to x_t and x_{t-1} respectively. When $\rho = 1$, they are consistent, that is $\hat{\rho} \overset{P}{\to} 1$ and $\hat{\sigma}^2 \overset{P}{\to} \sigma_0^2$ as $T \to \infty$. Moreover,

$$T(\hat{\rho} - 1) \overset{D}{\to} \int_0^1 \mathcal{W}d\mathcal{W} / \int_0^1 \mathcal{W}_u^2 du, \tag{III.26}$$

where $\mathcal{W}$ is a standard Brownian motion.

The LR test statistic for the hypothesis that $\rho = 1$ is given by,

$$LR(\rho = 1) = T\log(1 + W_T), \quad W_T = (\hat{\rho} - 1)^2 \frac{S_{zz}}{\hat{\sigma}^2}. \tag{III.27}$$

For $\rho = 1$, the LR statistic is asymptotically Dickey-Fuller type distributed,

$$LR(\rho = 1) \overset{D}{\to} (\int_0^1 \mathcal{W}d\mathcal{W})^2 / \int_0^1 \mathcal{W}_u^2 du \quad \text{as } T \to \infty. \tag{III.28}$$

Proof:

The expressions for the ML estimators and the LR statistic in (III.25) and (III.27) follow by the results for the AR(1) model established in Part 2 of the lecture notes. The result in (III.26), holds as

$$T(\hat{\rho} - 1) = S_{\varepsilon z}S_{zz}^{-1} = \frac{\frac{1}{T}\sum_{t=1}^{T}\varepsilon_t x_{t-1}}{\frac{1}{T^2}\sum_{t=1}^{T}x_{t-1}^2}.$$

Consider first the denominator,

$$\frac{1}{T}\sum_{t=1}^{T}\varepsilon_t x_{t-1} = \frac{1}{T}\sum_{t=1}^{T}\varepsilon_t\sum_{i=1}^{t-1}\varepsilon_i + x_0\frac{1}{T}\sum_{t=1}^{T}\varepsilon_t.$$

The first term converges in distribution to $\int_0^1 \mathcal{B}d\mathcal{B}$ by Theorem III.3.4 as $\Delta x_t = \varepsilon_t$ are i.i.d.N$(0, \sigma^2)$. The second term converges in probability to

$\mathbb{E}[\varepsilon_t] = 0$ by LLN for i.i.d. processes. For the numerator, applying Theorem III.3.4 gives the desired, as $\Delta x_t = \varepsilon_t$ is a martingale difference.

Finally, turn to the LR statistic, where

$$TW_T = TS_{\varepsilon z}S_{zz}^{-1}S_{\varepsilon z}/\hat{\sigma}^2 \xrightarrow{D} \int_0^1 \mathcal{B}d\mathcal{B}(\int_0^1 \mathcal{B}_u^2 du)^{-1}\int_0^1 \mathcal{B}d\mathcal{B}/\sigma_0^2, \qquad \text{(III.29)}$$

using the same arguments as before, and that $\hat{\sigma}^2$ is consistent. The result in (III.28) follows by setting $\mathcal{W} = \frac{1}{\sqrt{\sigma^2}}\mathcal{B}$. The consistency of $\hat{\sigma}^2$ can be seen by rewriting,

$$\hat{\sigma}^2 = S_{\varepsilon\varepsilon} - (\hat{\rho} - 1)^2 S_{zz} = S_{\varepsilon\varepsilon} + o_P(1) \xrightarrow{P} \sigma^2. \qquad \text{(III.30)}$$

That $(\hat{\rho} - 1)^2 S_{zz} = o_P(1)$, holds as $(\hat{\rho} - 1) S_{zz} = O_P(1)$, while $\hat{\rho} \xrightarrow{P} 1$. □

III.4 Testing for unit-roots in AR(k) models

It is discussed here what it means to allow for a unit root in the AR(k) model and how to test for it.

III.4.1 I(1) and I(0) Processes

First a definition of 'non-stationarity' and 'stationarity' suitable for the classification of AR processes is needed. Recall that the AR(1) process x_t is geometrically ergodic with stationary version $x_t^* = \sum_{i=0}^{\infty} \rho^i \varepsilon_{t-i}$ if $|\rho| < 1$, while if $\rho = 1$, $x_t - x_0$ is a random walk.

Definition III.4.1 *A geometrically ergodic process x_t is called I(0), if the stationary version x_t^* is a linear proces which satisfies $x_t^* = \phi(L)\varepsilon_t = \sum_{i=0}^{\infty} \phi_i \varepsilon_{t-i}$ with $\phi(1) = \sum_{i=0}^{\infty} \phi_i \neq 0$.*

Example III.4.1 *For $|\rho| < 1$, the AR(1) process $x_t^* = \sum_{i=0}^{\infty} \rho^i \varepsilon_{t-i}$ is I(0), because $\sum_{i=0}^{\infty} \rho^i = (1 - \rho)^{-1} \neq 0$.*

Example III.4.2 *Consider the AR(1) process with a constant μ and $|\rho| < 1$,*

$$x_t = \rho x_{t-1} + \mu + \varepsilon_t, \qquad \text{(III.31)}$$

which is geometrically ergodic with stationary solution x_t^* for which,

$$x_t^* - \mathbb{E}[x_t^*] = x_t^* - \frac{\mu}{1-\rho} = \sum_{i=0}^{\infty} \rho^i \varepsilon_{t-i}, \quad \text{and} \ \sum_{i=0}^{\infty} \rho^i = (1-\rho)^{-1}. \qquad \text{(III.32)}$$

Hence, $x_t - \frac{\mu}{1-\rho}$ is an I(0) process. Similarly, for $|\rho| < 1$ the AR(1) process with a linear trend,

$$x_t = \rho x_{t-1} + \mu_0 + \mu_1 t + \varepsilon_t, \tag{III.33}$$

can be written as $x_t = x_t + \frac{\mu_1}{1-\rho} t$ where

$$x_t = \rho x_{t-1} + \tilde{\mu} + \varepsilon_t, \tag{III.34}$$

and $\tilde{\mu} = \mu_0 - \frac{\mu_1}{1-\rho}$. Hence, $x_t - \frac{\tilde{\mu}}{1-\rho} t$ is an I(0) process, with

$$x_t^* = \frac{\tilde{\mu}}{1-\rho} + \sum_{i=0}^{\infty} \rho^i \varepsilon_{t-i}. \tag{III.35}$$

That is, x_t with its linear trend subtracted, $x_t - \frac{\mu_1}{1-\rho} t$ is an I(0) process and has a stationary version. Thus x_t is trend-I(0), and x_t^* trend-stationary.

As an example of a stationary process, which is not I(0) consider,

$$x_t = \Delta \varepsilon_t = \varepsilon_t - \varepsilon_{t-1} \tag{III.36}$$

Clearly, x_t is stationary and linear, but it is not I(0). The reason for not including $\Delta\varepsilon_t$ in the I(0) processes is that, when accumulating it, $\sum_{i=1}^{t} \Delta\varepsilon_i = \varepsilon_t - \varepsilon_0$, it is still stationary and no CLT, and hence FCLT, applies to the accumulated process. This is contrary to $x_t = \varepsilon_t$, where both a FCLT and CLT applies.

Next define I(1) and I(2) processes.

Definition III.4.2 *A stochastic process x_t is called integrated of order $d = 1, 2$, I(d), if $\Delta^d x_t$ is I(0).*

Example III.4.3 *A random walk, $x_t = \sum_{i=1}^{t} \varepsilon_i$, is indeed an I(1) process. For the random walk with drift,*

$$x_t = \sum_{i=1}^{t} \varepsilon_t + \mu t,$$

$x_t - \mu t$ is an I(1) processes.

Example III.4.4 *An example of an I(2) process is $\Delta^2 x_t = \varepsilon_t$. While processes integrated of order 2 are relevant for empirical applications, processes integrated of order higher than 2 have so far no practical applications.*

III.4.2 The AR(2) model

Consider the AR(2) model as given by

$$x_t = \rho_1 x_{t-1} + \rho_2 x_{t-2} + \varepsilon_t, \quad t = 1, 2, ..., T \tag{III.37}$$

with x_0 and x_{-1} fixed, ε_t i.i.d.N(0,σ^2) and parameters $(\rho_1, \rho_2, \sigma^2) \in \mathbb{R}^2 \times \mathbb{R}_+$.

The characteristic polynomial evaluated at $z = 1$,

$$A(1) = 1 - \rho_1 - \rho_2,$$

is zero if and only if $\rho_1 + \rho_2 = 1$. To make this restriction even simpler, reparametrize the AR(2) model as

$$\Delta x_t = \pi x_{t-1} + \gamma \Delta x_{t-1} + \varepsilon_t, \tag{III.38}$$

where $\pi = \rho_1 + \rho_2 - 1$, and $\gamma = -\rho_2$ are freely varying parameters, and $(x_0, \Delta x_0)$ fixed. In this parametrization, the characteristic polynomial is given by

$$A(z) = (1 - z) - \pi z - \gamma z (1 - z), \tag{III.39}$$

which is zero at $z = 1$ if and only if $\pi = 0$. Hence the hypothesis of a unit root is equivalent to,

$$H_0 : \pi = 0.$$

III.4.2.1 Properties of x_t when $\pi = 0$.

Under H_0, and if $|\gamma| < 1$, then $S_t \equiv \Delta x_t$ is geometrically ergodic with stationary solution,

$$S_t^* = \sum_{i=0}^{\infty} \gamma^i \varepsilon_{t-i}.$$

Next, use that by definition under H_0, $\gamma(z) = 1 - \gamma z = \gamma(1) + \gamma(1 - z)$, with $\gamma(L) \Delta x_t = \varepsilon_t$, such that

$$\gamma(L) \Delta x_t = \gamma(1) \Delta x_t + \gamma \Delta^2 x_t = \varepsilon_t.$$

Hence, with $|\gamma| < 1$,

$$\Delta x_t = \frac{1}{\gamma(1)} (-\gamma \Delta^2 x_t + \varepsilon_t),$$

and summation over t gives,

$$x_t = \frac{1}{\gamma(1)} \sum_{i=1}^{t} \varepsilon_i + \frac{\gamma}{\gamma(1)} S_t + (x_0 - \frac{\gamma}{\gamma(1)} \Delta x_0). \tag{III.40}$$

In other words, under the hypothesis of a unit root, or $\pi = 0$, and if $|\gamma| < 1$, then x_t can be represented as the sum of a random walk, a geometrically ergodic process S_t and initial values as a function of x_0 and x_{-1}. Thus x_t is non-stationary as an I(1) process.

Furthermore, with $u \in [0,1]$,

$$\frac{1}{\sqrt{T}} x_{[Tu]} = 1/\gamma(1) \frac{1}{\sqrt{T}} \sum_{i=1}^{[Tu]} \varepsilon_i + o_P(1) \overset{D}{\to} 1/\gamma(1) \mathcal{B}_u = \frac{1}{1-\gamma} \mathcal{B}_u, \qquad \text{(III.41)}$$

where $\mathcal{B}$ is a Brownian motion with variance σ^2. That is, a FCLT applies to x_t, and the limit is the same Brownian motion as in the AR(1) case but multiplied by a constant which reflects the further lag.

That (III.41) holds, follows by applying Theorem III.3.2 to ε_t, which shows that $\sum_{i=1}^{[T\cdot]} \varepsilon_i/\sqrt{T}$ converge in distribution to $\mathcal{B}$ on $C(0,1)$. Next, note that $\frac{1}{\sqrt{T}}(x_0 - \frac{\gamma}{\gamma-1}\Delta x_0) \overset{P}{\to} 0$ as x_0 and Δx_0 are fixed initial values. Finally, for the $o_P(1)$ term one needs to show $\frac{1}{\sqrt{T}} S_{[Tu]}$ converge to zero in probability on $C[0,1]$.

Some further results are needed for the product moments of x_t and $\Delta x_t = S_t$. These, as well as the previous considerations, are stated in the proposition:

Proposition III.4.1 *Consider x_t given by (III.38). If $\pi = 0$, then $\gamma(L)\Delta x_t = \varepsilon_t$, where $\gamma(z) = 1 - \gamma z$. If furthermore, $|\gamma| < 1$, $S_t \equiv \Delta x_t$ is geometrically ergodic, and x_t has the representation in (III.40). Moreover, as $T \to \infty$, it holds that with $u \in [0,1]$ and $\mathcal{B}_u$ a Brownian motion with variance σ^2,*

$$\begin{aligned}(\frac{1}{\sqrt{T}} \sum_{t=1}^{[T\cdot]} x_t, \frac{1}{T^2} \sum_{t=1}^{T} x_{t-1}^2, \frac{1}{T} \sum_{t=1}^{T} x_{t-1} S_t) &\overset{D}{\to} \qquad \text{(III.42)}\\ (\frac{1}{\gamma(1)} \mathcal{B}_\cdot, \frac{1}{\gamma(1)^2} \int_0^1 \mathcal{B}_u^2 du, \frac{1}{\gamma(1)^2} \int_0^1 \mathcal{B} d\mathcal{B} + \omega),&\end{aligned}$$

where $\omega = \sum_{h=1}^{\infty} Cov(S_t^*, S_{t+h}^*)$. Also, jointly,

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} S_{t-1} \varepsilon_t \overset{D}{\to} N\left(0, \sigma^2 \mathbb{V}(S_t^*)\right). \qquad \text{(III.43)}$$

Proof: The convergence to $\mathcal{B}$ and $\int \mathcal{B}^2 du$ hold by (III.41) and the continuity of $f(x) = \int_0^1 x_u^2 du$, $f : C(0,1) \to \mathbb{R}$. The convergence towards the stochastic integral holds by Theorem 4.1 in Hansen (1992), in combination with the LLN. The Gaussian limit holds by the CLT for martingale differences. □

The following result is used repeatedly:

Lemma III.4.1 *Let S_t be a geometrically ergodic process, with $\mathbb{E}\left[|S_t^*|^3\right] < \infty$. Then*

$$\frac{1}{\sqrt{T}} S_{[T\cdot]} \xrightarrow{P} 0$$

as $T \to \infty$.

Proof: By definition of convergence in probability on $C(0,1)$ it has to be shown that

$$P(\sup_{u\in[0,1]} \frac{1}{\sqrt{T}} |S_{[T\cdot]}| > \delta) \to 0.$$

As before, this follows by,

$$P(\sup_{u\in[0,1]} \frac{1}{\sqrt{T}} |S_{[T\cdot]}| > \delta) = P(\max_{t=1,..,T} |S_t| > \delta\sqrt{T}) \tag{III.44}$$

$$\leq \frac{1}{T^{3/2}\delta^3} \sum_{t=1}^{T} E(|S_t|^3 \mathbb{I}(|S_t| > \delta\sqrt{T})) = O(T^{-1/2})$$

□

III.4.2.2 Testing $\pi = 0$

With $Y_t \equiv \Delta x_t$, $\beta' = (\pi, \gamma)$ and $Z_t = (x_{t-1}, \Delta x_{t-1})'$, the AR(2) model in (III.38) can be analyzed by OLS in the linear regression, $Y_t = \beta' Z_t + \varepsilon_t$ and the following theorem holds:

Theorem III.4.1 *Consider the AR(2) model in (III.38) with ML estimators given by,*

$$\hat{\beta}' = (\hat{\pi}, \hat{\gamma}) = S_{yz} S_{zz}^{-1}, \quad \textit{and } \hat{\sigma}^2 = S_{yy\cdot z}. \tag{III.45}$$

where y, z in the product moments refer to $Y_t = \Delta x_t$ and $Z_t = (x_{t-1}, \Delta x_{t-1})'$ respectively.

When $\pi = 0$, and $|\gamma| < 1$, they are consistent, that is $\hat{\pi} \xrightarrow{P} 1, \hat{\gamma} \xrightarrow{P} \gamma$ and $\hat{\sigma}^2 \xrightarrow{P} \sigma^2$ as $T \to \infty$. Moreover,

$$(T\hat{\pi}, \sqrt{T}(\hat{\gamma} - \gamma)) \xrightarrow{D} ((1-\gamma) \int_0^1 \mathcal{W} d\mathcal{W} / \int_0^1 \mathcal{W}_u^2 du,\ N\left(0, 1-\gamma^2\right)) \tag{III.46}$$

where $\mathcal{W}$ is a standard Brownian motion.

The LR statistic for the hypothesis $H_0 : \pi = 0$, is given by,

$$LR\,(\pi = 0) = T \log\left(1 + W_T\right), \tag{III.47}$$

$$W = TW_T = TS_{yz}S_{zz}^{-1}R(R'S_{zz}^{-1}R)^{-1}R'S_{zz}^{-1}S_{zy}/\hat{\sigma}^2, \tag{III.48}$$

where $R' = (1, 0)$. When $\pi = 0$, and $|\gamma| < 1$, the LR statistic is asymptotically Dickey-Fuller type distributed,

$$LR\,(\pi = 0) \xrightarrow{D} (\int_0^1 \mathcal{W}d\mathcal{W})^2 / \int_0^1 \mathcal{W}_u^2 du \quad \textit{as } T \to \infty. \tag{III.49}$$

Note that the limiting distributions are not affected by the presence of the additional lag as can be seen by comparing with for example (III.29) and Theorem III.3.5. However, the small sample distributions may be severely affected by (estimation of) γ.

Observe that $\hat{\pi}$ is super-consistent, while $\hat{\gamma}$ is the usual $\sqrt{T}$ consistent.

Note also that the LR statistic can be written in many ways, the one in (III.48) has been chosen to emphasize the role of the restriction matrix $R' = (0, 1)$. With $Z_{1t} = x_{t-1}, Z_{2t} = \Delta x_{t-1}$, the formula may also be written as

$$W = TW_T = TS_{y1\cdot 2}S_{11\cdot 2}^{-1}S_{1y\cdot 2}/\hat{\sigma}^2, \tag{III.50}$$

where, for example, $S_{y1\cdot 2} = S_{yz_1} - S_{yz_2}S_{z_2z_2}^{-1}S_{z_2z_1}$.

This corresponds to the usual regression interpretation of $\hat{\pi}$, that is, $\hat{\pi}$ can be found in two steps. First regress Δx_t and x_{t-1} on Δx_{t-1} and obtain the residuals,

$$R(\Delta x_t|\Delta x_{t-1}) = \Delta x_t - S_{yz_2}S_{z_2z_2}^{-1}\Delta x_{t-1},$$
$$R(x_{t-1}|\Delta x_{t-1}) = x_{t-1} - S_{z_1z_2}S_{z_2z_2}^{-1}\Delta x_{t-1}.$$

Next, $\hat{\pi}$ is obtained by OLS regression of $R(\Delta x_t|\Delta x_{t-1})$ on $R(x_{t-1}|\Delta x_{t-1})$, giving $\hat{\pi} = S_{y1\cdot 2}S_{11\cdot 2}^{-1}$.

Proof of Theorem III.4.1:

Using Theorems III.6 and III.9, the ML estimators are given by (III.45). Likewise the LR statistic of $\pi = 0$ can be written as in (III.47) using for example equation (III.43) and the proof of Theorem III.9.

Corresponding to the different rates of convergence, define the normalization matrix N_T as,

$$N_T \equiv \begin{pmatrix} \frac{1}{\sqrt{T}} & 0 \\ 0 & 1 \end{pmatrix}.$$

By definition, $\hat{\beta} - \beta = S_{zz}^{-1} S_{z\varepsilon}$ and hence,

$$\begin{pmatrix} T\hat{\pi} \\ \sqrt{T}(\hat{\gamma} - \gamma) \end{pmatrix} = (N_T S_{zz} N_T)^{-1} \sqrt{T} N_T S_{z\varepsilon}$$

Note that,

$$\begin{aligned} N_T S_{zz} N_T = & \begin{pmatrix} \frac{1}{T^2} \sum_{t=1}^T x_{t-1}^2 & \frac{1}{T^{3/2}} \sum_{t=1}^T x_{t-1} \Delta x_{t-1} \\ \frac{1}{T^{3/2}} \sum_{t=1}^T x_{t-1} \Delta x_{t-1} & \frac{1}{T} \sum_{t=1}^T \Delta x_{t-1}^2 \end{pmatrix} \\ & \overset{D}{\rightarrow} \begin{pmatrix} (\frac{1}{1-\gamma})^2 \int \mathcal{B}_u^2 du & 0 \\ 0 & \frac{\sigma^2}{(1-\gamma^2)} \end{pmatrix}, \end{aligned}$$

where $\mathcal{B}$ is a Brownian motion with variance σ^2. This follows by Proposition III.4.1. Likewise,

$$\sqrt{T} N_T S_{z\varepsilon} = \begin{pmatrix} \frac{1}{T} \sum_{t=1}^T x_{t-1} \varepsilon_t \\ \frac{1}{\sqrt{T}} \sum_{t=1}^T \Delta x_{t-1} \varepsilon_t \end{pmatrix} \overset{D}{\rightarrow} \begin{pmatrix} \frac{1}{1-\gamma} \int \mathcal{B} d\mathcal{B} \\ N(0, \frac{\sigma^4}{(1-\gamma^2)}) \end{pmatrix},$$

Collecting terms,

$$\begin{pmatrix} T\hat{\pi} \\ \sqrt{T}(\hat{\gamma} - \gamma) \end{pmatrix} \overset{D}{\rightarrow} \begin{pmatrix} (1-\gamma) \left(\int \mathcal{B}_u^2 du \right)^{-1} \int \mathcal{B} d\mathcal{B} \\ N\left(0, 1 - \gamma^2\right) \end{pmatrix}$$

and (III.46) follows. For the LR statistic, consider TW_T,

$$TW_T = T S_{\varepsilon z} S_{zz}^{-1} R (R' S_{zz}^{-1} R)^{-1} R' S_{zz}^{-1} S_{z\varepsilon} / \hat{\sigma}^2.$$

Write the denominator as,

$$\begin{gathered} \sqrt{T} S_{\varepsilon z} N_T (N_T S_{zz} N_T)^{-1} N_T R (R' N_T (N_T S_{zz} N_T)^{-1} N_T R)^{-1} \times \\ R' N_T (N_T S_{zz} N_T)^{-1} N_T S_{z\varepsilon} \sqrt{T} \\ = \sqrt{T} S_{\varepsilon z} N_T (N_T S_{zz} N_T)^{-1} R (R' (N_T S_{zz} N_T)^{-1} R)^{-1} R' (N_T S_{zz} N_T)^{-1} N_T S_{z\varepsilon} \sqrt{T} \end{gathered}$$

where it has been used that $N_T' R = \frac{1}{\sqrt{T}} R$ as $R' = (1, 0)$. The just applied arguments gives that this converge in distribution to,

$$\int \mathcal{B} d\mathcal{B} \left(\int \mathcal{B}^2 du \right)^{-1} \int \mathcal{B} d\mathcal{B} \overset{D}{=} \sigma^2 \int \mathcal{W} d\mathcal{W} \left(\int \mathcal{W}^2 du \right)^{-1} \int \mathcal{W} d\mathcal{W}$$

with $\mathcal{W} = \mathcal{B}/\sigma$. Using that $\hat{\sigma}^2$ is consistent, the result in (III.49) follows as $W_T = O_P(T^{-1})$ and the usual Taylor expansion of $\log(1 + w)$ can be applied. □

III.4.3 The AR(k) model

Next turn to the AR(k) model which will only briefly be discussed since the main arguments have been given in connection with the discussion of the AR(2) model. The AR(k) model is given by,

$$x_t = \rho_1 x_{t-1} + \ldots + \rho_k x_{t-k} + \varepsilon_t, \; t = 1, ..., T \tag{III.51}$$

ε_t i.i.d.N$(0, \sigma^2)$ and $x_0, ..., x_{-k}$ fixed. To anticipate the unit root analysis reparametrize (III.51) as

$$\Delta x_t = \pi x_{t-1} + \gamma_1 \Delta x_{t-1} + \ldots + \gamma_{k-1} \Delta x_{t-k+1} + \varepsilon_t$$

where $\pi = -A(1) = \sum_{i=1}^{k} \rho_i - 1$ and $\gamma_i = -\sum_{j=i+1}^{k} \rho_j$. Thus the characteristic polynomial takes the form

$$A(z) = (1 - z) - \pi z - \gamma_1 (1 - z) z - \ldots - \gamma_{k-1} (1 - z) z^{k-1}$$

which has a unit root at $z = 1$ if and only if $\pi = 0$.

III.4.3.1 Representation for the AR(k)

Suppose that $\pi = 0$, or equivalently,

$$\Delta x_t = \gamma_1 \Delta x_{t-1} + .. + \gamma_{k-1} \Delta x_{t-k+1} + \varepsilon_t$$

Set $\gamma(z)$ equal to the characteristic polynomial of Δx_t, that is

$$\gamma(z) \equiv 1 - \gamma_1 z - ... - \gamma_{k-1} z^{k-1}, \tag{III.52}$$

and $A(z) = (1 - z)\gamma(z)$ when $\pi = 0$. Then if $|\gamma(z)| = 0$ implies $|z| > 1$, $(\Delta x_t, ..., \Delta x_{t-k+1})$ is geometrically ergodic. In particular, Δx_t admits a stationary representation of the form,

$$\Delta x_t^* = \theta(L) \varepsilon_t \tag{III.53}$$

where $\theta(z) = \gamma(z)^{-1} = \sum_{i=0}^{\infty} \theta_i z^i$, with θ_i given in Theorem II.9.

Similar to the AR(2) case the following result holds:

Theorem III.4.2 *Let x_t be an AR(k) process given by (III.51). Assume that $A(z)$ has one, and only one, unit-root, while the remaining roots, that is the roots of $\gamma(z)$ in (III.52), are larger than one in absolute value. Then x_t is an I(1) process, with representation,*

$$x_t = \phi \sum_{i=1}^{t} \varepsilon_t + \lambda' S_t + \lambda_0, \tag{III.54}$$

where $\phi = \gamma(1)^{-1}$, $S_t = (\Delta x_t, ..., \Delta x_{t-k+1})'$ is geometrically ergodic, and $\lambda' S_t$ is a linear combination of S_t. The constant λ_0 depends on initial values of the AR(k) process and is given by $\lambda_0 = \lambda' S_0 + x_0$.

The vector $\lambda \in \mathbb{R}^{k-1}$ is given by $\lambda' = \phi\left(\gamma_0^*, ..., \gamma_{k-2}^*\right)$, where $\gamma_0^* = 1 - \gamma(1)$, and $\gamma_i^* = \gamma_{i-1}^* - \gamma_i$ for $i = 1, .., k-2$.

Thus under the assumption of a unit root, and the additional assumption that the remaining roots are outside the unit circle, x_t has a representation as a random walk plus a linear combination of a geometrically ergodic process. The LLN applies to the $\gamma' S_t$ term, see Lemma III.4.1, and the FCLT apply to the random walk part such that,

$$\frac{1}{\sqrt{T}} x_{[T\cdot]} \overset{D}{\to} \phi \mathcal{B}, \tag{III.55}$$

where $\mathcal{B}$ is a Brownian motion on [0,1] with variance σ^2.

Hence testing for a unit root in the AR(k) model, that is $\pi = 0$, is equivalent to testing if the autoregressive process is an I(1) process provided that the remaining roots are larger than one in absolute value.

Proof:

The arguments are analogous to the AR(2) case. Expand $\gamma(z)$ as follows,

$$\gamma(z) = \gamma(1) + \gamma^*(z)(1-z), \tag{III.56}$$

where $\gamma^*(z) = (\gamma(z) - \gamma(1))/(1-z)$ is a polynomial of order $k-2$, $\gamma^*(z) = \gamma_0^* + \gamma_1^* z + ... + \gamma_{k-2}^* z^{k-2}$. Simple identification of coefficients show that γ_i^* and γ_i are related by

$$\gamma_0^* = 1 - \gamma(1), \quad \gamma_i^* = \gamma_{i-1}^* - \gamma_i \text{ for } i = 1, 2, ..., k-2. \tag{III.57}$$

Use this to rewrite $\gamma(L)\Delta x_t = \varepsilon_t$ as,

$$\gamma(L)\Delta x_t = \gamma(1)\Delta x_t + \gamma^*(L)\Delta^2 x_t = \varepsilon_t. \tag{III.58}$$

Next, divide by $\gamma(1)$, and consider $\sum_{i=1}^t \Delta x_i$ which, with $\phi = 1/\gamma(1)$, equals,

$$x_t = x_0 + \phi(\sum_{i=1}^t \varepsilon_i + \gamma^*(L)(\Delta x_t - \Delta x_0)). \tag{III.59}$$

Now, $\gamma^*(L)\Delta x_t = \gamma_0^* \Delta x_t + ... + \gamma_{k-2}^* \Delta x_{t-k+2}$, and hence $S_t \equiv \phi\gamma^*(L)\Delta x_t$ is a linear combination of the geometrically ergodic process $(\Delta x_t, ..., \Delta x_{t-k+2})'$ such that the LLN applies to S_t. Likewise, S_0 is a linear combination of the initial values $(\Delta x_0, ..., \Delta x_{-k+2})'$. □

III.4.3.2 Testing for a unit-root

The assumption of a unit root in the AR(k) model in (III.51) is equivalent to the assumption that

$$H_0: \ \pi = 0. \tag{III.60}$$

The analysis of the AR(k) model is identical to the analysis of the AR(2) model, with $Y_t \equiv \Delta x_t$, $\beta' = (\pi, \gamma')$, where $\gamma' \equiv (\gamma_1, .., \gamma_{k-1})$ and $Z_t = (x_{t-1}, \Delta x_{t-1}, ..., \Delta x_{t-k+1})'$. Using Theorem III.4.2, mimicking the proof for the AR(2) case of Theorem III.4.1 and noting that Proposition III.4.1 generalizes immediately to the AR(k) case, the following holds:

Theorem III.4.3 *Consider the AR(k) model in (III.51) with ML estimators given by,*

$$\hat{\beta}' = (\hat{\pi}, \hat{\gamma}') = S_{yz}S_{zz}^{-1}, \quad \text{and } \hat{\sigma}^2 = S_{yy\cdot z}. \tag{III.61}$$

where z, y in the product moments refer to $Z_t = (x_{t-1}, \Delta x_{t-1}, ..., \Delta x_{t-k+1})'$ and $Y_t = \Delta x_t$ respectively.

When $\pi = 0$, and $|\gamma(z)| = 0$ implies $|z| > 1$, $\hat{\beta}$ and $\hat{\sigma}^2$ are consistent, that is $\hat{\pi} \xrightarrow{P} 1, \hat{\gamma} \xrightarrow{P} \gamma$ and $\hat{\sigma}^2 \xrightarrow{P} \sigma^2$ as $T \to \infty$. Moreover,

$$(T\hat{\pi}, \sqrt{T}(\hat{\gamma} - \gamma)') \xrightarrow{D} (\phi^{-1} \int_0^1 \mathcal{W} d\mathcal{W} / \int_0^1 \mathcal{W}_u^2 du, \ N_{k-1}(0, V)) \tag{III.62}$$

where $\mathcal{W}$ is a standard Brownian motion and $V = V(S_t^*)$, see Theorem III.4.2.

The LR statistic for the hypothesis $H_0 : \pi = 0$, or with $R'\beta = (1, 0)\beta = \pi = 0$, is given by,

$$LR(\pi = 0) = T \log(1 + W_T), \ \ \text{where} \tag{III.63}$$

$$W = TW_T = TS_{yz}S_{zz}^{-1}R(R'S_{zz}^{-1}R)^{-1}R'S_{zz}^{-1}S_{zy}/\hat{\sigma}^2, \tag{III.64}$$

When $\pi = 0$, and $|\gamma(z)| = 0$ implies $|z| > 1$, the LR statistic is asymptotically Dickey-Fuller type distributed,

$$LR(\pi = 0) \xrightarrow{D} (\int_0^1 \mathcal{W} d\mathcal{W})^2 / \int_0^1 \mathcal{W}_u^2 du \quad \text{as } T \to \infty. \tag{III.65}$$

III.5 The role of deterministic terms

Consider the simple AR(1) model with a constant regressor

$$\Delta x_t = \pi x_{t-1} + \mu + \varepsilon_t$$

Under the assumption $\pi = 0$,

$$x_t = x_0 + \sum_{i=1}^{t} \varepsilon_i + \mu t, \tag{III.66}$$

whereas, if $\pi \in (-2, 0)$ or equivalently $|\rho| < 1$,

$$x_t^* = \sum_{i=0}^{\infty} \rho^i \varepsilon_{t-i} + \frac{\mu}{1-\rho}. \tag{III.67}$$

Thus testing the assumption of $\pi = 0$ against the alternative $\pi \in (-2, 0)$ essentially tests if x_t is a random walk with drift against it being a stationary process with a constant mean. This lack of deterministic 'balance' shows up in the limit distribution of the likelihood ratio test which will have nuisance parameters in the sense that there are two different limit distributions depending on whether or not $\mu = 0$. Apart from this in most practical situations such as the analysis of the US-GNP below it is of interest to test if the process is trend-stationary against it being a random walk with 'stationary noise'. This can be accomplished by first testing whether the process is I(1) in a model which allows a deterministic trend both under and outside the alternative. Next, the determination of whether there is a linear trend or not, can be done by the standard χ^2 distributed likelihood ratio test statistics.

Similar considerations lead as in Dickey and Fuller (1979) to consider the following AR(k) model when analyzing the I(1) hypothesis in the case of deterministic linear trends,

$$\Delta x_t = \pi_1 x_{t-1} + \pi_2 t + \gamma_1 \Delta x_{t-1} + \ldots + \gamma_{k-1} \Delta x_{t-k-1} + \mu + \varepsilon_t$$

where the hypothesis of interest is,

$$H: \ \pi_1 = \pi_2 = 0.$$

Under this hypothesis it follows, under the assumption of Theorem III.4.2, that

$$x_t = \phi \sum_{i=1}^{t} \varepsilon_i + \phi \mu t + \lambda' S_t + a + c,$$

where c is a constant. That is,x_t is an I(1) process with a linear trend. Under the alternative hypothesis that all roots of the characteristic polynomial are outside the unit circle, x_t is trend-I(0) and has a representation as,

$$x_t^* = \theta(L)\varepsilon_t + \theta(1)\mu t + \text{constant}$$

i.e. trend-stationary or a stationary process with a linear trend. Having determined whether or not $\pi_1 = \pi_2 = 0$ a successive test of the presence of a linear trend is as mentioned the usual χ^2 test. Note that if x_t is I(1) this is a test of $\mu = 0$, whereas if x_t is trend-stationary it is a test of $\pi_2 = 0$.

The likelihood ratio test statistic of the hypothesis, $\pi_1 = \pi_2 = 0$ has the same form as before, and is based on OLS regression of Δx_t on $(x_{t-1}, t)'$ corrected for a constant and the lagged differences, Δx_{t-i}. The limit distribution of the likelihood ratio test statistic is in this case given by

$$\mathrm{LR}(\pi_1 = \pi_2 = 0) \overset{D}{\rightarrow} (\int_0^1 F d\mathcal{W})'(\int_0^1 F_u F_u')^{-1} \int_0^1 F d\mathcal{W},$$

where $\mathcal{W}$ is a standard Brownian motion and the two-dimensional process F is given by

$$F_u = \begin{pmatrix} \mathcal{W}_u - \int_0^1 \mathcal{W}_s ds \\ u - \int_0^1 sds \end{pmatrix}.$$

This distribution is tabulated in Johansen (1996) and some quantiles are reported below.

Mimicking the ideas above, the model which allows for a constant level is given by

$$\Delta x_t = \pi_1 x_{t-1} + \pi_2 + \gamma_1 \Delta x_{t-1} + \ldots + \gamma_{k-1} \Delta x_{t-k-1} + \varepsilon_t.$$

In this case, under the hypothesis of $\pi_1 = \pi_2 = 0$, the likelihood ratio test statistic has a limit distribution as above, but with $F_u = (\mathcal{W}_u, 1)$.

III.5.1 Quantiles for LR testing

Summarizing the discussion above three different models were of interest. With the notation $V_t = (\Delta x_{t-1}, .., \Delta x_{t-k+1})'$ and $\gamma' = (\gamma_1, .., \gamma_{k-1})$ these can be rewritten as,

$$\begin{aligned} H_0 &: \ \Delta x_t = \pi x_{t-1} + \gamma' V_t + \varepsilon_t \\ H_1 &: \ \Delta x_t = (\pi_1, \pi_2)(x_{t-1}, 1)' + \gamma' V_t + \varepsilon_t \\ H_2 &: \ \Delta x_t = (\pi_1, \pi_2)'(x_{t-1}, t)' + \gamma' V_t + \mu + \varepsilon_t \end{aligned}$$

The hypotheses of interest are $H_0^* : \pi = 0$, $H_1^* : \pi_1 = \pi_2 = 0$ and $H_2^* : \pi_1 = \pi_2 = 0$ respectively.

For $i = 0, 1, 2$, the limit distributions of the likelihood ratio test statistics of H_i^* against H_i are given by

$$\int d\mathcal{W} F (\int FF du)^{-1} \int F d\mathcal{W} \qquad \text{(III.68)}$$

under the assumptions of Theorem 4.3. Here $\mathcal{W}$ is a standard Brownian motion and F takes the forms,

$$\begin{aligned} &H_0^* : F_u = \mathcal{W}_u \\ &H_1^* : F_u = (\mathcal{W}_u, 1)' \\ &H_2^* : F_u = (\mathcal{W}_u - \int \mathcal{W}_s ds, u - 1/2)'. \end{aligned}$$

The quantiles of $(III.68)$ given below are from Johansen (1996). For comparison also the quantiles of the χ^2 distributions with 1 and 2 degrees of freedom are given.

Quantiles of the Likelihood Ratio Tests for Unit Roots

Hypothesis	95% quantile	97.5 % quantile	99 % quantile
H_0^*	4.2	5.3	7.0
H_1^*	9.1	10.7	12.7
H_2^*	12.4	14.1	16.4
χ_1^2	3.84	5.02	6.64
χ_2^2	6.0	7.4	9.21

III.5.2 The US-GNP example

A preliminary analysis of log(US-GNP) quarterly data from the period 1959:3 – 1996:4 indicates that an AR(6) model with a linear trend describe well the dynamics. This is in accordance with the economic literature on growth. But it is also found that one of the roots in the characteristic polynomial is close to one, $z = 0.916$.

The interpretation of μ changes dramatically depending on whether or not there is a unit root. Write the AR(6) model to anticipate the I(1) analysis as

$$\Delta x_t = \pi_1 x_{t-1} + \pi_2 t + \gamma_1 \Delta x_{t-1} + .. + \gamma_5 \Delta x_{t-5} + \mu + \varepsilon_t \qquad \text{(III.69)}$$

The reported values are

$$\begin{array}{rrr} \text{Parameter} & \text{Estimate} & t\text{-value.} \\ \pi_1 & -0.055 & -3.05 \\ \pi_2 & 0.004 & 2.74 \end{array} \qquad \text{(III.70)}$$

Clearly, based on the discussion so far, it is not clear how the reported t-values should be interpreted if indeed "0.916 = 1" and $\pi_2 = 0$.

The likelihood ratio test statistic of the hypothesis $\pi_1 = \pi_2 = 0$ equals,

$$\text{LR}(\pi_1 = \pi_2 = 0) = 14.3 \qquad \text{(III.71)}$$

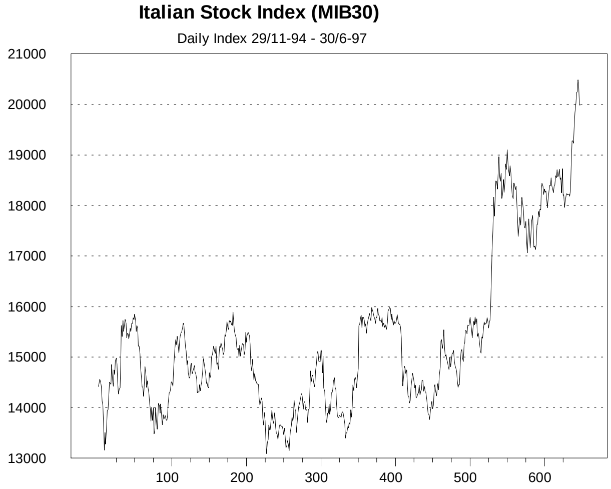
which according to the χ^2_2 distribution is clearly rejected. However, it is not the χ^2 distribution but the Dickey-Fuller type distribution that should be used with a 95% quantile of 12.4 and a 97.5% quantile of 14.13. Hence it is not clear if the hypothesis should be rejected. Thus a model for log(US-GNP) is a random walk with drift.

To test whether indeed the drift is present is very simple, since this is a simple hypothesis in the stationary model of $\Delta \log$(US-GNP). The reported t-value is 2.6 and, based on the normal distribution, the drift is present.

On the other hand if one rejects the unit root then it is of interest to test if the linear trend is present in the stationary model. Clearly the trend plays a significant role since the $t-$value is 2.7.

III.5.3 Stock market index

Consider the log of Italian stock market index, x_t in figure below and compare with the simulated realization of the random walk. It appears that there is a random walk part in the series.



Indeed an overly simplified implication of the efficient market hypothesis is that the log of stock prices follow a random walk with a drift. In particular, in the basic Black-Scholes set-up,

$$dx_u = \mu du + d\mathcal{B}_u \tag{III.72}$$

with $\mathcal{B}_u$ a Brownian motion with variance σ^2, which in discrete time can be represented as

$$\Delta x_t = \mu + \varepsilon_t \tag{III.73}$$

To see if this describes the variation in the data an AR(4) model was fitted ,

$$\Delta x_t = (\pi_1, \pi_2)(x_{t-1}, t)' + \gamma_1 \Delta x_{t-1} + \ldots + \gamma_3 \Delta x_{t-3} + \mu + \varepsilon_t \tag{III.74}$$

to the log of the Italian MIB30 index.

One finds that indeed there is a root in the characteristic polynomial of 0.98 (and the remaining roots outside the unit circle) and that

$$\mathrm{LR}(\pi_1 = \pi_2 = 0) = 5.6 \tag{III.75}$$

and hence based on the 95 % quantile of the Dickey-Fuller type distribution this is accepted. Hence, we maintain the hypothesis that x_t has a drift and a random walk part.

Note in particular that the misspecification test for ARCH is significant which will be explored further when discussing ARCH models. Thus the AR model does not describe fully the variation in the data and a different model is needed, see later. Also the above analysis is based on the assumption of no ARCH and hence is by no means a "valid" analysis.

References

[1] Chan & Wei, 1988, ”Limiting Distributions of Least Squares Estimates of Unstable Autoregressive Processes”, Annals of Math. Statistics

[2] Billingsley, 1968, ” Convergence of Probability Measures”, Wiley.

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

[4] Dickey and Fuller, 1979, ”Distributions of the Estimators for Autoregressive Time Series with a Unit Root”, JASA.

[5] Hansen, B., 1992, ”Convergence to Stochastic Integrals for Dependent Heterogenous Processes”, Econometric Theory, 8:489-500

[6] Johansen, S., 1996, "Likelihood-Based Inference in Cointegrated Vector Autoregressive Models", Oxford University Press

Anders Rahbek
Rasmus Søndergaard Pedersen
University of Copenhagen

September 2024

Part IV
Cointegration Analysis in Vector Autoregressions

IV.1 Introduction

The literature on cointegration is by now enormous, see for example the survey by Johansen (2005), and cointegration analysis is applied everywhere in applied and theoretical time series analysis. This part gives a brief introduction to the rich theory of cointegration analysis in VAR models, with emphasis on analysis of the cointegrated VAR(1) model.

IV.1.1 Futures and no arbitrage

Let K denote the expiry date of a futures contract at time t on the index with spot price S_t, and $F(t,K)$ the price of the future contract. Considering forward pricing as equivalent to futures pricing, the condition for no arbitrage from financial theory can be written as the identity,

$$F(t,K)=S_t\exp(r_{t,K}(t-K))$$

where $r_{t,K}$ is the zero coupon rate for the period between t and expiry date K. With $f_t=\log(F(t,K)/r_{t,K}(t-K))$ and $s_t=\log(S_t)$, the identity reduces to,

$$f_t-s_t=0.$$

While often applied, the assumption of equivalence of future and forward pricing is actually not correct, essentially due to the different types of settlements of the contracts. It is therefore of interest to see to what extend it may be correct, and to find out if there is an empirical relationship between s_t and f_t.

Italian MIB30 daily data are shown below for the period 29/11-94 - 30/6-97. Previously it was concluded that the spot price s_t seemed to be modelled well by a random walk plus drift type model, that is, as an I(1) process with

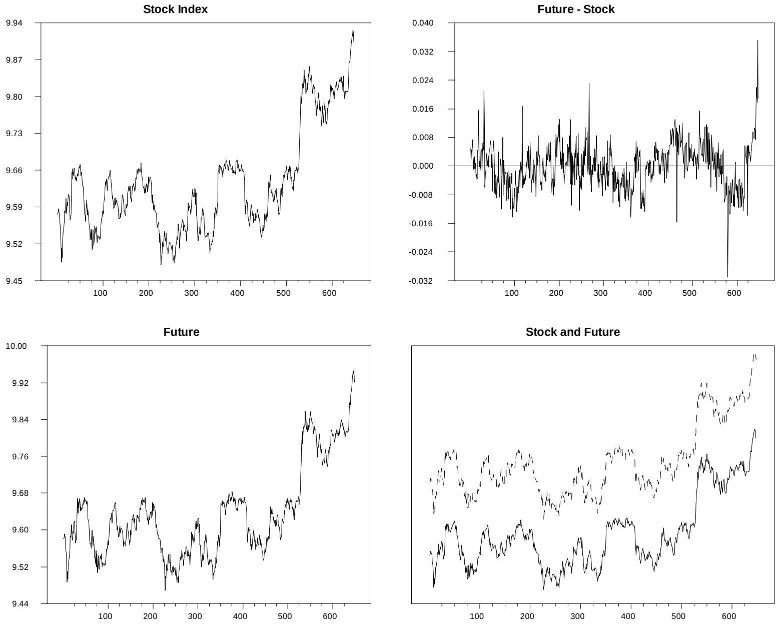


Figure 1: The time series f_t, s_t and $f_t - s_t$.

a linear drift, ignoring the ARCH type misspecification. Hence a VAR model for $X_t = (s_t, f_t)'$ must allow for unit root(s), and, at the same time, for 'identities', or 'stable relations', such as the above. In terms of cointegration, the identity or cointegration relation, is interpreted as $s_t - f_t$ being asymptotically stable or stationary, as opposed to I(1). The series f_t and s_t are shown in Figure 1 below, together with the cointegrating relation as given by the linear combination,

$$\beta' X_t = (1, -1) X_t = f_t - s_t. \tag{IV.1}$$

Clearly, both s_t and f_t resemble random walks, while the cointegrating relation, $\beta' X_t$, seems more stable.

More can be seen by considering the cross-plots of s_t and f_t in Figure 2. As expected the series are clearly correlated. Furthermore, they seem to be pushed, or moved, up and down around the straight line $s_t = f_t$. Essentially, if $|s_t - f_t| > 0$, then $|s_{t+1} - f_{t+1}| < |s_t - f_t|$ as illustrated for a part of the sample, $t = 40, ..., 50$. Thus the 'error' at time t, as measured by $s_t - f_t$, is being corrected such that at time $t + 1$, the spread, $s_{t+1} - f_{t+1}$, is 'smaller'.

Thus a VAR model for X_t must allow for I(1) type non-stationary (integrated) variables with a stable or 'cointegrated' linear combination, $\beta' X_t$. At the same time the dynamics of the model must be such that ΔX_{t+1} can adjust to $\beta' X_t$, that is, it must allow error correction.

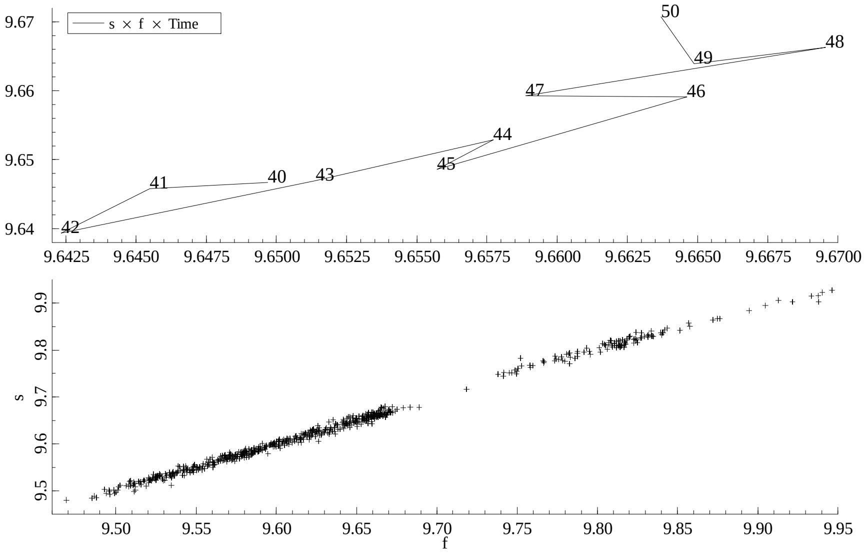


Figure 2: Cross-plots of s_t and f_t for $t = 1, ..., 648$ and $t = 40, ..., 50$.

The cointegrated VAR model allows exactly for this as will be shown below. Moreover, inference in the VAR model makes it possible to determine the number of stable or cointegrated relations β, and also to do hypothesis testing on these.

IV.2 VAR(1) model

Consider the p-dimensional VAR model of order one, VAR(1), model as given by,

$$X_t = AX_{t-1} + \varepsilon_t, \quad t = 1, 2, ..., T \tag{IV.2}$$

with $A \in \mathbb{R}^{p \times p}$, X_0 fixed and ε_t i.i.d. $\mathrm{N}_p(0, \Omega)$ distributed, $\Omega > 0$.

Recall the condition $\rho(A) < 1$ which implies X_t in (IV.2) is geometrically ergodic. The condition may equivalently be stated in terms of the roots of the so-called characteristic polynomial, $A(z) = I - Az$, $z \in \mathbb{C}$,

$$\det(A(z)) = \det(I_p - Az) = 0 \;\Rightarrow\; |z| > 1.$$

Allowing eigenvalues of A at one implies $\rho(A) = 1$, and can be stated in terms of $A(z)$ as

$$\det(A(1)) = \det(I_p - A) = 0.$$

That is, with

$$\Pi = A - I_p,$$

$A(z)$ has one or more roots at $z = 1$ if, and only if, the $(p \times p)$ dimensional matrix Π has reduced rank $r < p$. This is central to the formulation of cointegration.

One may reparameterize the VAR(1) model in terms Π, using $\Delta X_t = X_t - X_{t-1}$,

$$\Delta X_t = \Pi X_{t-1} + \varepsilon_t, \quad t = 1, ..., T, \tag{IV.3}$$

where $\Pi \in \mathbb{R}^{p \times p}$, X_0 is fixed and ε_t are i.i.d.$\mathrm{N}_p(0, \Omega)$.

We denote the hypothesis that Π has rank less than or equal to r by H_r, where $0 \leq r \leq p$. It follows that under H_r the $p \times p$ matrix Π can be factorized as

$$\Pi = \alpha\beta',$$

where α, β are $p \times r$ matrices, see Lemma B.1 in the appendix. The following two examples illustrate this:

Example IV.2.1 *Consider the case of a 2×2 matrix Π of rank $r = 1$, given by*

$$\Pi = \begin{pmatrix} 1 & -1 \\ 1 & -1 \end{pmatrix}.$$

Simple calculations show,

$$\Pi = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \begin{pmatrix} 1 \\ -1 \end{pmatrix}' = \alpha\beta'.$$

Note that β, and hence α, are unique up to a normalization as

$$\Pi = (\alpha m)(\beta m^{-1})',$$

for any $m \in \mathbb{R}$, $m \neq 0$. Stated differently, α spans the column space of Π, $\mathrm{sp}(\Pi)$, while β spans the row space of Π, $\mathrm{sp}(\Pi')$.

Example IV.2.2 *Consider next the case of a 3×3 matrix Π of rank $r = 2$,*

$$\Pi = \begin{pmatrix} 1 & 1 & 1 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ -1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}' = \alpha\beta',$$

where as before, $\mathrm{sp}(\alpha) = \mathrm{sp}(\Pi)$, $\mathrm{sp}(\beta) = \mathrm{sp}(\Pi')$. In particular, $\Pi = (\alpha m')(\beta m^{-1})'$, with m any 2×2 dimensional invertible matrix.

IV.2.1 Stochastic behavior of X_t

As illustrated, the fact that the characteristic polynomial has one or more roots at $z = 1$, implies that the matrix parameter Π has reduced rank, and hence that $\Pi = \alpha\beta'$ for some $(p \times r)$-dimensional matrices α and β. For α and β of full rank r, we define their orthogonal complements $\alpha_\perp$ and $\beta_\perp$. These are $(p \times (p - r))$-dimensional matrices of full rank $(p - r)$, for which $\alpha'_\perp\alpha = \beta'_\perp\beta = 0$, $\det(\alpha, \alpha_\perp) \neq 0$, and $\det(\beta, \beta_\perp) \neq 0$.

Consider next what this implies for the stochastic behavior of X_t.

Example IV.2.3 *Consider the* 2*-dimensional VAR(1) process* X_t *as given by,*

$$\Delta X_t = \alpha\beta' X_{t-1} + \varepsilon_t, \tag{IV.4}$$

where $\alpha = (-1, 0)'$, $\beta' = (1, -1)$ *and* $\varepsilon_t = (\varepsilon_{1t}, \varepsilon_{2t})'$ *i.i.d.*$N_2(0, \Omega)$. *Note initially that, with* $\alpha_\perp = (0, 1)'$,

$$\alpha'_\perp \Delta X_t = \Delta X_{2t} = \alpha'_\perp (\alpha\beta' X_{t-1} + \varepsilon_t) = \alpha'_\perp \varepsilon_t = \varepsilon_{2t}.$$

That is, $X_{2t} = \sum_{i=1}^t \varepsilon_{2i} + X_{20}$. *Equivalently,* X_{2t} *is the sum of a random walk and the initial value, and* X_{2t} *is an I(1) process.*

Next, note that

$$\beta' X_t = X_{1t} - X_{2t} = \beta'\varepsilon_t = \varepsilon_{1t} - \varepsilon_{2t},$$

that is, the difference, or 'spread', between X_{1t} *and* X_{2t} *is i.i.d. Gaussian, and therefore, in particular, geometrically ergodic with a stationary solution (often simply referred to as stationary).*

Collecting terms, we find

$$\begin{pmatrix} X_{1t} - X_{2t} \\ X_{2t} \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} X_t = \begin{pmatrix} \varepsilon_{1t} - \varepsilon_{2t} \\ \sum_{i=1}^t \varepsilon_{2i} + X_{20} \end{pmatrix},$$

or, re-organizing terms,

$$\begin{aligned} X_t = \begin{pmatrix} X_{1t} \\ X_{2t} \end{pmatrix} &= \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}^{-1} \begin{pmatrix} \varepsilon_{1t} - \varepsilon_{2t} \\ \sum_{i=1}^t \varepsilon_{2i} + X_{20} \end{pmatrix} \\ &= \begin{pmatrix} 1 \\ 1 \end{pmatrix} \sum_{i=1}^t \varepsilon_{2t} + \begin{pmatrix} 1 \\ 0 \end{pmatrix} (\varepsilon_{1t} - \varepsilon_{2t}) + \begin{pmatrix} 1 \\ 1 \end{pmatrix} X_{02}. \end{aligned}$$

As $\beta_\perp = (1, 1)'$, *this may be stated as.*

$$X_t = \beta_\perp \alpha'_\perp \sum_{i=1}^t \varepsilon_t + \alpha\beta'\varepsilon_t + \beta_\perp\alpha'_\perp X_0. \tag{IV.5}$$

Thus X_t has a representation as a sum of a "common trend" (i.e. the random walk $\alpha_\perp' \sum_{i=1}^t \varepsilon_t$), a "stationary" (geometrically ergodic) process, $\beta'\varepsilon_t$, and the initial value X_0. Hence X_t is a non-stationary I(1) process.

Note that, in line with the mentioned example of spot and futures rates, where $X_t = (s_t, f_t)'$, X_t is error correcting to $\beta' X_{t-1}$ with adjustment vector α, which can be seen by rewriting (IV.4) as,

$$\Delta X_t = \alpha\beta' X_{t-1} + \varepsilon_t. \tag{IV.6}$$

Simulating a process with these parameter values will give a cross plot of the form in Figure 2. More precisely, by the first factor $\beta_\perp \alpha_\perp' \sum_{i=1}^t \varepsilon_t$ in (IV.5), the points (X_{1t}, X_{2t}) will be moved up and down by $\alpha_\perp' \sum_{i=1}^t \varepsilon_t$ on the attractor given by $\beta' X_t = 0$, or along the line spanned by $\beta_\perp = (1, -1)'$. The movements away from, or around, the attractor, as given by $\alpha\beta'\varepsilon_t$, are then error corrected by the term $\alpha\beta' X_{t-1}$ in (IV.6).

To generalize Example IV.2.3 to the case where the cointegrating relations $\beta' X_t$ are not i.i.d. the following assumption plays a key role throughout:

Assumption IV.2.1 *Let X_t be a vector autoregressive process with characteristic polynomial $A(z)$, $z \in \mathbb{C}$. Assume that $A(z)$ has exactly $(p - r)$ roots at $z = 1$ and the remaining roots are larger than one in absolute value, $|z| > 1$.*

Remark IV.2.1 *The condition may equivalently be stated for the VAR(1) process as A has $(p - r)$ eigenvalues equal to one, while the remaining eigenvalues are smaller than one in absolute value, $\rho(I_r + \beta'\alpha) < 1$; see also the proof of Proof of Theorem IV.2.1.*

One has the following theorem which is a version of the so-called 'Granger representation theorem', see also Johansen (1996).

Theorem IV.2.1 *Consider the VAR(1) process given by (IV.3) under the hypothesis*

$$H_r : \Pi = \alpha\beta', \quad \alpha, \beta \in \mathbb{R}^{p\times r}. \tag{IV.7}$$

Under Assumption IV.2.1, X_t is an I(1) process, and has the representation,

$$X_t = C\sum_{i=1}^t \varepsilon_i + C_S S_t + C_0, \tag{IV.8}$$

where $C = \beta_\perp (\alpha_\perp' \beta_\perp)^{-1} \alpha_\perp'$ is a $p\times p-$dimensional matrix of rank $(p-r)$, and $C_S = \alpha(\beta'\alpha)^{-1}$. Moreover, C_0 depends on the initial values and is given by $C_0 = CX_0$. The r cointegrating relations $\beta' X_t = S_t$ are geometrically ergodic. In particular, the initial values $\beta' S_0$ can be given an initial distribution such that S_t has the stationary representation, $S_t^* = \sum_{i=0}^\infty (I + \beta'\alpha)^i \beta'\varepsilon_i$.

Remark IV.2.2 *Recall that by the geometric ergodicty, the LLN and CLT apply to functions of $\{S_t\}$ as will be used throughout.*

The theorem states that when Π has rank r, and if there are exactly $(p-r)$ unit roots, and the remaining roots correspond to the asymptotically stable case, the p dimensional process X_t is non-stationary and I(1). Moreover, it has r cointegrating relations, $\beta' X_t$, which are geometrically ergodic.

Assumption IV.2.1 is indeed vital for the theorem to hold. For example, if there are exactly $p-r$ roots at $z=1$, but all, or some of, the remaining roots are smaller than one in absolute value, $\beta' X_t$ is an explosive process, and hence not 'cointegrating'. Also, by allowing for more than $p-r$ roots at $z=1$, the process may be integrated of order two, instead of one. *In other words, the assumption of reduced rank r of Π is a necessary, but not sufficient, condition for the VAR process to be an I(1) process which is cointegrated.*

Note also that under H_p, the theorem states that X_t is asymptotically stable, provided that all roots of the characteristic polynomial are larger than one in absolute value, which is in accordance with the previous theory for geometrically ergodic VAR processes.

Proof of Theorem IV.2.1: The arguments are identical to the ones used in Example IV.2.3: Note first, that by definition, $\alpha'_\perp \Delta X_t = \alpha'_\perp \varepsilon_t$, and hence, $\alpha'_\perp X_t = \alpha'_\perp \sum_{i=1}^t \varepsilon_i + \alpha'_\perp X_0$. Next,

$$\beta' X_t = \left(I_r + \beta'\alpha\right) \beta' X_{t-1} + \beta' \varepsilon_t, \tag{IV.9}$$

which is geometricallty ergodic provided $\rho\left(I_r + \beta'\alpha\right) < 1$. This holds by Assumption IV.2.1, as

$$\det\left(A\left(z\right)\right) = 0 \Leftrightarrow \det\left(I_r - (I_r + \beta'\alpha)z\right) \det\left(\left(1-z\right) I_{p-r}\right) = 0, \tag{IV.10}$$

which follows by pre- and post multiplying $A(z)$ by $(\beta, \beta_\perp)'$ and $(\beta, \beta_\perp)$ respectively.

The final result holds by using the (skew-projection) identity

$$I_p = \alpha\left(\beta'\alpha\right)^{-1}\beta' + \beta_\perp(\alpha'_\perp\beta_\perp)^{-1}\alpha'_\perp = C_S\beta' + C, \tag{IV.11}$$

such that X_t has the decomposition,

$$X_t = CX_t + C_S\beta' X_t = CX_t + C_S S_t, \tag{IV.12}$$

as claimed. That (IV.11) holds, follows by $\beta'\alpha$ having full rank r as implied by (IV.10). The full rank of $\beta'\alpha$ implies that $\alpha'_\perp\beta_\perp$ has full rank $p-r$, and that $(\beta, \alpha_\perp)$ has full rank. Multiplying in (IV.11) by $(\beta, \alpha_\perp)'$ from the left establishes the identity. □

IV.2.2 Econometric analysis

Next, turn to ML estimation of the parameters in the VAR(1) model in (IV.3) under H_r, given by the equations,

$$\Delta X_t = \Pi X_{t-1} + \varepsilon_t,$$
$$H_r : \Pi = \alpha\beta'.$$

In terms of the usual regression model with $Y_t = \Delta X_t$ and $Z_t = X_{t-1}$, the hypothesis H_r of reduced rank of Π is a nonlinear restriction, and the previous linear regression results cannot be used. Instead the following result holds:

Theorem IV.2.2 *Under H_r, the ML estimators of α, β and Ω are given by,*

$$\hat{\alpha} = S_{yz}\hat{\beta}(\hat{\beta}'S_{zz}\hat{\beta})^{-1}, \tag{IV.13}$$

$$\hat{\Omega} = S_{yy\cdot\hat{\beta}} = S_{yy} - S_{yz}\hat{\beta}(\hat{\beta}'S_{zz}\hat{\beta})^{-1}\hat{\beta}'S_{zy}. \tag{IV.14}$$

Here $Y_t = \Delta X_t$ and $Z_t = X_{t-1}$ in the product moment matrices, such that e.g. $S_{yz} = T^{-1}\sum_{t=1}^{T} Y_t Z_t$. The MLE $\hat{\beta}$ is found by solving the eigenvalue problem

$$\det\left(\lambda S_{zz} - S_{zy}S_{yy}^{-1}S_{yz}\right) = 0, \tag{IV.15}$$

with eigenvalues $1 > \hat{\lambda}_1 > ... > \hat{\lambda}_r > ...\hat{\lambda}_p > 0$ and corresponding eigenvectors $\hat{V} = (\hat{v}_1, ..., \hat{v}_p)$, for which $\hat{V}'S_{zz}\hat{V} = I_p$ and $\hat{V}'S_{zy}S_{yy}^{-1}S_{yz}\hat{V} = \hat{\Lambda}$, where $\hat{\Lambda} = diag\left(\hat{\lambda}_1, ..., \hat{\lambda}_p\right)$. Then

$$\hat{\beta} = (\hat{v}_1, ..., \hat{v}_r), \tag{IV.16}$$

and the maximized likelihood function is given by,

$$L(\hat{\alpha}, \hat{\beta}, \hat{\Omega}) = (\det(S_{yy})\prod_{i=1}^{r}(1-\hat{\lambda}_i))^{-T/2}. \tag{IV.17}$$

This is an example of so-called reduced rank regression, RRR, which was developed in Anderson (1951), and the ML estimators in Theorem IV.2.2 are said to be found by RRR of ΔX_t on X_{t-1}. Note that the expressions in (IV.13) and (IV.14) show that $\hat{\alpha}$ and $\hat{\Omega}$ are found by OLS regression with β known, and the nonlinear restriction implies that in order to find $\hat{\beta}$ an eigenvalue problem has to be solved.

Proof: Under H_r the log-likelihood function is, apart from a constant, given by

$$\log L(\alpha, \beta, \Omega) = -\frac{T}{2}\log\det(\Omega) - \frac{1}{2}\sum_{t=1}^{T}(\Delta X_t - \alpha\beta'X_{t-1})'\Omega^{-1}(\Delta X_t - \alpha\beta'X_{t-1}). \tag{IV.18}$$

With $\beta \in \mathbb{R}^{p \times r}$ fixed, this is the likelihood function for the usual linear regression model, and $\hat{\alpha}(\beta)$ and $\hat{\Omega}(\beta)$ are found by OLS regression of $Y_t = \Delta X_t$ on $\beta' Z_t = \beta' X_{t-1}$,

$$\hat{\alpha}\,(\beta) = S_{yz}\beta(\beta' S_{zz}\beta)^{-1}, \tag{IV.19}$$

$$\hat{\Omega}\,(\beta) = S_{yy\cdot\beta} = S_{yy} - S_{yz}\beta(\beta' S_{zz}\beta)^{-1}\beta' S_{zy}. \tag{IV.20}$$

Inserting these in $L(\alpha, \beta, \Omega)$ gives the concentrated likelihood function,

$$\log L(\beta, \hat{\alpha}\,(\beta)\,, \hat{\Omega}(\beta)) = -\frac{T}{2}\det\left(\hat{\Omega}\,(\beta)\right). \tag{IV.21}$$

Next, exploiting the determinant of a block-matrix gives,

$$\det\begin{pmatrix} S_{yy} & S_{yz}\beta \\ \beta' S_{zy} & \beta' S_{zz}\beta \end{pmatrix} = \det\left(S_{yy\cdot\beta}\right)\det\left(S_{\beta\beta}\right) = \det\left(S_{yy}\right)\det\left(S_{\beta\beta\cdot y}\right), \tag{IV.22}$$

where the index β refers to $\beta' Z_t$, such that e.g. $S_{\beta\beta\cdot y} = \beta' S_{zz}\beta - \beta' S_{zy} S_{yy}^{-1} S_{yz}\beta$. Use this to see that, as $\hat{\Omega}\,(\beta) = S_{yy\cdot\beta}$, $\hat{\beta}$ is found by minimization of,

$$\det\left(\hat{\Omega}\,(\beta)\right) = \det\left(S_{yy}\right)\frac{\det\left(S_{\beta\beta\cdot y}\right)}{\det\left(S_{\beta\beta}\right)} = \det\left(S_{yy}\right)\frac{\det\left(\beta' S_{zz\cdot y}\beta\right)}{\det\left(\beta' S_{zz}\beta\right)}. \tag{IV.23}$$

That this ratio of quadratic forms is minimized by solving the eigenvalue problem in (IV.15), and setting $\hat{\beta} = (\hat{v}_1, ..., \hat{v}_r)$, that is the eigenvectors corresponding to the r largest eigenvalues, follows by Johansen (1996, Theorem 6.1). □

IV.2.3 Hypothesis testing

Hypothesis testing for two hypotheses will be discussed here: First the hypothesis of Π having reduced rank r, and next, given that the rank r is known, linear restrictions on the cointegration vectors β. For other kinds of hypotheses, see Section IV.3 for a discussion of the asymptotic properties of $\hat{\alpha}$ and $\hat{\beta}$.

Turn first to the hypothesis of Π having reduced rank r.

IV.2.3.1 Rank test

By definition, the LR test statistic for the hypothesis H_r against the unrestricted model, H_p, is given by $\text{LR}_r = -2\log Q$, where Q is the ratio of the maximized likelihood functions. This equals, by Theorem IV.2.2,

$$\text{LR}_r \equiv \text{LR}(H_r|H_p) = -T\sum_{i=r+1}^{p}\log(1-\hat{\lambda}_i). \tag{IV.24}$$

That is, the computation of the test statistics LR_i for $i = 0, 1, ..., p-1$ is based on solving one, and only one, eigenvalue problem. Now to address the limiting distribution of the test statistic LR_r, the properties of the process X_t are needed. By Theorem IV.2.1, X_t has a representation as a sum of a random walk $\sum_{i=1}^{t} \varepsilon_i$, and a geometrically ergodic process. Under the assumptions in Theorem IV.2.1,

$$\frac{1}{\sqrt{T}} X_{[Tu]} \overset{D}{\rightarrow} C\mathcal{B}_u, \tag{IV.25}$$

as $T \rightarrow \infty$, by application of the invariance principle as in the univariate case. Here $\mathcal{B}$ is a $(p-r)$ dimensional Brownian motion with variance Ω, while $C\mathcal{B}$ has variance,

$$\Omega_\infty = C\Omega C'. \tag{IV.26}$$

The $(p \times p)$ variance Ω_∞ is referred to as the long-run variance of X_t. By Theorem IV.2.1 $C = \beta_\perp(\alpha'_\perp\beta_\perp)^{-1}\alpha'_\perp$, and hence Ω_∞ has rank $p - r < p$ and it is therefore *singular.* This corresponds to the assumption of exactly $p - r$ unit roots, and reflects that the $(p - r)$ dimensional random walk, or the $p - r$ common trends, $\alpha'_\perp \sum_{i=1}^{t} \varepsilon_i$ enter X_t by the $(p \times (p-r))$ dimensional coefficient matrix, $\beta_\perp(\alpha'_\perp\beta_\perp)^{-1}$.

Thus, in addition to the assumption H_r of rank Π less than or equal to r, the additional assumptions in Theorem IV.2.1 are vital for interpreting the (limiting) behavior of X_t. Denote therefore by H_r^0 the hypothesis H_r where the additional assumptions hold, i.e. that $A(z)$ has exactly $p - r$ roots at $z = 1$, while the remaining roots are larger than one in absolute value.

The following result holds:

Theorem IV.2.3 *Under H_r^0, it holds that*

$$LR_r \overset{D}{\rightarrow} DF_{p-r}(\mathcal{W}), \tag{IV.27}$$

where $\mathcal{W}$ is a $p - r$ dimensional standard Brownian motion, and

$$DF_{p-r}(\mathcal{W}) = \mathrm{tr}\{\int_0^1 d\mathcal{W}\mathcal{W}'(\int_0^1 \mathcal{W}_u\mathcal{W}'_u du)^{-1}\int_0^1 \mathcal{W}d\mathcal{W}'\}. \tag{IV.28}$$

The limiting distribution $\mathrm{DF}_1(\mathcal{W})$ is identical to the limiting distribution of the LR test for a unit root in the univariate AR model, see Theorem **??** and $\mathrm{DF}_{p-r}(\mathcal{W})$ is the multivariate generalization.

Some quantiles of $\mathrm{DF}_{p-r}(\mathcal{W})$ for different values of $p - r$ are reported in Section IV.5.4.

Proof: Consider here the case of $r = 0$ only, see Johansen (1996) for the general case. That $r = 0$ is equivalent to $\Pi = 0$ which is a linear restriction.

That is, as for the linear regression model with $Y_t = \Delta X_t$ and $Z_t = X_{t-1}$, the LR_0 test statistic can be written as,

$$\text{LR}_0 = -T \log \det \left(I_p - \tilde{W}_T \right), \quad \tilde{W}_T = S_{yy}^{-1} \hat{\Pi} S_{zz} \hat{\Pi}' = S_{yy}^{-1} S_{yz} S_{zz}^{-1} S_{zy}. \tag{IV.29}$$

Equivalently, this also follows by noting that by definition $\text{LR}_0 = -T \sum_{i=1}^{p} \log(1-\hat{\lambda}_i) = -T \log \det \left(I - \hat{\Lambda} \right)$, where $\hat{\Lambda}$ is given in Theorem IV.2.2. Next, under $H_0^0 = H_0$, $\Delta X_t = \varepsilon_t$ and $X_t = \sum_{i=1}^{t} \varepsilon_i + X_0$. Hence,

$$S_{yy} = \frac{1}{T} \sum_{t=1}^{T} \Delta X_t \Delta X_t' \xrightarrow{P} \Omega, \tag{IV.30}$$

by the LLN for i.i.d. variables. By the FCLT for i.i.d. variables, and as in Theorem **??**,

$$\frac{1}{\sqrt{T}} X_{[Tu]} \xrightarrow{D} \mathcal{B}_u, \tag{IV.31}$$

$$T^{-1} S_{zz} = T^{-2} \sum_{t=1}^{T} X_{t-1} X_{t-1}' \xrightarrow{D} \int_0^1 \mathcal{B}_u \mathcal{B}_u' du \quad \text{and} \tag{IV.32}$$

$$S_{zy} = T^{-2} \sum_{t=1}^{T} X_{t-1} \Delta X_t' \xrightarrow{D} \int_0^1 \mathcal{B} d\mathcal{B}', \tag{IV.33}$$

where $\mathcal{B}$ is a Brownian motion on $[0,1]$ with covariance Ω. Thus $\text{tr}\{\tilde{W}_T\} = O_P(T^{-1})$, and a stochastic Taylor expansion of $\log \det \left(I - \tilde{W}_T \right)$ gives,

$$\text{LR}_0 = -T \log \det \left(I_p - \tilde{W}_T \right) = \text{tr}\{T \tilde{W}_T\} + o_P(1) \xrightarrow{D} \tag{IV.34}$$

$$\text{tr}\{\Omega^{-1} \int_0^1 d\mathcal{B}\mathcal{B}' (\int_0^1 \mathcal{B}\mathcal{B}' du)^{-1} \int_0^1 \mathcal{B} d\mathcal{B}'\} \overset{D}{=} \text{DF}_p \left(\mathcal{W} \right), \tag{IV.35}$$

where $\mathcal{W} = \Omega^{-1/2} \mathcal{B}$, is a p dimensional standard Brownian motion. □

When testing for the rank r we proceed as follows:

Considering the test statistic LR_r, the limiting distribution will be different from the one in Theorem IV.2.3 if H_r^0 does not hold. In particular, H_r^0 states that the rank of Π equals r, while H_r states that the rank is less than or equal to r. Thus the limiting distribution of LR_r will be different for each H_i^0 for $i = 0, 1, .., r$ which are all nested in the H_r hypothesis. This problem is avoided by using a sequential testing procedure to find the rank $\hat{r}$.

The idea is to compute LR_i for all $i = 0, 1, ..., p-1$. Then the rank r is accepted, $\hat{r} = r$, provided that for the test statistics concerning lower ranks it holds that, $\mathrm{LR}_0 > c_0$, $\mathrm{LR}_1 > c_1$, ...,$\mathrm{LR}_{r-1} > c_{r-1}$, while $\mathrm{LR}_r < c_r$, where c_j are the critical values corresponding to, say, the 95% quantile of $\mathrm{DF}_{p-j}(\mathcal{W})$ in Theorem IV.2.3. This way rank r is accepted, if previous smaller ranks have all been rejected. The idea is that, when evaluating LR_r, all previous cases have been rejected, and hence H_{r-1}, or rank Π less than or equal to $r-1$, does not hold. It is therefore sufficient to consider LR_r under H_r^0, where the limiting distribution is given in Theorem IV.2.3.

Example IV.2.4 *For the spot and futures data, analysis of a bivariate VAR model with $X_t = (s_t, f_t)'$ gives $LR_0 > c_0$, and hence rank $r = 0$ is rejected. Next, $LR_1 < c_1$ and H_1 is accepted. As H_0, or rank equal to zero, was rejected, accepting H_1, or rank less than or equal to 1, implies $\hat{r} = 1$.*

IV.2.3.2 Linear hypotheses on β

Next, with the rank r given, consider here linear hypotheses on β of the form,

$$H_{\mathrm{lin}} : \beta = H\varphi, \tag{IV.36}$$

where H is a $p \times s$ dimensional known matrix, while φ is a $s \times r$ dimensional matrix with freely varying parameters, with $r \leq s \leq p$. The hypothesis is equivalent to

$$H_{\mathrm{lin}} : R'\beta = 0, \tag{IV.37}$$

where $R = H_\perp$ and is a $p \times (p-s)$ dimensional matrix. For example, this kind of hypothesis allows for testing if a variable X_{it} can be excluded in all cointegrating relations, by setting $R' = (0, ..0, 1, 0, ..., 0)$ with the '1' in place i.

Example IV.2.5 *Continuing with the $X_t = (s_t, f_t)'$ example, with $r = 1$, the hypothesis that, say, f_t does not enter in the cointegrating relation can be formulated as*

$$\beta = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \varphi = H\varphi. \tag{IV.38}$$

Rewriting the VAR(1) under H_{lin},

$$\Delta X_t = \alpha\beta' X_{t-1} + \varepsilon_t = \alpha\varphi' H' X_{t-1} + \varepsilon_t, \tag{IV.39}$$

shows that α, φ and Ω can be found as in Theorem IV.2.2 by setting $Y_t = \Delta X_t$ as there, while $Z_t = H'X_{t-1}$ and $\beta = H\varphi$. That is, RRR of ΔX_t on $H'X_{t-1}$ gives the ML estimators $\hat{\alpha}$, $\hat{\varphi}$ and $\hat{\Omega}$ in this case.

Using this, it follows directly that the likelihood ratio test statistic of H_{lin} against H_r is given by,

$$\text{LR}_{\text{lin}} = \text{LR}(H_{\text{lin}}|H_r) = T\sum_{i=1}^{r}\log(\frac{1-\hat{\lambda}_i}{1-\tilde{\lambda}_i}), \tag{IV.40}$$

which is asymptotically χ^2 distributed with $r\,(p-s)$ degrees of freedom. Here the eigenvalues $\hat{\lambda}_i$ solve the eigenvalue problem in (IV.15) with $Z_t = X_{t-1}$, while $\tilde{\lambda}_i$ solve the eigenvalue problem with $Z_t = H'X_{t-1}$, corresponding to ML estimation under H_{lin}.

That LR_{lin} is asymptotically χ^2 distributed may seem somewhat surprising. It is an implication of the result that under H_r^0, $\hat{\beta}$, appropriately normalized, is super-consistent and asymptotically *mixed* Gaussian distributed, while $\hat{\alpha}$ is consistent at the standard rate and asymptotically Gaussian, see Section IV.3.

In general, hypotheses of the form,

$$\beta = (\beta_1, ..., \beta_r) = (H_1\varphi_1, ..., H_r\varphi_r), \tag{IV.41}$$

that is where each cointegrating vector β_i is restricted by a linear restriction given by H_i, are of interest. Estimation, and asymptotic distributions of the LR tests, are not discussed here. However, observe that the hypothesis of a single variable X_{it} being stationary, or asymptotically stable, can be written in exactly this form by setting $H_i = (0, ..0, 1, 0, ..., 0)'$ with the '1' in place i, and leaving the other cointegrating vectors unrestricted. Moreover, in the simple case of $r = 1$, this collapses to a restriction of the form H_{lin} discussed above.

Example IV.2.6 *Continuing with the $X_t = (s_t, f_t)'$ example, with $r = 1$, the hypothesis that, $s_t - f_t$ is a cointegrating relation can be formulated as*

$$\beta = \begin{pmatrix} 1 \\ -1 \end{pmatrix}\varphi = H\varphi. \tag{IV.42}$$

IV.3 Asymptotics for the MLEs of β and α

Having derived the MLEs for α and β given in Theorem IV.2.2, and secondly being able to test for cointegration (and the order r) from the result in Theorem IV.2.3, we now discuss the asymptotic properties of $\hat{\alpha}$ and $\hat{\beta}$ under H_0^r.

Before discussing distributional theory for $\hat{\alpha}$ and $\hat{\beta}$, consider first the issue of identification. As mentioned in the previous section the sub-spaces sp (β)

and $\mathrm{sp}(\alpha)$ are identified, which follows by noting that with m any $(r \times r)$ matrix of full rank r, we have

$$\Pi = \alpha\beta' = [\alpha m'] \left[\left(\beta m^{-1} \right)' \right] = \alpha_m \beta_m',$$

with $\beta_m = \beta m^{-1}$ and $\alpha_m = \alpha m'$. Thus, while the spaces spanned by (the columns of) α and β are identified, the individual parameters in α and β are not identified. To identify these, some normalization as e.g. given by a specific known choice of m can be imposed.

For example, with $p = 2$ and $r = 1$, we have for $\beta = (\beta_1, \beta_2)'$. With $m = \beta_1$, we get

$$\beta_m = \begin{pmatrix} 1 \\ b \end{pmatrix}, \quad \alpha_m = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix},$$

such that three parameters b, a_1 and a_2 are identified. That is, while the 4 parameters in β and α are not identified, the 3 parameters in the new parametrization in terms of b, a_1 and a_s identified, where $b = \beta_2/\beta_1$ and $a_1 = \alpha_1\beta_1, a_2 = \alpha_2\beta_1$. We say that β (and hence α) are identified by the normalization with $m = \beta_1$, or simply, by $\beta_1 = 1$.

For general p and r, often m is chosen as $m = c'\beta$ with c some known $(p \times r)$ matrix. Corresponding to the simple case above, consider $c = (I_r, 0)'$, such that

$$\beta_m = \beta \left(c'\beta \right)^{-1} = \begin{pmatrix} I_r \\ b \end{pmatrix},$$

where b $((p-r) \times r)$ and α_m $(p \times r)$ are identified. Note that, as $m = c'\beta$, this requires in particular $c'\beta$ to have full rank.

IV.3.1 Preliminary considerations

Consider again the simple case of $p = 2$ and $r = 1$, with normalisation $m = c'\beta$, $c = (1, 0)'$, and hence

$$\beta_m = (1, b)' \text{ and } \alpha_m = (a_1, a_2)'.$$

The VAR model can then be stated in terms of the new identified parameters α_m and β_m (that is, b) as,

$$\Delta X_t = \alpha_m \beta_m' X_{t-1} + \varepsilon_t$$

Consider the MLE $\hat{b}$, or equivalently, $\hat{\beta}_m = (1, \hat{b})'$ with Ω and α_m fixed for simplicity. By definition, $\hat{b}$ satisfies

$$\partial \log L \left(\alpha_m, \beta_m, \Omega \right) / \partial b|_{\hat{\beta}_m} = 0,$$

where

$$\log L\left(\alpha_m,\beta_m,\Omega\right)=-\frac{1}{2}\left[T\log\det\left(\Omega\right)+\sum_{t=1}^{T}(\Delta X_t-\alpha_m\beta_m'X_{t-1})'\Omega^{-1}(\Delta X_t-\alpha_m\beta_m'X_{t-1})\right].$$

Using $\partial\left(\beta_m'X_{t-1}\right)/\partial b=\partial\left(1+bX_{2t-1}\right)/\partial b=X_{2t-1}$, simple differentiation gives

$$\partial\log L\left(\alpha_m,\beta_m,\Omega\right)/\partial b|_{\hat{\beta}_m}=\sum_{t=1}^{T}X_{2t-1}\alpha_m'\Omega^{-1}(\Delta X_t-\alpha_m\hat{\beta}_m'X_{t-1})$$

Next, insert $\Delta X_t=\alpha_m\beta_m'X_{t-1}+\varepsilon_t$, such that we get

$$\begin{aligned}\partial\log L\left(\alpha_m,\beta_m,\Omega\right)/\partial b|_{\hat{\beta}_m}&=\sum_{t=1}^{T}X_{2t-1}\alpha_m'\Omega^{-1}(\varepsilon_t-\alpha_m(\hat{\beta}_m-\beta_m)'X_{t-1})\\&=\sum_{t=1}^{T}X_{2t-1}\alpha_m'\Omega^{-1}(\varepsilon_t-\alpha_m(\hat{b}-b)X_{2t-1})\end{aligned}$$

We conclude that $\partial\log L\left(\alpha_m,\beta_m,\Omega\right)/\partial b|_{\hat{\beta}_m}=0$ is equivalent to,

$$\alpha_m'\Omega^{-1}\sum_{t=1}^{T}\varepsilon_tX_{2t-1}=\left(\alpha_m'\Omega^{-1}\alpha_m\right)\left(\hat{b}-b\right)\sum_{t=1}^{T}X_{2t-1}^2$$

such that solving for $\hat{b}-b$ gives,

$$\begin{aligned}\hat{b}-b&=\left(\alpha_m'\Omega^{-1}\alpha_m\right)^{-1}\alpha_m'\Omega^{-1}\sum_{t=1}^{T}\varepsilon_tX_{2t-1}\left[\sum_{t=1}^{T}X_{2t-1}^2\right]^{-1}\\&=\left[\sum_{t=1}^{T}X_{2t-1}^2\right]^{-1}\sum_{t=1}^{T}X_{2t-1}\varepsilon_t'\Omega^{-1}\alpha_m\left(\alpha_m'\Omega^{-1}\alpha_m\right)^{-1}\end{aligned}\qquad\text{(IV.43)}$$

Observe that the stochastic behaviour of $\hat{b}$ is given by the limiting behaviour of the two key quantities,

$$\sum_{t=1}^{T}\varepsilon_tX_{2t-1},\quad\sum_{t=1}^{T}X_{2t-1}^2.$$

For the properties of X_{2t}, note initially that by (IV.8) in Theorem IV.2.1, it follows directly that X_{2t} has the representation,

$$X_{2t}=\left(0,1\right)X_t=\left(0,1\right)\left[C\sum_{i=1}^{t}\varepsilon_i+C_SS_t+C_0\right],$$

where $C = \beta_{m\perp}\left(\alpha'_{m\perp}\beta_{m\perp}\right)^{-1}\alpha'_{m\perp}$. By defintion of β_m, we may choose

$$\beta_{m\perp} = \begin{pmatrix} -b \\ 1 \end{pmatrix},$$

such that $(0,1)\,\beta_{m\perp} = 1 \neq 0$ and hence $(0,1)\,C \neq 0$, implying that $X_{2t} = (0,1)\,X_t$ is I(1). Using the previous results for convergence to the Brownian motion in (IV.25), we find with $u \in (0,1)$,

$$\begin{aligned}\frac{1}{\sqrt{T}}X_{2[Tu]} &= \frac{1}{\sqrt{T}}(0,1)\,X_{[Tu]} \\ &= \left(\alpha'_{m\perp}\beta_{m\perp}\right)^{-1}\frac{1}{\sqrt{T}}\sum_{i=1}^{[Tu]}\alpha'_{m\perp}\varepsilon_i + o_p(1) \\ &\xrightarrow{D} \gamma'\mathcal{B}_u, \qquad \gamma = \alpha_{m\perp}\left(\beta'_{m\perp}\alpha_{m\perp}\right)^{-1}.\end{aligned}$$

Multiplying $(\hat{b}-b)$ by T, we get

$$T\left(\hat{b}-b\right) = \left[T^{-2}\sum_{t=1}^{T}X_{2t-1}^2\right]^{-1}T^{-1}\sum_{t=1}^{T}X_{2t-1}\varepsilon'_t\Omega^{-1}\alpha_m\left(\alpha'_m\Omega^{-1}\alpha_m\right)^{-1}$$

and hence, collecting terms,

$$T\left(\hat{b}-b\right) \xrightarrow{D} \left[\int_0^1\gamma'\mathcal{B}_u\mathcal{B}'_u\gamma du\right]^{-1}\int_0^1\gamma'\mathcal{B}_u d\mathcal{B}'_u\Omega^{-1}\alpha_m\left(\alpha'_m\Omega^{-1}\alpha_m\right)^{-1},$$

with $\gamma = \alpha_{m\perp}\left(\beta'_{m\perp}\alpha_{m\perp}\right)^{-1}$. We note that as $\mathrm{Cov}\left(\alpha'_m\Omega^{-1}\mathcal{B}_u, \gamma'\mathcal{B}_u\right) = \alpha'_m\Omega^{-1}\Omega\gamma = \alpha'_m\gamma = 0$, then $\gamma'\mathcal{B}_u$ and $\alpha'_m\Omega^{-1}\mathcal{B}_u$ are independent, which means $\hat{b}$ is asymptotically mixed Gaussian (MG) distributed. Moreover, note that $\hat{b}$ is super consistent due to the rate T of convergence.

A different way of stating this is, using the properties of stochastic integrals and MG,

$$T\left(\hat{b}-b\right) \xrightarrow{D} \mathrm{MG}\left(0, 1/\sigma_b^2\right), \quad \text{with } \sigma_b^2 = \left[\int_0^1\gamma'\mathcal{B}_u\mathcal{B}'_u\gamma du\right]\left[\alpha'_m\Omega^{-1}\alpha_m\right]$$

Likewise, in terms of the t-ratio τ,

$$\tau = T\left(\hat{b}-b\right)\sqrt{\sigma_b^2} \xrightarrow{D} \mathrm{N}(0,1).$$

Finally, in terms of the vector $\hat{\beta}_m = (1,\hat{b})'$, we get, using that $m = c'\beta$ with $c = (1,0)'$, we can set $c_\perp = (0,1)'$ and hence,

$$T\left(\hat{\beta}_m - \beta_m\right) = \begin{pmatrix} 0 \\ T\left(\hat{b}-b\right) \end{pmatrix} = c_\perp T\left(\hat{b}-b\right) \xrightarrow{D} c_\perp\,\mathrm{MG}\left(0, 1/\sigma_b^2\right).$$

In other words, $c_\perp' \hat{\beta}_m = \hat{b}_m$ is asymptotically non-standard distributed with a limiting mixed Gaussian distribution, such that the t-ratio τ is standard Gaussian distributed, and, as can be shown, in general LR statistics for hypothesis testing on β are asymptotically χ^2 distributed.

IV.3.2 Asymptotics for MLE of β

The previous considerations can be extended to the general case of dimension p and rank r, with $\beta_m = \beta m^{-1}$:

Theorem IV.3.1 *Under the I(1) conditions H_0^r, then with $m = c'\beta$ of full rank r,*

$$T\left(\hat{\beta}_m - \beta_m\right) \overset{D}{\to} c_\perp \left[\int_0^1 \gamma' \mathcal{B}_u \mathcal{B}_u' \gamma du\right]^{-1} \int_0^1 \gamma' \mathcal{B}_u d\mathcal{B}_u' \Omega^{-1} \alpha_m \left(\alpha_m' \Omega^{-1} \alpha_m\right)^{-1},$$

where $\gamma = \alpha_{m\perp} \left(\beta_{m\perp}' \alpha_{m\perp}\right)^{-1}$.

Note that while the considerations in the previous section were made under the assumption of fixed α_m and Ω, Theorem IV.3.1 holds for α_m and Ω estimated. A proof, similar to the proof for the case of $p = 2$ and $r = 1$ is given in Section IV.3.4 below (see also Johansen, 1996, proof of Theorem 13.3).

Note also that Theorem IV.3.1 states that $\hat{\beta}_m$ is asymptotically mixed Gaussian,

$$T\left(\hat{\beta}_m - \beta_m\right) \overset{D}{\to} c_\perp \,\mathrm{MG}\left(0, \Sigma^{-1}\right)$$

where $\Sigma = \left[\int_0^1 \gamma' \mathcal{B}_u \mathcal{B}_u' \gamma du\right]^{-1} \otimes [\alpha_m' \Omega^{-1} \alpha_m]$. Also note that as $c'c_\perp \,\mathrm{MG}\left(0, \Sigma^{-1}\right) = 0$, the limiting distribution of $\hat{\beta}_m$ is singular, reflecting that $\beta_m = (1, b)' = c_\perp b$, and hence that it is the distribution of $\hat{b}$ in $\hat{\beta}_m$ which is non-singular.

Note Recall that when estimating β by solving the eigenvalue problem in (IV.15), by definition we have

$$\hat{\beta}' S_{zz} \hat{\beta} = I_r \quad \text{and} \quad \hat{\beta}' S_{zy} S_{yy}^{-1} S_{yz} \hat{\beta} = \hat{\Lambda}_r,$$

where $\hat{\Lambda}_r = \mathrm{diag}(\hat{\lambda}_1, ..., \hat{\lambda}_r)$. These normalizations – which are more complicated than using m^{-1} for β_m – ensure identification as well. However, the limiting distribution of $\hat{\beta}$ normalized this way we do not consider as it is not needed for our considerations.

IV.3.3 Asymptotics for MLE of α

With β normalized by m, $\beta_m = \beta m^{-1}$, recall that $\alpha_m = \alpha m'$, and hence by (IV.13),

$$\hat{\alpha}_m - \alpha_m = S_{\varepsilon z}\hat{\beta}_m(\hat{\beta}_m' S_{zz}\hat{\beta}_m)^{-1}$$

Thus with β_m fixed, it follows that $\hat{\alpha}_m(\beta_m)$ satisfies,

$$\begin{aligned}\sqrt{T}(\hat{\alpha}_m(\beta_m) - \alpha_m) &= \sqrt{T}S_{\varepsilon z}\beta_m(\beta_m' S_{zz}\beta_m)^{-1} \\ &= \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\varepsilon_t X_{t-1}'\beta_m(\frac{1}{T}\sum_{t=1}^{T}\beta_m' X_{t-1}X_{t-1}'\beta_m)^{-1}.\end{aligned}$$

And, using standard arguments for the CLT and LLN,

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\varepsilon_t X_{t-1}'\beta_m \to_d \mathrm{N}\left(0, \Omega \otimes \Sigma_{\beta\beta}\right), \qquad \frac{1}{T}\sum_{t=1}^{T}\beta_m' X_{t-1}X_{t-1}'\beta_m \to_p \Sigma_{\beta\beta},$$

where $\Sigma_{\beta\beta} = \mathbb{V}(\beta_m' X_t^*)$, we find

$$\sqrt{T}(\hat{\alpha}_m(\beta_m) - \alpha_m) \to_d \mathrm{N}\left(0, \Omega \otimes \Sigma_{\beta\beta}^{-1}\right).$$

For $\hat{\alpha}_m = \hat{\alpha}_m(\hat{\beta}_m)$, that is with $\hat{\beta}_m$ inserted, we find the equivalent result:

Theorem IV.3.2 *Under the I(1) conditions, then with $m = c'\beta$ of full rank r,*

$$\sqrt{T}(\hat{\alpha}_m - \alpha_m) \to_d \mathrm{N}\left(0, \Omega \otimes \Sigma_{\beta\beta}^{-1}\right),$$

with $\Sigma_{\beta\beta} = \mathbb{V}(\beta_m' X_t^*)$.

A proof can be found in Johansen (1996, proof of Theorem 13.3) which is based on arguments as for the case with β_m known ($\hat{\alpha}_m(\beta_m)$) treated above, using that $\hat{\beta}_m$ is super consistent, and hence, $\hat{\alpha}_m - \alpha_m = \hat{\alpha}_m(\hat{\beta}_m) - \alpha_m = \hat{\alpha}_m(\beta_m) - \alpha_m + o_p(1)$.

In the next section the distribution is derived for $\hat{\beta}_m$ for general p and r, with α_m and Ω fixed.

IV.3.4 General proof for $\hat{\beta}_m$

Consider the VAR model is given by

$$\Delta X_t = \alpha_m \beta_m' X_{t-1} + \varepsilon_t,$$

where $\beta_m = \beta m^{-1} = (I_r, b')$, such that $m = c'\beta$ and $c = (I_r, 0)'$.

Consider the MLE $\hat{b}$, or equivalently, $\hat{\beta}_m = (I_r, \hat{b}')'$ with Ω and α_m fixed. By definition, $\hat{b}$ satisfies (with d denoting differential),

$$d \log L\left(\alpha_m, \beta_m, \Omega; db\right)|_{\hat{\beta}_m} = 0, \tag{IV.44}$$

where db is $(p-r) \times r$ dimensional, and

$$\log L\left(\alpha_m, \beta_m, \Omega\right) = -\frac{T}{2} \log \det\left(\Omega\right) - \frac{1}{2} \sum_{t=1}^{T} (\Delta X_t - \alpha_m \beta_m' X_{t-1})' \Omega^{-1} (\Delta X_t - \alpha_m \beta_m' X_{t-1}).$$

Set, $c_\perp = (0, I_{p-r})'$, and note that as $\beta_m = c_\perp b$, then $d\beta_m = c_\perp db$, and therefore $d\left(\beta' X_{t-1}\right) = db' c_\perp' X_t$. We find that (IV.44) is given by

$$\begin{aligned} d \log L\left(\alpha_m, \beta_m, \Omega : db\right)|_{\hat{\beta}_m} &= \sum_{t=1}^{T} (\Delta X_t - \alpha_m \hat{\beta}_m' X_{t-1})' \Omega^{-1} \alpha_m db' c_\perp' X_{t-1} \\ &= \sum_{t=1}^{T} (\varepsilon_t - \alpha_m (\hat{\beta}_m - \beta_m)' X_{t-1})' \Omega^{-1} \alpha_m db' c_\perp' X_{t-1} \\ &= \sum_{t=1}^{T} (\varepsilon_t - \alpha_m (\hat{b} - b)' c_\perp' X_{t-1})' \Omega^{-1} \alpha_m db' c_\perp' X_{t-1} = 0 \end{aligned}$$

Hence,

$$\sum_{t=1}^{T} \varepsilon_t' \Omega^{-1} \alpha_m db' c_\perp' X_{t-1} = \sum_{t=1}^{T} X_{t-1}' c_\perp \left(\hat{b} - b\right) \alpha_m' \Omega^{-1} \alpha_m db' c_\perp' X_{t-1} \tag{IV.45}$$

Using that $\operatorname{tr}\left(a\right) = a$ with a scalar, and $\operatorname{tr}\left(AB\right) = \operatorname{tr}\left(BA\right)$,

$$\operatorname{tr}(\sum_{t=1}^{T} \varepsilon_t' \Omega^{-1} \alpha_m db' c_\perp' X_{t-1}) = \operatorname{tr}(\sum_{t=1}^{T} c_\perp' X_{t-1} \varepsilon_t' \Omega^{-1} \alpha_m db')$$

$$\operatorname{tr}(\sum_{t=1}^{T} X_{t-1}' c_\perp \left(\hat{b} - b\right) \alpha_m' \Omega^{-1} \alpha_m db' c_\perp' X_{t-1}) = \operatorname{tr}(\sum_{t=1}^{T} c_\perp' X_{t-1} X_{t-1}' c_\perp \left(\hat{b} - b\right) \alpha_m' \Omega^{-1} \alpha_m db')$$

and as (IV.45) holds for all db,

$$T^{-1} \sum_{t=1}^{T} c_\perp' X_{t-1} \varepsilon_t' \Omega^{-1} \alpha_m = T^{-1} \sum_{t=1}^{T} c_\perp' X_{t-1} X_{t-1}' c_\perp \left(\hat{b} - b\right) \alpha_m' \Omega^{-1} \alpha_m,$$

or

$$c_\perp' S_{z\varepsilon} \Omega^{-1} \alpha_m = c_\perp' S_{zz} c_\perp \left(\hat{b} - b\right) \alpha_m' \Omega^{-1} \alpha_m,$$

That is, analogous to the univariate case,

$$\hat{b}-b=(c_{\perp}'S_{zz}c_{\perp})^{-1}c_{\perp}'S_{z\varepsilon}\Omega^{-1}\alpha_m\left(\alpha_m'\Omega^{-1}\alpha_m\right)^{-1}$$

Again by (IV.8),

$$c_{\perp}'X_t=c_{\perp}'\left[C\sum_{i=1}^{t}\varepsilon_i+C_SS_t+C_0\right],$$

where $C_{\Sigma}=\beta_{m\perp}\left(\alpha_{m\perp}'\beta_{m\perp}\right)^{-1}\alpha_{m\perp}'$, with

$$\beta_{m\perp}=\begin{pmatrix}-b'\\ I_{p-r}\end{pmatrix},\quad\text{where }\beta_m'\beta_{m\perp}=-b'+b'=0.$$

In particular, $c_{\perp}'\beta_{m\perp}=I_{p-r}\neq 0$, and hence, as above,

$$\begin{aligned}\frac{1}{\sqrt{T}}c_{\perp}'X_{[Tu]}&=\left(\alpha_{m\perp}'\beta_{m\perp}\right)^{-1}\frac{1}{\sqrt{T}}\sum_{i=1}^{[Tu]}\alpha_{m\perp}'\varepsilon_i+o_p(1)\\&\xrightarrow{D}\left(\alpha_{m\perp}'\beta_{m\perp}\right)^{-1}\alpha_{m\perp}'\mathcal{B}_u=\gamma'\mathcal{B}_u.\end{aligned}$$

Collecting terms,

$$T\left(\hat{b}-b\right)\xrightarrow{D}\mathrm{MG}:=\left[\int_0^1\gamma'\mathcal{B}_u\mathcal{B}_u'\gamma du\right]^{-1}\int_0^1\gamma'\mathcal{B}_ud\mathcal{B}_u'\Omega^{-1}\alpha_m\left(\alpha_m'\Omega^{-1}\alpha_m\right)^{-1}$$

That is, as $\mathrm{Cov}\left(\alpha_m'\Omega^{-1}\mathcal{B}_u,\gamma'\mathcal{B}_u\right)=\alpha_m'\Omega^{-1}\Omega\gamma=\alpha_m'\gamma=0$, $\hat{b}$ is superconsistent and asymptotically mixed Gaussian distributed. This can also be stated as in Theorem IV.3.1, using $\beta_m=c_{\perp}b$,

$$T\left(\hat{\beta}_m-\beta_m\right)\xrightarrow{D}c_{\perp}\,\mathrm{MG}=\left[\int_0^1\gamma'\mathcal{B}_u\mathcal{B}_u'\gamma du\right]^{-1}\int_0^1\gamma'\mathcal{B}_ud\mathcal{B}_u'\Omega^{-1}\alpha_m\left(\alpha_m'\Omega^{-1}\alpha_m\right)^{-1}$$

IV.4 Orthogonal complements

Having obtained the asymptotic distribution for the normalized $\hat{\alpha}_m$ and $\hat{\beta}_m$, we now consider their orthogonal complements.

Recall that $\alpha_{\perp}$ and $\beta_{\perp}$ are $(p\times(p-r))$-dimensional matrices of full rank $p-r$ and such that

$$\beta_{\perp}'\beta=0\text{ and }\alpha_{\perp}'\alpha=0,$$

with $\det(\alpha,\alpha_{\perp})\neq 0$ and $\det(\beta,\beta_{\perp})\neq 0$.

A main challenge is that, as for α and β, the orthogonal complements $\alpha_{\perp}$ and $\beta_{\perp}$ are not identified in the sense that the parameters are not unique, and a normalization is needed; importantly, this holds even if α and β are identified.

IV.4.1 Considering $\beta_\perp$

To overcome the identification issue observe that with $\beta_m = \beta m^{-1}$ where $m = c'\beta$, and $c = (I_r, 0)'$ then the analog

$$\beta_{m\perp} = \beta_\perp \left(c_\perp' \beta_\perp\right)^{-1}, \quad c_\perp = \left(0, I_{p-r}\right)',$$

is identified. Next, use the simple identity, or skew-projection,

$$\begin{aligned} I_p &= \beta_\perp \left(c_\perp' \beta_\perp\right)^{-1} c_\perp' + c\left(\beta' c\right)^{-1} \beta' \\ &= \beta_{m\perp} c_\perp' + c\beta_m' \end{aligned}$$

to solve for $\beta_{m\perp}$,

$$\beta_{m\perp} = (I - c\beta_m^{'})c_\perp.$$

In other words, we can find a unique $\beta_{m\perp}$ by using the identified β_m, and by this specific choice

$$\begin{aligned} T(\hat{\beta}_{m\perp} - \beta_{m\perp}) &= T((I - c\hat{\beta}_m^{'})\bar{c}_\perp - (I - c\beta_m^{'})\bar{c}_\perp) \\ &= -Tc(\hat{\beta}_m - \beta_m)'\bar{c}_\perp \end{aligned}$$

From above, we have $T\left(\hat{\beta}_m - \beta_m\right) \xrightarrow{D} c_\perp \, \mathrm{MG}$, and hence, as for $\hat{\beta}_m$, we find

$$T(\hat{\beta}_{m\perp} - \beta_{m\perp}) \xrightarrow{D} -c\,\mathrm{MG}' = -c\left(\alpha_m' \Omega^{-1} \alpha_m\right)^{-1} \alpha_m' \Omega^{-1} \int_0^1 d\mathcal{B}_u \mathcal{B}_u' \gamma \left[\int_0^1 \gamma' \mathcal{B}_u \mathcal{B}_u' \gamma du\right]^{-1}$$

Recall that $c = (I_r, 0)'$, and hence with $c_\perp = (0, I_{p-r})'$ we can state the above as "dual" results:

$$\begin{aligned} T(\hat{\beta}_{m\perp} - \beta_{m\perp}) &= -Tc(\hat{b} - b)' \xrightarrow{D} -c\,\mathrm{MG}' \\ T(\hat{\beta}_m - \beta) &= Tc_\perp(\hat{b} - b) \xrightarrow{D} c_\perp \, \mathrm{MG} \end{aligned}$$

IV.4.2 Asymptotics for $\hat{\alpha}_\perp$

As for $\beta_\perp$, we have $\alpha_{\perp n} = \alpha_\perp n^{-1}$ is identified, where n is a known $(p - r) \times (p - r)$ dimensional matrix of full rank. Which kind of normalization is of interest depends on the application. Thus if for example estimation of the C_Σ in (IV.8) is of interest, we note by definition,

$$C_\Sigma = \beta_\perp(\alpha_\perp' \beta_\perp)^{-1} \alpha_\perp = \beta_{m\perp}(\alpha_{n\perp}' \beta_{m\perp})^{-1} \alpha_{n\perp}.$$

That is, any normalization for $\beta_\perp$ and $\alpha_\perp$ respectively will work; i.e. C_Σ is invariant to the choices. A different situation is if $\alpha_\perp$ itself is of interest, in which case the concrete application implies which normalization may be of interest.

IV.4.2.1 Case of $p = 2, r = 1$

Consider initially a simple choice of $\alpha_\perp$ for the $p = 2, r = 1$ case previously initially considered.

Here $\beta_m = \beta m^{-1} = (1, b)'$, with $m = c'\beta$ and $c = (1, 0)'$. Hence β_m and $\alpha_m = (a_1, a_2)'$ are identified, and we can choose,

$$\alpha_{m\perp} = \begin{pmatrix} \alpha_{m\perp,1} \\ \alpha_{m\perp,2} \end{pmatrix} = \begin{pmatrix} -a_2 \\ a_1 \end{pmatrix} = A\alpha, \quad \text{where } A = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}.$$

For this particular choice it follows immediately that with $\sigma^2_{\beta\beta} = \mathbb{V}(\beta'_m X^*_t)$,

$$\sqrt{T}(\hat{\alpha}_{m\perp} - \alpha_{m\perp}) = A\sqrt{T}(\hat{\alpha}_m - \alpha_m) \overset{D}{\to} A\,\mathrm{N}(0, \Omega/\sigma^2_{\beta\beta}) = \mathrm{N}(0, A\Omega A'/\sigma^2_{\beta\beta}).$$

Alternatively, normalizing as $\alpha_{n\perp} = \alpha_\perp n^{-1}$, with $n = c'_\perp \alpha_\perp = a_1$, assuming $a_1 \neq 0$, we get.

$$\alpha_{\perp n} = \begin{pmatrix} a_\perp \\ 1 \end{pmatrix} = \begin{pmatrix} -a_2/a_1 \\ 1 \end{pmatrix}.$$

Hence,

$$\hat{\alpha}_{n\perp} - \alpha_{n\perp} = \begin{pmatrix} \hat{a}_\perp - a_\perp \\ 0 \end{pmatrix} = c\left(\hat{a}_\perp - a_\perp\right),$$

where, by definition

$$\hat{a}_\perp - a_\perp = -\left(\frac{\hat{a}_2}{\hat{a}_1} - \frac{a_2}{a_1}\right)$$

which does not have a standard limiting Gaussian distribution it seems at a first sight. However, note that as

$$\frac{\hat{a}_2}{\hat{a}_1} - \frac{a_2}{a_1} = \frac{\hat{a}_2 a_1 - \hat{a}_1 a_2}{a_1 \hat{a}_1} = \frac{\left(\hat{a}_2 - a_2\right) a_1 - \left(\hat{a}_1 - a_1\right) a_2}{a_1 \hat{a}_1}$$

we get,

$$-\sqrt{T}\left(\frac{\hat{a}_2}{\hat{a}_1} - \frac{a_2}{a_1}\right) = \frac{\sqrt{T}\left(\hat{a}_1 - a_1\right) a_2 - \sqrt{T}\left(\hat{a}_2 - a_2\right) a_1}{a_1 \hat{a}_1} = \frac{(a_2, -a_1)}{a_1 \hat{a}_1}\sqrt{T}\left(\hat{\alpha}_m - \alpha_m\right)$$
$$\to_D -\frac{(-a_2/a_1, 1)}{a_1}\,\mathrm{N}\left(0, \Omega/\sigma^2_{\beta\beta}\right) = -\frac{1}{a_1}\alpha'_{n\perp}\,\mathrm{N}\left(0, \Omega/\sigma^2_{\beta\beta}\right).$$

That is, a one dimensional Gaussian limiting distribution.

Yet another normalization, as sometimes applied in the so-called price discovery litterature, is given by

$$\alpha_{p\perp} = \begin{pmatrix} \alpha_{m\perp,1} \\ \alpha_{m\perp,2} \end{pmatrix} = \alpha_{m\perp}\left(p'\alpha_{m\perp}\right)^{-1} = \begin{pmatrix} \alpha_{m\perp,1} \\ \alpha_{m\perp,2} \end{pmatrix} / \left(\alpha_{m\perp,1} + \alpha_{m\perp,2}\right), \;\; p = (1, 1)'.$$

This way, $p'\alpha_{p\perp} = \alpha_{p\perp,1} + \alpha_{p\perp,2} = 1$. As above we get,

$$\sqrt{T}(\hat{\alpha}_{p\perp} - \alpha_{p\perp}) = \sqrt{T}\left(\hat{\alpha}_{m\perp}\left(p'\hat{\alpha}_{m\perp}\right)^{-1} - \alpha_{m\perp}\left(p'\alpha_{m\perp}\right)^{-1}\right)$$

to be asymptotically Gaussian as well.

Next we consider different normalizations of $\alpha_\perp$ by some $(p-r)\times(p-r)$ dimensional full rank n to obtain identication, $\alpha_{\perp n} = \alpha_\perp n^{-1}$.

IV.4.3 A normalization of theoretical interest

A choice of $\hat{\alpha}_\perp$, which may be of theoretical interest, would be

$$\hat{\alpha}_{m\perp} = (I_p - \alpha_m(\hat{\alpha}_m'\alpha_m)^{-1}\hat{\alpha}_m')\alpha_{m\perp} = \alpha_\perp - \alpha_m(\hat{\alpha}_m'\alpha_m)^{-1}\hat{\alpha}_m'\alpha_{m\perp},$$

where $\alpha_{m\perp}'\alpha_m = 0$ and $\det(\alpha_m, \alpha_{m\perp}) \neq 0$.

We note that by definition, $\hat{\alpha}_m'\hat{\alpha}_{m\perp} = 0$, and hence

$$\hat{\alpha}_{m\perp} - \alpha_\perp = -\alpha_m(\hat{\alpha}_m'\alpha_m)^{-1}(\hat{\alpha}_m - \alpha_m)'\alpha_{m\perp}$$

A Taylor expansion of $f(a) = (a'\alpha_m)^{-1}$ gives, with $\bar{\alpha}_m = \alpha_m\left(\alpha_m'\alpha_m\right)^{-1}$,

$$\sqrt{T}(\hat{\alpha}_{m\perp} - \alpha_\perp) = -\bar{\alpha}_m\sqrt{T}(\hat{\alpha}_m - \alpha_m)'\alpha_{m\perp} + o_p(1),$$

where the last term (by a Taylor expansion) is $o_p(1)$ as $\bar{\alpha}_m(\hat{\alpha}_m - \alpha_m)'\bar{\alpha}_m(\hat{\alpha}_m - \alpha_m)'\alpha_{m\perp}$ is $O_p(T^{-1})$.

We conclude,

$$\sqrt{T}(\hat{\alpha}_{m\perp} - \alpha_\perp) \to_D -\bar{\alpha}_m \,\mathrm{N}\left(0, \Sigma_{\beta\beta}\otimes\Omega\right)\alpha_{m\perp} = \mathrm{N}\left(0, \bar{\alpha}_m\Sigma_{\beta\beta}\bar{\alpha}_m' \otimes \alpha_{m\perp}'\Omega\alpha_{m\perp}\right).$$

IV.4.4 Normalization with $c_\perp$

With $\alpha_{n\perp} = \alpha_\perp n^{-1}$, consider here $n = c_\perp'\alpha_\perp$, with $c_\perp = (0, I_{p-r})'$ as used for β and $\beta_\perp$, assuming n of full rank. In line with skew-projections repeatedly used, we have

$$\begin{aligned} I_p &= \alpha_\perp(c_\perp'\alpha_\perp)^{-1}c_\perp' + c\left(\alpha'c\right)^{-1}\alpha' \\ &= \alpha_{\perp n}c_\perp' + c\left(\alpha_m'c\right)^{-1}\alpha_m' \end{aligned}$$

using $\alpha_m = \alpha\beta'c$. It thus follows that, a candidate with is

$$\alpha_{\perp n} = (I_p - c\left(\alpha_m'c\right)^{-1}\alpha_m')c_\perp$$

Hence with $\hat{\alpha}_{\perp n} = (I_p - c\left(\hat{\alpha}_m'c\right)^{-1}\hat{\alpha}_m')c_\perp$, we find

$$\sqrt{T}(\hat{\alpha}_{\perp n} - \alpha_{\perp n}) = -c\sqrt{T}[\left(\hat{\alpha}_m'c\right)^{-1}\hat{\alpha}_m' - (\alpha_m'c)^{-1}\alpha_m']c_\perp$$

Rewriting, we get

$$\sqrt{T}(\hat{\alpha}_{\perp n} - \alpha_{\perp n}) = -c(\alpha_m' c)^{-1}\sqrt{T}[\hat{\alpha}_m - \alpha_m]' c_\perp + V_T$$

where

$$\begin{aligned} V_T &= -c\sqrt{T}[(\hat{\alpha}_m' c)^{-1} - (\alpha_m' c)^{-1}]\hat{\alpha}_m' c_\perp \\ &= c(\alpha_m' c)^{-1}\sqrt{T}(\hat{\alpha}_m - \alpha_m)' c(\alpha_m' c)^{-1}\alpha_m' c_\perp + o_p(1) \end{aligned}$$

with the $o_p(1)$ term as above. Collecting terms,

$$\sqrt{T}(\hat{\alpha}_{\perp n} - \alpha_{\perp n}) = -c(\alpha_m' c)^{-1}\sqrt{T}[\hat{\alpha}_m - \alpha_m]'(I - c(\alpha_m' c)^{-1}\alpha_m')c_\perp$$

and, asymptotic normailty holds by that of $\hat{\alpha}_m$,

$$\sqrt{T}(\hat{\alpha}_{\perp n} - \alpha_{\perp n}) \to_D -c(\alpha_m' c)^{-1}\,\mathrm{N}\left(0, \Omega \otimes \Sigma_{\beta\beta}^{-1}\right)(I - c(\alpha_m' c)^{-1}\alpha_m')c_\perp$$

Remark IV.4.1 *Another choice matching the expression for* $C = \beta_\perp (\alpha_\perp' \beta_\perp)^{-1} \alpha_\perp = \beta_{m\perp} (\alpha_\perp' \beta_{m\perp})^{-1} \alpha_\perp'$*, would be to use* $n = \beta_{m\perp}' \alpha_\perp$*, and hence*

$$\alpha_{n\perp} = \alpha_\perp (\beta_{m\perp}' \alpha_\perp)^{-1}.$$

A detailed discussion of this choice, and more general normalizations can be found in Paruolo (1997).

IV.5 Deterministic terms and VAR(k)

IV.5.1 Constant level

Similar to the univariate case, consider initially the VAR(1) model with a constant regressor,

$$\Delta X_t = \Pi X_{t-1} + \mu + \varepsilon_t, \quad t = 1, 2, ..., T \tag{IV.46}$$

$\Pi \in \mathbb{R}^{p \times p}$, $\mu \in \mathbb{R}^p$, X_0 fixed and ε_t i.i.d.$\mathrm{N}_p(0, \Omega)$. Under the hypothesis $H_r : \Pi = \alpha\beta'$, and Assumption IV.2.1, X_t has the representation,

$$X_t = C\sum_{i=1}^{t}(\varepsilon_i + \mu) + C_S S_{t,c} + C_0, \tag{IV.47}$$

where $S_{t,c} = \beta' X_t$ is the asymptotically stable process given by $S_{t,c} = (I_r + \beta'\alpha)S_{t-1,c} + \beta'\mu$. In other words, the reduced rank assumption on Π implies that X_t has a linear trend $C\mu t$.

The linear trend vanishes provided $C\mu = 0$, or $\alpha'_{\perp}\mu = 0$. This leads to the hypothesis of interest in the case of an included $c$onstant regressor to be given by,

$$H_{r,c} : \Pi = \alpha\beta',\ \mu = \alpha\mu_c^{'}, \tag{IV.48}$$

where μ_c' is an r dimensional vector. Denote by $H_{r,c}^0$, $H_{r,c}^0 \subseteq H_{r,c}$, the hypothesis that $H_{r,c}$ and Assumption IV.2.1 holds. Then under $H_{r,c}^0$, by Theorem IV.2.1,

$$H_{r,c}^0 : X_t = C_\Sigma \sum_{i=1}^{t} \varepsilon_i + C_S S_{t,c} + C_0. \tag{IV.49}$$

In particular, the mean of the stationary version of the cointegrating relations $\beta' S_t$ equals,

$$\mathbb{E}\left[S_{t,c}^*\right] = \sum_{i=0}^{\infty} (I + \beta'\alpha)^i \beta'\mu = -(\beta'\alpha)^{-1}\beta'\mu = -\mu_c^{'}, \tag{IV.50}$$

and X_t has the representation as an I(1) process with a constant level given by $C_0 - C_S\mu_c'$.

Under $H_{r,c}$ the VAR(1) model is given by,

$$\begin{aligned} \Delta X_t &= \alpha\beta' X_{t-1} + \alpha\mu_c^{'} + \varepsilon_t & \text{(IV.51)} \\ &= \alpha \begin{pmatrix} \beta \\ \mu_c \end{pmatrix}' \begin{pmatrix} X_{t-1} \\ 1 \end{pmatrix}' + \varepsilon_t & \text{(IV.52)} \\ &= \alpha\beta_c^{'} X_{t-1,c} + \varepsilon_t, & \text{(IV.53)} \end{aligned}$$

and hence the MLE of α, $\beta_c = (\beta', \mu_c')$ and Ω are found by RRR of ΔX_t on $X_{t-1,c}$. The unrestricted model $H_{p,c}$ is given by (IV.46), and limiting distribution of the LR test of $H_{r,c}$ against $H_{p,c}$ converge in distribution under $H_{r,c}^0$,

$$\text{LR}_{r,c} \equiv \text{LR}(H_{r,c}|H_{p,c}) \overset{D}{\to} DF_{p-r}^c(\mathcal{W}), \tag{IV.54}$$

where

$$DF_{p-r}^c(\mathcal{W}) = \text{tr}\{\int_0^1 d\mathcal{W}\mathcal{W}^{c\prime}(\int_0^1 \mathcal{W}_u^c\mathcal{W}_u^{c\prime}du)^{-1}\int_0^1 \mathcal{W}^c d\mathcal{W}'\}, \tag{IV.55}$$

with $\mathcal{W}$ a $p - r$ dimensional standard Brownian motion, and $\mathcal{W}^c = (\mathcal{W}', 1)'$. Some quantiles of DF^c are reported in Section IV.5.4.

IV.5.2 Linear trend

Similar considerations for a model which allows for a linear trend $\tau \in \mathbb{R}^p$,

$$\Delta X_t = \Pi X_{t-1} + \tau t + \mu + \varepsilon_t, \tag{IV.56}$$

leads to consider the model $H_{r,l}$ as given by,

$$\Delta X_t = \alpha\beta' X_{t-1} + \alpha\tau_l' t + \mu + \varepsilon_t, \tag{IV.57}$$

where $\tau_l' \in \mathbb{R}^r$. Denote by $H_{r,l}^0$, $H_{r,l}^0 \subseteq H_{r,l}$, the case where $H_{r,l}$ and Assumption IV.2.1 hold. Then under $H_{r,l}^0$, X_t has the representation,

$$X_t = C_\Sigma \sum_{i=1}^{t} \varepsilon_t + C_\Sigma t + C_0 + C_S S_{t,l}, \tag{IV.58}$$

where the cointegrating relations $S_{t,l} = \beta' X_t$ are asymptotically stable around a linear trend,

$$S_{t,l} = (I + \beta'\alpha) S_{t-1,l} + \beta'\tau t + \beta'\mu + \varepsilon_t. \tag{IV.59}$$

That is, X_t is trending-I(1) and $\beta' X_t$ also have a linear trend.

Under $H_{r,l}$ the VAR(1) model is given by,

$$\begin{aligned}
\Delta X_t &= \alpha\beta' X_{t-1} + \alpha\tau_l' t + \mu + \varepsilon_t &\text{(IV.60)}\\
&= \alpha \begin{pmatrix} \beta \\ \tau_l \end{pmatrix}' \begin{pmatrix} X_{t-1} \\ t \end{pmatrix}' + \mu + \varepsilon_t &\text{(IV.61)}\\
&= \alpha\beta_l' X_{t-1,l} + \mu + \varepsilon_t, &\text{(IV.62)}
\end{aligned}$$

and hence the MLE of α, $\beta_l = (\beta', \tau_l')$, μ and Ω are found by RRR of ΔX_t on $X_{t-1,l}$, both corrected by OLS regression on the constant. The unrestricted model $H_{p,l}$ is given by (IV.56), and limiting distribution of the LR test of $H_{r,l}$ against $H_{p,l}$ converge in distribution under $H_{r,l}^0$,

$$\text{LR}_{r,l} \equiv \text{LR}(H_{r,l}|H_{p,l}) \xrightarrow{D} DF_{p-r}^l(\mathcal{W}), \tag{IV.63}$$

where

$$DF_{p-r}^l(\mathcal{W}) = \text{tr}\{\int_0^1 d\mathcal{W}\mathcal{W}^{l\prime}(\int_0^1 \mathcal{W}_u^l \mathcal{W}_u^{l\prime} du)^{-1} \int_0^1 \mathcal{W}^l d\mathcal{W}'\}, \tag{IV.64}$$

with $\mathcal{W}$ a $p-r$ dimensional standard Brownian motion, and $\mathcal{W}_u^l = (\mathcal{W}_u - \int_0^1 \mathcal{W}_s ds, u - 1/2)'$. Some quantiles of DF^l are reported in Section IV.5.4.

IV.5.3 The VAR(k) model

Consider the p-dimensional VAR(k) model given by,

$$X_t = A_1 X_{t-1} + ... + A_k X_{t-k} + \varepsilon_t, \quad t = 1, ..., T \tag{IV.65}$$

where $A_i \in \mathbb{R}^{p \times p}$, $X_0, ...X_{-k+1}$ are fixed and ε_t are i.i.d.$\text{N}_p(0, \Omega)$, $\Omega > 0$. As before, to allow for roots at $z = 1$ in the characteristic polynomial, reparametrize the model as,

$$\Delta X_t = \Pi X_{t-1} + \Gamma_1 \Delta X_{t-1} + ... + \Gamma_{k-1} \Delta X_{t-k+1} + \varepsilon_t, \tag{IV.66}$$

where $\Pi, \Gamma_i \in \mathbb{R}^{p \times p}$ with $\Pi = \sum_{i=1}^k A_i - I_p \in$ and $\Gamma_i = -\sum_{j=i+1}^k A_j \in \mathbb{R}$ for $i = 1, ..., k-1$. The characteristic polynomial is given by

$$A(z) = (1-z) I_p - \Pi z - \Gamma_1 (1-z) z - ... - \Gamma_{k-1} (1-z) z^k, \tag{IV.67}$$

and it immediately follows that $\det(A(1)) = 0$ if, and only if, Π has reduced rank.

IV.5.3.1 Representation of VAR(k) processes

The generalization of Theorem IV.2.2 is given by:

Theorem IV.5.1 *Consider the VAR(k) process given by (IV.66) under the hypothesis*

$$H_r : \Pi = \alpha\beta', \quad \alpha, \beta \in \mathbb{R}^{p \times r}, \ r < p. \tag{IV.68}$$

Then if Assumption IV.2.1 holds, X_t is an I(1) process. Moreover, it has the representation,

$$X_t = C_\Sigma \sum_{i=1}^t \varepsilon_i + C_S S_t + C_0, \tag{IV.69}$$

where $C_\Sigma = \beta_\perp (\alpha'_\perp \Gamma \beta_\perp)^{-1} \alpha'_\perp$ is a $p \times p$ dimensional matrix of rank $(p - r)$ and $\Gamma = (I - \sum_{j=1}^{k-1} \Gamma_j)$. The $(r + p(k-1))$ dimensional process $S_t = (X_t'\beta, \Delta X'_{t-1}, ..., \Delta X'_{t-k+1})'$ is asymptotically stable. In particular, S_0 can be given an initial distribution such that S_t and the cointegrating relations $\beta' X_t$ have a stationary representation. Moreover, C_S is a $p \times (r + p(k-1))$ matrix, and C_0 depends on the initial values $X_0, \Delta X_0, ..., \Delta X_{-k+2}$, and satisfies $\beta' C_0 = 0$.

Remark IV.5.1 *With*

$$A = \begin{pmatrix} A_1 & \cdots & A_{k-1} & A_k \\ I_p & & & \\ & \ddots & & \\ & & I_p & 0 \end{pmatrix}$$

the condition is equivalent to A having $(p - r)$ eigenvalues equal to one, and the remaining smaller than one in absolute value.

Proof of Theorem IV.5.1: The proof follows by mimicking the proof of Theorem IV.2.1 by writing the VAR(k) process as a $pk-$dimensional VAR(1) process. Specifically, for $k = 2$, define $X_t^* = (X_t', X_{t-1}')'$, then $\Pi^* = A - I$, with A given by (). Hence,

$$\Delta X_t^* = \Pi^* X_{t-1}^* + \varepsilon_t, \quad \Pi^* = \alpha^*(\beta^*)', \text{ and} \tag{IV.70}$$

$$\alpha^* = \begin{pmatrix} \alpha & \Gamma_1 \\ 0 & I_p \end{pmatrix}, \beta^* = \begin{pmatrix} \beta & I_p \\ 0 & -I_p \end{pmatrix}, \tag{IV.71}$$

with $\varepsilon_t^* = (\varepsilon_t', 0)'$. By the proof of Theorem IV.2.1, X_t^* has the representation,

$$X_t^* = C^* \sum_{i=1}^{t} \varepsilon_t^* + C_S^* S_t^* + C_0^*, \tag{IV.72}$$

with $C^* = \beta_\perp^*(\alpha_\perp^{*\prime}\beta_\perp^*)^{-1}\alpha_\perp^{*\prime}$, $C_S^* = \alpha^*(\beta^{*\prime}\alpha^*)^{-1}$ and $C_0^* = C^* X_0^*$. Note that $\alpha_\perp^*$ and $\beta_\perp^*$ are given by,

$$\alpha_\perp^* = \begin{pmatrix} \alpha_\perp \\ -\Gamma_1'\alpha_\perp \end{pmatrix}, \beta_\perp^* = \begin{pmatrix} \beta_\perp \\ \beta_\perp \end{pmatrix} \tag{IV.73}$$

Finally, use that $X_t = (I, 0)X_t^*$ to derive the result for X_t as desired. □

IV.5.3.2 Estimation in the VAR(k) model

Rewrite the VAR(k) model in (IV.66) under H_r as,

$$\Delta X_t = \alpha\beta' X_{t-1} + \Gamma^* \Delta X_{t-1}^* + \varepsilon_t, \tag{IV.74}$$

with $\Gamma^* = (\Gamma_1, ..., \Gamma_k)$, and similarly, $\Delta X_{t-1}^* = (\Delta X_{t-1}', ..., \Delta X_{t-k+1}^{'})'$. With α, β known, the MLE $\hat{\Gamma}^*(\alpha, \beta)$ is found by OLS regression of $\Delta X_t - \alpha\beta' X_{t-1}$ on ΔX_{t-1}^*. Introduce therefore the corresponding residuals,

$$Y_t = R(\Delta X_t | \Delta X_{t-1}^*) \text{ and } Z_t = R(X_{t-1} | \Delta X_{t-1}^*), \tag{IV.75}$$

from the OLS regressions. Then $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\Omega}$ are found by RRR of Y_t on Z_t, see Theorem IV.2.2.

IV.5.4 Examples and quantiles for rank testing

Summarizing the discussion above, three different models were of interest which with $\Delta X_{t-1}^{*} = (\Delta X_{t-1}', .., \Delta X_{t-k+1}')'$ and $\Gamma^{*} = (\Gamma_1, .., \Gamma_{k-1})$ can be rewritten as,

$$\begin{aligned} H_p : \ & \Delta X_t = \Pi X_{t-1} + \Gamma^{*} \Delta X_{t-1}^{*} + \varepsilon_t & \text{(IV.76)} \\ H_{p,c} : \ & \Delta X_t = \Pi X_{t-1} + \mu + \Gamma^{*} \Delta X_{t-1}^{*} + \varepsilon_t & \text{(IV.77)} \\ H_{p,l} : \ & \Delta X_t = \Pi X_{t-1} + \tau t + \mu + \Gamma^{*} \Delta X_{t-1}^{*} + \varepsilon_t & \text{(IV.78)} \end{aligned}$$

The hypotheses of interest are for each model given by,

$$\begin{aligned} & H_r : \Pi = \alpha\beta' & \text{(IV.79)} \\ & H_{r,c} : \Pi = \alpha\beta', \ \mu = \alpha\mu_c' & \text{(IV.80)} \\ & H_{r,l} : \Pi = \alpha\beta', \ \tau = \alpha\tau_l' & \text{(IV.81)} \end{aligned}$$

The limit distributions of the likelihood ratio tests of H_r against H_p, $H_{r,c}$ against $H_{p,c}$ and $H_{r,l}$ against $H_{p,l}$, under the assumption of Theorem IV.5.1 are given by,

$$\mathrm{DF}_{p-r}(\mathcal{W}), \ \mathrm{DF}_{p-r}^{c}(\mathcal{W}) \ \text{ and } \mathrm{DF}_{p-r}^{l}(\mathcal{W}) \qquad \text{(IV.82)}$$

respectively, where $\mathcal{W}$ is a $(p-r)$ dimensional standard Brownian motion. The quantiles of (IV.82) given below are from Johansen (1996).

IV.5.4.1 The spot and futures data

Returning to the MIB30 data, consider the analysis of $X_t = (s_t, f_t)'$ by a VAR(6) model with a linear trend:

$$\Delta X_t = \Pi X_{t-1} + \tau t + \mu + \Gamma_1 \Delta X_{t-1} + ... + \Gamma_5 \Delta X_{t-5} + \varepsilon_t. \qquad \text{(IV.83)}$$

In order to find the number of possible cointegrating relations, the following rank statistics were computed:

$$\mathrm{LR}_{0,l} = 21.5, \quad \mathrm{LR}_{1,l} = 5.1. \qquad \text{(IV.84)}$$

As $\mathrm{LR}_{0,l} = 21.5 > ...$, but $\mathrm{LR}_{1,l} = 5.1 < 9.1$, the hypothesis of one cointegrating vector is accepted.

The cointegrating vector, and its linear trend coefficient, are given by

$$(\hat{\beta}', \hat{\tau}_l) = \left(1, -0.99, 6 \cdot 10^{-6}\right). \qquad \text{(IV.85)}$$

Consider the LR test for the linear hypothesis given by,

$$H_{\text{lin}} : (\hat{\beta}', \hat{\tau}_l) = \varphi(1, -1, 0), \qquad \text{(IV.86)}$$

that is, the linear trend can be omitted in the cointegrating relation, and the spread $s_t - f_t$ is indeed asymptotically stable, or stationary. Computation gives, $\text{LR}_{\text{lin}} = 1.9$, and as $1.9 < 6$, the 95% quantile of the χ^2_2 distribution, the hypothesis is clearly accepted.

Asymptotic Quantiles of LR Test for Rank (IV.87)

$\text{LR}_r \equiv \text{LR}(H_r \mid H_p)$			
$p - r$	95% quantile	97.5 % quantile	99 % quantile
1	4.2	5.3	7.0
2	12.2	13.9	16.1
3	24.0	26.4	29.1
4	39.7	42.5	46.0
5	59.2	62.6	66.7

(IV.88)

$\text{LR}_{r,c} \equiv \text{LR}(H_{r,c} \mid H_{p,c})$			
$p - r$	95% quantile	97.5 % quantile	99 % quantile
1	9.1	10.7	12.7
2	20.0	22.0	24.7
3	34.8	37.5	40.8
4	53.4	56.5	60.4
5	75.7	79.6	83.9

(IV.89)

$\text{LR}_{r,l} \equiv \text{LR}(H_{r,l} \mid H_{p,l})$			
$p - r$	95% quantile	97.5 % quantile	99 % quantile
1	12.3	14.1	16.3
2	25.4	27.8	30.6
3	42.2	45.0	48.5
4	62.6	66.0	70.2
5	86.9	90.8	95.3

(IV.90)

IV.5.4.2 Money demand

Return to the first page of Part I, where it was emphasized that it is of interest to see whether relations such as the money demand type relations exist.

With money stock M, prices P, real income Y and interest rate i, it was postulated that often one finds 'stable' relations between observed variables of the form such as,

$$m_t - p_t - b_1 y_t - b_2 i_t,$$

where the coefficients b_i are estimated from data, and $m_t = \log M_t$, $p_t = \log P_t$

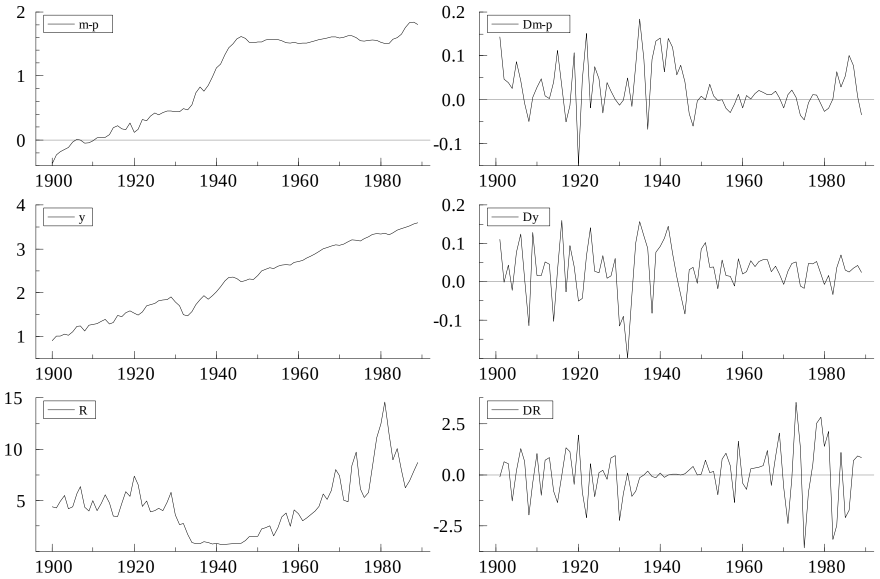


Figure 3: Data from Hayashi (2000). In the graphs $m - p = \log(M_1/P)$, R is the annual rate, $y = \log Y$, and Y is the net national product.

and $y_t = \log Y_t$. Consider annual US data for the period 1900-1989 shown in Figure 3.

By estimation of a VAR(2) model which allows for a linear trend for $X_t = ((m-p)_t, y_t, R_t)$ the following rank test statistics were found,

$$\mathrm{LR}_{0,l} = 48.5, \quad \mathrm{LR}_{1,l} = 18.3, \quad \mathrm{LR}_{2,l} = 3.2. \tag{IV.91}$$

Hence $\hat{r} = 1$, with

$$\hat{\beta}' X_t + \hat{\tau}_l t = (m-p)_t - 1.61 y_t + 0.11 R_t + 0.02t. \tag{IV.92}$$

The LR test statistic of the exclusion of the linear trend, $\tau_l = 0$, and $(m - p - y)$ cointegrating is given by

$$\mathrm{LR}\,(\tau_l = 0) = 3.3, \tag{IV.93}$$

and hence as the LR test is asymptotically χ^2_2 distributed this is accepted. The cointegrating relation is given by

$$\hat{\beta}' X_t = (m-p)_t - y_t + 0.11 R_t. \tag{IV.94}$$

References

[1] Hayashi, F.(2000), Econometrics, Princeton University Press.

[2] Johansen, S. (1996), Likelihood-based Inference in Cointegrated Vector Autoregressive Models, Oxford University Press.

[3] Johansen, S. (2005), Cointegration: A Survey, Handbook of Econometrics.

[4] Paruolo, P. (1997), Asymptotic Inference on the Moving Average Impact Matrix in Cointegrated I(1) VAR Systems, *Econometric Theory,* 13:79-118.

A Results from Linear Algebra

Some well-known results that are used in the text are briefly mentioned here.

Let M be a $m \times n$ matrix of rank r. Then with $rank(M) = r(M) = r$, it holds that:

1. $r(M) = r \leq \min(m, n)$
2. $r(M) = r(MM') = r(M'M) = r(M') = r$
3. $r(MB) \leq \min(r(M), r(B))$
4. $r(MB) = r(M) = r$ if B is a $n \times n$ matrix of full rank

For the column space of M,

$$sp(M) = \{Mx \mid x \in \mathbb{R}^n\},$$

it holds that,

$$\begin{aligned} \dim(sp(M)) &= r(M) = r \\ sp(M) &= sp(MM') \end{aligned}$$

Also the direct sum applies, that is:

$$\begin{aligned} m &= \dim(sp(M)) + \dim(sp(M)_\perp) \\ &= r + (m - r) \end{aligned}$$

This is used explicitly in cointegration in the following way. Let α be a $m \times r$ matrix with rank r so that $sp(\alpha) = sp(M)$. Then define $\alpha_\perp$ as a $m \times (m - r)$ matrix of full rank $(m - r)$, such that $sp(\alpha_\perp) = sp(M)_\perp$ ie.

$$\alpha'\alpha_\perp = 0 \quad \text{and} \quad sp(\alpha, \alpha_\perp) = \mathbb{R}^m$$

In particular, the orthogonal projection is often applied:

$$I_m = \alpha(\alpha'\alpha)^{-1}\alpha' + \alpha_\perp(\alpha'_\perp\alpha_\perp)^{-1}\alpha'_\perp.$$

A.1 Diagonalization

With M be a symmetric $p \times p$ matrix, then M is diagonalizable. The eigenvalue problem is given by,

$$\det(\lambda I_p - M) = 0, \tag{IV.95}$$

which is solved for eigenvalues $\lambda_1 \geq \lambda_2 \geq .. \geq \lambda_p$ with corresponding eigenvectors $v_1, .., v_p$. Then with $V = (v_1, .., v_p)$ and $\Lambda =$diag$(\lambda_1, .., \lambda_p)$,

$$V'MV = \Lambda \tag{IV.96}$$

$$MV = V\Lambda \tag{IV.97}$$

$$VV' = V'V = I_p. \tag{IV.98}$$

A generalized version of the eigenvalue problem, used in cointegration analysis and reduced rank regression (RRR), is given by:

$$\det(\rho N - M) = 0, \tag{IV.99}$$

where M is as before, while N is a positive definite (and hence in particular symmetric) $p \times p$ matrix. The generalized eigenvalue problem has eigenvalues $\rho_1 \geq ... \geq \rho_p$, with $W = (w_1, ..., w_p)$ the corresponding eigenvectors, and it holds that

$$W'MW = R \tag{IV.100}$$

$$MW = NWR \tag{IV.101}$$

$$W'NW' = I_p, \tag{IV.102}$$

where $R =$diag$(\rho_1, ..., \rho_p)$.

That the generalized eigenvalue problem has this solution, follows easily by noting that $N^{-1/2}MN^{-1/2}$ is symmetric and hence diagonalizable, and,

$$\det(\rho N - M) = 0 \Leftrightarrow \det\left(\rho I - N^{-1/2}MN^{-1/2}\right) = 0.$$

B General Decomposition

Let Π be a $p \times p$ matrix with rank r possibly less than p.

Then a 'Singular Value Decomposition' holds:

Lemma B.1 *With Π a $p \times p$ matrix of rank r, there exist $p \times r$ matrices A, B with full rank r and a $r \times r$ diagonal matrix $\Lambda = diag(\lambda_1, .., \lambda_r)$, where $\lambda_i > 0$ such that*

$$\Pi = A\Lambda^{1/2}B' \tag{IV.103}$$

It holds that $sp(B) = sp(\Pi')$ (row space) and $sp(A) = sp(\Pi)$ (column space).

The decomposition in cointegration of Π as $a\beta'$ then holds as a corollary:

Corollary B.1 *With* Π *a* $p \times p$ *matrix of rank* r, $p \times r$ *matrices* α, β *of rank* r *exist such that*

$$\Pi = \alpha\beta' \tag{IV.104}$$

where $sp(\beta) = sp(\Pi')$*.*

Proof of Lemma B.1:

$\Pi\Pi'$ is symmetric and positive definite with rank r. Diagonalize $\Pi\Pi'$, which has eigenvalues and eigenvectors,

$$\begin{aligned} \lambda_1 \geq .. \geq \lambda_r \ , V_1 = (v_1, .., v_r) \\ \lambda_{r+1} = ... = \lambda_p = 0, \ V_2 = (v_{r+1}, .., v_p) \end{aligned}$$

Then with $\Lambda = diag(\lambda_1, .., \lambda_r)$,

$$\Pi\Pi' V_1 = V_1 \Lambda \tag{IV.105}$$

$$\Pi\Pi' V_2 = 0 \tag{IV.106}$$

Note that (IV.106) implies $V_2 V_2' \Pi = 0$ (use e.g.. $r(V_2'\Pi) = r(V_2'\Pi\Pi' V_2) = 0$). Define $B = \Pi' V_1 \Lambda^{-1/2}$ and set $A = V_1$. Then the desired decomposition holds by

$$\begin{aligned} \Pi &= (V_1, V_2)(V_1, V_2)'\Pi = V_1 V_1' \Pi \\ &= V_1 \Lambda^{1/2} \Lambda^{-1/2} V_1' \Pi = A \Lambda^{1/2} B' \end{aligned}$$

□

Anders Rahbek
Rasmus Søndergaard Pedersen
University of Copenhagen

September 2024

Part V
SREs: Stationarity, Ergodicity, Tails and Limit Theory

V.1 Stochastic Recurrence Equations

In the previous chapters, we considered the stochastic properties of Markov chains $\{X_t\}_{t=0,1,\ldots}$. It was shown that geometric ergodicity implies that X_0 can be assigned a particular distribution, $X_0 \overset{D}{=} X_0^*$ such that the resulting process $\{X_t^*\}_{t=0,1,\ldots}$ is stationary. Moreover, the LLN in Theorem I.4.2 stated that for any initial value of X_0, $T^{-1}\sum_{t=1}^{T} X_t \xrightarrow{p} \mathbb{E}[X_t^*]$ provided that $\mathbb{E}[|X_t^*|] < \infty$. That is, under geometric ergodicity the initial value of the Markov chain plays no role for the stochastic limit results. An important consequence of the geometric ergodicity is that $\{X_t^*\}_{t=0,1,\ldots}$ is so-called ergodic (which we define below). In particular, there are LLNs and CLTs that apply to stationary and ergodic processes parallel to Theorems I.4.2 and I.4.4.

In this chapter we consider a specific class of Markov chains for $X_t \in \mathbb{R}^d$ given by

$$X_t = A_t X_{t-1} + B_t,$$

for some initial value X_0, where A_t is a $d \times d$ random matrix and B_t is a d-dimensional random vector, and $\{(A_t, B_t)\}_{t=1,2,\ldots}$ is an i.i.d. process such that (A_t, B_t) and $\{X_{t-1}, X_{t-2}, \ldots, X_0\}$ are independent for all $t \geq 1$. The above equation for X_t is a so-called *stochastic recurrence equation* (SRE), and, as illustrated later in Section V.3, this class of processes covers a wide range of processes applied in financial econometric modelling. The aim of the chapter is to provide conditions such that the stationary version, $\{X_t^*\}_{t=0,1,\ldots}$, of $\{X_t\}_{t=0,1,\ldots}$ exists, with $X_t^* = A_t X_{t-1}^* + B_t$. Moreover, we present an *explicit expression* for the stationary solution to the SRE, X_t^*, and provide conditions ensuring finite moments and ergodicity such that a LLN and a CLT apply to $\{X_t^*\}_{t=0,1,\ldots}$. The conditions for stationarity and finite moments are sharp in the sense that they are essentially necessary. Moreover, the stated conditions enable us to characterize the tail shape of the unconditional distribution of X_t^*. Importantly, under mild conditions the distribution is *heavy-tailed* which is of particular relevance in actuarial sciences and risk management and mimics the tails of the empirical distribution of, say, equity returns observed in practice.

Most of the results stated in the chapter can be found in the recent textbook by Burazcewski et al. (2016) [BDM henceforth] that provides a comprehensive treatment of SREs as well as detailed proofs of the stated results. We emphasize that conditions ensuring geometric ergodicity of $\{X_t\}_{t=0,1,\ldots}$ can be derived by applying the drift criterion; cf. Part I and BDM (Section 2.2). The conditions for stationarity and finite moments provided in this chapter are – in addition to provide further knowledge about X_t^* – typically milder as they, for instance, do not require the Markov chain to have a positive, continuous transition density.

Our main focus is to consider cases where the SRE process is stationary, and we state limit results only in terms of the stationary version of the process. In order to ease the notation, throughout, whenever a given process is stationary we write it in terms of X_t and not X_t^*. Moreover, as is standard in the literature, we let stationary processes run over all integers $\mathbb{Z}$ instead of $\{0, 1, \ldots\}$.[1] Consequently, we pay attention to processes $\{X_t\}_{t\in\mathbb{Z}}$ of the form

$$X_t = A_t X_{t-1} + B_t, \quad t \in \mathbb{Z}, \tag{V.1}$$

with $X_t \in \mathbb{Z}^d$ and with A_t and B_t given as above. It is assumed that $\{(A_t, B_t)\}_{t\in\mathbb{Z}}$ is an i.i.d. process and (A_t, B_t) and $\{X_{t-1}, X_{t-2}, \ldots\}$ are independent for all $t \in \mathbb{Z}$. We say that X_t obeys an SRE, if it satisfies (V.1).

V.2 Ergodicity and Limit Theory

In this section we define the notion of ergodicity and present limit theorems for stationary and ergodic processes. These results will be used throughout in the remainder of the course when considering statistical inference in models for time-varying conditional volatility, such as ARCH. We emphasize that there are different ways of defining ergodicity. We say that a stationary process $\{X_t\}_{t\in\mathbb{Z}}$, with $X_t \in \mathbb{R}^d$, is ergodic if and only if for every measurable function $f : (\mathbb{R}^d)^\infty \to \mathbb{R}$ with $\mathbb{E}[|f(\ldots, X_{t-1}, X_t, X_{t+1}, \ldots)|] < \infty$, as $T \to \infty$

$$\frac{1}{T}\sum_{t=1}^{T} f(\ldots, X_{t-1}, X_t, X_{t+1}, \ldots) \overset{a.s.}{\to} \mathbb{E}\left[f(\ldots, X_{t-1}, X_t, X_{t+1}, \ldots)\right], \tag{V.2}$$

[1] This is without loss of generality: The process $\{X_t^*\}_{t=0,1,\ldots}$ is stationary if and only if, there exists a stationary process $\{Y_t\}_{t\in\mathbb{Z}}$ satisfying $\{X_t^*\}_{t=0,1,\ldots} \overset{D}{=} \{Y_t\}_{t=0,1,\ldots}$.

where $\overset{a.s.}{\rightarrow}$ denotes almost sure convergence[2]. By definition, ergodicity ensures that a (strong) LLN applies to the stationary process $\{X_t\}_{t\in\mathbb{Z}}$. This parallels geometric ergodicity of Markov chains that ensures that a LLN holds (Theorem I.4.2). Recall that geometric ergodicity implies the "mixing" property that X_t and X_{t+h} are independent as $h \to \infty$. Likewise, ergodicity implicitly ensures that observations cannot be "too dependent". To see this, consider the process $\{X_t\}_{t\in\mathbb{Z}}$, with $X_t \in \mathbb{R}$ such that $\mathbb{P}(X_t = Z) = 1$ for some non-degenerate random variable Z with $\mathbb{E}[|Z|] < \infty$. Clearly the process is stationary, but it is not ergodic, as the average of the perfectly dependent process $T^{-1}\sum_{t=1}^{T} X_t = Z$ (almost surely) does not converge to $\mathbb{E}[X_t] = \mathbb{E}[Z]$.

As mentioned in the introduction, the stationary version of a geometric ergodic process is ergodic:

Corollary V.2.1 *Suppose that a Markov chain $\{Y_t\}_{t=0,1,\ldots}$ satisfies the drift criterion in Part I such that it is geometrically ergodic. Then there exists a stationary and ergodic process $\{X_t\}_{t\in\mathbb{Z}}$ with $\{X_t\}_{t=0,1,\ldots} \overset{d}{=} \{Y_t^*\}_{t=0,1,\ldots}$.*

Before presenting a LLN and CLT for stationary and ergodic processes, we state the following result that any nice deterministic function of a stationary and ergodic process yields itself a stationary and ergodic process:

Theorem V.2.1 *With $X_t \in \mathbb{R}^d$, let $\{X_t\}_{t\in\mathbb{Z}}$ be a stationary and ergodic process. For some measurable $f : (\mathbb{R}^d)^\infty \to \mathbb{R}$, let*

$$Y_t = f(\ldots, X_{t-1}, X_t, X_{t+1}, \ldots), \quad t \in \mathbb{Z}.$$

Suppose that Y_t is finite almost surely (e.g., $\mathbb{E}[|Y_t|] < \infty$) for some t. Then the process $\{Y_t\}_{t\in\mathbb{Z}}$ is stationary and ergodic.

Note that by definition of ergodicity, (V.2) ensures that a (strong) LLN applies. For convenience, we state this as a theorem, typically referred to as the Ergodic Theorem.[3]

Theorem V.2.2 *With $X_t \in \mathbb{R}^d$, let $\{X_t\}_{t\in\mathbb{Z}}$ be a stationary and ergodic process. For some measurable $f : (\mathbb{R}^d)^\infty \to \mathbb{R}$, let $Y_t = f(\ldots, X_{t-1}, X_t, X_{t+1}, \ldots)$, $t \in \mathbb{Z}$. Suppose that $\mathbb{E}[|Y_t|] < \infty$. Then as $T \to \infty$,*

$$\frac{1}{T}\sum_{t=1}^{T} Y_t \xrightarrow{p} \mathbb{E}\left[Y_t\right].$$

[2] That is, $P\left(\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} f(\ldots, X_{t-1}, X_t, X_{t+1}, \ldots) = \mathbb{E}\left[f(\ldots, X_{t-1}, X_t, X_{t+1}, \ldots)\right]\right) = 1$. You may recall that almost sure convergence implies convergence in probability.

[3] Note that by definition of ergodicity, the convergence holds almost surely. We state the weaker result that the average converges with probability approaching one, as we nowhere need almost sure convergence when considering the stochastic properties of estimators later on.

We also have the following CLT for martingale differences.

Theorem V.2.3 *With $X_t \in \mathbb{R}^d$, let $\{X_t\}_{t\in\mathbb{Z}}$ be a stationary and ergodic process. For some measurable $f : (\mathbb{R}^d)^\infty \to \mathbb{R}$, let*

$$Y_t = f(\dots, X_{t-1}, X_t, X_{t+1}, \dots), \quad t \in \mathbb{Z},$$

and let $\mathcal{F}_t$ denote the natural filtration generated by $\{X_s\}_{s\leq t}$. Assume that $\mathbb{E}[Y_t|\mathcal{F}_{t-1}] = 0$ a.s. and $0 < \mathbb{E}[Y_t^2] < \infty$. Then as $T \to \infty$,

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} Y_t \xrightarrow{D} N(0, \mathbb{E}[Y_t^2]).$$

The result is essentially a corollary to Theorem I.4.4. In particular with $Y_t = X_t$, we see that the conditions in that theorem are easily checked:

$$\text{(i)} \quad \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[X_t^2|\mathcal{F}_{t-1}] \xrightarrow{p} \mathbb{E}[\mathbb{E}[X_t^2|\mathcal{F}_{t-1}]] = \mathbb{E}[X_t^2] > 0,$$

where we have used that $\mathbb{E}[\mathbb{E}[X_t^2|\mathcal{F}_{t-1}]] < \infty$ and Theorem V.2.2. Likewise,

$$\begin{aligned}\text{(ii)} \quad \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[X_t^2 \mathbb{I}(|X_t| > \delta\sqrt{T})\right] &= \mathbb{E}[X_t^2 \mathbb{I}(|X_t| > \delta\sqrt{T})] \\ &= \int x^2 \mathbb{I}(|x| > \delta\sqrt{T}) dP \to 0 \quad \text{as } T \to \infty,\end{aligned}$$

where we have used that $\mathbb{I}(|x| > \delta\sqrt{T}) \to 0$, $\mathbb{E}[X_t^2] = \int x^2 dP < \infty$ as well as dominated convergence.

Remark V.2.1 *Ergodicity does not ensure that a CLT for non-martingale differences, such as Theorem I.4.3, holds. In order to have such a result, one would typically rely on showing that the process is geometrically ergodic by relying on the drift criterion.*

V.3 Examples

The following examples illustrate that many time series processes within financial econometrics belongs to the class of SREs.

Example V.3.1 (AR(1)) *Consider the AR(1) process from Part I with*

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

with $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ *and i.i.d. process with* $\mathbb{E}[\varepsilon_t] = 0$ *and* $\mathbb{V}[\varepsilon_t] = \sigma^2$ *with* $0 < \sigma^2 < \infty$*. We see that this process is given by and SRE with* $d = 1$*,* $A_t = \rho$ *(constant) and* $B_t = \varepsilon_t$*.*

Example V.3.2 (VAR(1)) *Consider the VAR(1) from Part I with*

$$X_t = AX_{t-1} + \varepsilon_t,$$

with $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ *an i.i.d. process with* $\mathbb{E}[\varepsilon_t] = 0$ *and* $\mathbb{V}[\varepsilon_t] = \Omega$ *positive definite. The process obeys an SRE with* $A_t = A$ *and* $B_t = \varepsilon_t$*.*

Example V.3.3 (ARCH(1)) *Consider Engle's (1982) ARCH(1) process from Part I given by*

$$\begin{aligned} x_t &= \sigma_t z_t, \\ \sigma_t^2 &= \omega + \alpha x_{t-1}^2, \end{aligned}$$

with $\alpha \geq 0$*,* $\omega > 0$*,and* $\{Z_t\}_{t\in\mathbb{Z}}$ *an i.i.d. process with* $Z_t \overset{D}{=} N(0,1)$*. Recall that (almost surely)* $\mathbb{E}[x_t|x_{t-1}] = 0$ *and* $\mathbb{E}[x_t^2|x_{t-1}] = \sigma_t^2$*. In fact,*

$$x_t|x_{t-1} \overset{D}{=} N(0, \sigma_t^2).$$

Consider the random vector

$$\begin{pmatrix} A_t \\ B_t \end{pmatrix} \overset{D}{=} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \alpha & 0 \\ 0 & \omega \end{bmatrix} \right),$$

Then with $X_t = x_t \in \mathbb{R}$ *satisfying (V.1), we have that*

$$x_t|x_{t-1} \overset{D}{=} N(\mathbb{E}[x_t|x_{t-1}], \mathbb{V}[x_t|x_{t-1}]),$$

with (almost surely)

$$\begin{aligned} \mathbb{E}[x_t|x_{t-1}] &= \mathbb{E}[A_t x_{t-1} + B_t|x_{t-1}] \\ &= \mathbb{E}[A_t|x_{t-1}]x_{t-1} + \mathbb{E}[B_t|x_{t-1}] \\ &= 0x_{t-1} + 0 \\ &= 0, \end{aligned}$$

and

$$\begin{aligned}\mathbb{V}[x_t|x_{t-1}] &= \mathbb{E}[x_t^2|x_{t-1}] \\ &= \mathbb{E}[(A_t x_{t-1} + B_t)^2|x_{t-1}] \\ &= \mathbb{E}[A_t^2|x_{t-1}]x_{t-1}^2 + \mathbb{E}[B_t^2|x_{t-1}] + 2\mathbb{E}[A_t B_t|x_{t-1}]x_{t-1} \\ &= \alpha x_{t-1}^2 + \omega + 2 \times 0 \times x_{t-1} \\ &= \omega + \alpha x_{t-1}^2.\end{aligned}$$

We conclude that the ARCH(1) process as an SRE representation. Consequently the stochastic properties of the ARCH(1) process, such as stationarity, ergodicity and finite moments, can be derived by making use of the SRE representation.

Example V.3.4 (BEKK) *Consider the $(d+1)$-dimensional vector*

$$\begin{pmatrix} m_t \\ B_t \end{pmatrix} \overset{D}{=} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & \Omega \end{bmatrix}\right),$$

with Ω constant and positive definite $(d \times d)$ matrix. For a constant $(d \times d)$ matrix A, let $A_t = m_t A$. Then with $X_t \in \mathbb{R}^d$ given by (V.1),

$$X_t|X_{t-1} \overset{D}{=} N(\mathbb{E}[X_t|X_{t-1}], \mathbb{V}[X_t|X_{t-1}]),$$

with (almost surely)

$$\begin{aligned}\mathbb{E}[X_t|X_{t-1}] &= \mathbb{E}[A_t X_{t-1} + B_t|X_{t-1}] \\ &= \mathbb{E}[A_t|X_{t-1}]X_{t-1} + \mathbb{E}[B_t|X_{t-1}] \\ &= 0AX_{t-1} + 0_{d\times 1} \\ &= 0_{d\times 1},\end{aligned}$$

and

$$\begin{aligned}&\mathbb{V}[X_t|X_{t-1}] \\ &= \mathbb{E}[X_t X_t'|X_{t-1}] \\ &= \mathbb{E}[(A_t X_{t-1} + B_t)\,(A_t X_{t-1} + B_t)'\,|X_{t-1}] \\ &= \mathbb{E}[A_t X_{t-1} X_{t-1}' A_t'|X_{t-1}] + \mathbb{E}[B_t B_t'|X_{t-1}] + \mathbb{E}[A_t X_{t-1} B_t' + B_t X_{t-1}' A_t'|X_{t-1}] \\ &= \mathbb{E}[m_t^2|X_{t-1}]AX_{t-1}X_{t-1}'A' + \mathbb{E}[B_t B_t'|X_{t-1}] + AX_{t-1}\mathbb{E}[m_t B_t'|X_{t-1}] + \mathbb{E}[B_t m_t|X_{t-1}]\,(AX_{t-1})' \\ &= \mathbb{E}[m_t^2]AX_{t-1}X_{t-1}'A' + \mathbb{E}[B_t B_t'] + AX_{t-1}\mathbb{E}[m_t B_t'] + \mathbb{E}[B_t m_t]\,(AX_{t-1})' \\ &= AX_{t-1}X_{t-1}'A' + \Omega.\end{aligned}$$

Similar to the ARCH(1) example given above, we may alternatively write the process as

$$\begin{aligned} X_t &= \Omega_t^{1/2} Z_t \\ \Omega_t &= \Omega + A X_{t-1} X_{t-1}' A, \end{aligned}$$

with $\{Z_t\}_{t\in\mathbb{Z}}$ *an i.i.d. process with* $Z_t \overset{D}{=} N(0, I_d)$ *and* Z_t *independent of* $\{X_{t-1}, X_{t-2}, \dots\}$*, and* $\Omega_t^{1/2}$ *is the (symmetric) square-root of* Ω_t*. Here* X_t *is given in terms of a multivariate ARCH process with a so-called Baba-Engle-Kraft-Kroner (BEKK) formulation of the conditional covariance matrix* $\mathbb{V}[X_t|X_{t-1}]$*, as originally considered in Engle and Kroner (1995). We return to this type of process later on when considering modelling of time varying conditional covariance matrices as used within dynamic allocation of assets.*

Example V.3.5 (DAR) *Consider the bivariate vector*

$$\begin{pmatrix} A_t \\ B_t \end{pmatrix} \overset{D}{=} N\left(\begin{bmatrix} \phi \\ 0 \end{bmatrix}, \begin{bmatrix} \alpha & 0 \\ 0 & \omega \end{bmatrix} \right),$$

with constants $\phi \in \mathbb{R}$, $\alpha \geq 0$, $\omega > 0$. *Then with* $X_t = x_t \in \mathbb{R}$ *satisfying (V.1), we have that*

$$x_t|x_{t-1} \overset{D}{=} N(\mathbb{E}[x_t|x_{t-1}], \mathbb{V}[x_t|x_{t-1}]),$$

with (almost surely),

$$\begin{aligned} \mathbb{E}[x_t|x_{t-1}] &= \mathbb{E}[A_t x_{t-1} + B_t|x_{t-1}] \\ &= \mathbb{E}[A_t|x_{t-1}]x_{t-1} + \mathbb{E}[B_t|x_{t-1}] \\ &= \mathbb{E}[A_t]x_{t-1} + 0 \\ &= \phi x_{t-1}, \end{aligned}$$

and

$$\begin{aligned} &\mathbb{V}[x_t|x_{t-1}] \\ &= \mathbb{E}[x_t^2|x_{t-1}] - (\mathbb{E}[x_t|x_{t-1}])^2 \\ &= \mathbb{E}[(A_t x_{t-1} + B_t)^2|x_{t-1}] - (\phi x_{t-1})^2 \\ &= \mathbb{E}[A_t^2 x_{t-1}^2|X_{t-1}] + \mathbb{E}[B_t^2|x_{t-1}] + \mathbb{E}[A_t B_t x_{t-1}|X_{t-1}] - (\phi x_{t-1})^2 \\ &= \mathbb{E}[A_t^2]x_{t-1}^2 + \mathbb{E}[B_t^2] + 0 - (\phi x_{t-1})^2 \\ &= (\alpha + \phi^2)x_{t-1}^2 + \omega - (\phi x_{t-1})^2 \\ &= \omega + \alpha x_{t-1}^2. \end{aligned}$$

Here X_t is given in terms of a so-called double autoregressive (DAR) process of order one with Gaussian innovations. An alternative formulation of X_t is

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

where $\{z_t\}_{t\in\mathbb{Z}}$ is an i.i.d. process with $z_t \overset{D}{=} N(0,1)$, and z_t is independent of $\{x_{t-1}, x_{t-2}, \dots\}$ for all t. This type of process has been studied in detail by Ling (2004) and extended to a multivariate setting in Nielsen and Rahbek (2014) for the modelling interest rate dynamics; see also Hansen (2021).

V.4 Properties of univariate SREs

In this section, we consider properties of $\{X_t\}_{t\in\mathbb{Z}}$ given by the SRE in (V.1). With $V_t = (A_t, B_t)$, we note that the process $\{V_t\}_{t\in\mathbb{Z}}$ generates a natural filtration $\mathcal{F}_t = \sigma(V_t, V_{t-1}, \dots)$, and clearly, by independence, V_{t+1} is unpredictable with respect to $\mathcal{F}_t$. Likewise, we may consider cases where $X_t \in \mathcal{F}_t$ and X_t is independent of $\{V_{t+1}, V_{t+2}, \dots\}$. When the latter holds, we say that $\{X_t\}_{t\in\mathbb{Z}}$ is a *causal* process (with respect to $\mathcal{F}_t$). From the recursions in (**??**),

$$X_t = \sum_{i=0}^{t-1} \prod_{j=0}^{i-1} A_{t-j} B_{t-i} + \prod_{j=0}^{t-1} A_{t-j} X_0,$$

It shows up that the process $\{X_t\}_{t\in\mathbb{Z}}$ has a stationary causal solution when the last term term becomes asymptotically negligible, that is, the term vanishes as $t \to \infty$. In this case, X_t is given by the infinite series

$$X_t = \sum_{i=0}^{\infty} \prod_{j=0}^{i-1} A_{t-j} B_{t-i}, \tag{V.3}$$

which by construction is causal. Heuristically, in the one-dimensional case we may think of asymptotic negligibility as $\prod_{j=0}^{t-1} A_{t-j} \xrightarrow{P} 0$ as $t \to \infty$. Using Markov's inequality and that $\{A_t\}_{t\in\mathbb{Z}}$ is an i.i.d. process for any $\epsilon > 0$ and some $\delta > 0$,

$$\mathbb{P}\left(\left| \prod_{j=0}^{t-1} A_{t-j} \right| > \epsilon \right) \leq \frac{\mathbb{E}\left[\left| \prod_{j=0}^{t-1} A_{t-j} \right|^{\delta} \right]}{\epsilon^{\delta}} = \frac{\left(\mathbb{E}\left[|A_t|^{\delta}\right]\right)^t}{\epsilon^{\delta}},$$

and we conclude that $\prod_{j=0}^{t-1} A_{t-j} \xrightarrow{P} 0$, if $\mathbb{E}[|A_t|^{\delta}] < 1$.

The next section provides conditions ensuring the existence of a stationary and ergodic causal solution to the SRE. For the ease of presentation we start out by considering results for the one dimensional case.

V.4.1 Stationarity and ergodicity

Let $\log^+(x) = \max\{\log(x), 0\}$, and note that, for instance, $\log^+(x) \leq |x|$. We have the following result.

Theorem V.4.1 (BDM, Theorem 2.1.3) *Consider the i.i.d. process* $\{(A_t, B_t)\}_{t\in\mathbb{Z}}$ *with* $(A_t, B_t)' \in \mathbb{R}^2$*-valued. Suppose that one of the following conditions holds:*

1. $\mathbb{P}(A_t = 0) > 0$;
2. $\mathbb{P}(A_t = 0) = 0$, $-\infty \leq \mathbb{E}[\log|A_t|] < 0$ *and* $\mathbb{E}[\log^+|B_t|] < \infty$.

Then the SRE in (V.1) has a stationary and ergodic causal solution given by (V.3).The infinite series in (V.3) converges almost surely for any $t \in \mathbb{Z}$.
On the contrary, assume that one of the following conditions holds:

1. $\mathbb{P}(A_t = 0) = 0$, $\mathbb{P}(B_t = 0) < 1$ *and* $0 \leq \mathbb{E}[\log|A_t|] < \infty$;
2. $\mathbb{E}[\log|A_t|] > -\infty$ *and* $\mathbb{E}[\log^+|B_t|] = \infty$.

Then no stationary causal solution to ($V.1$) *exists.*

One may loosely say that the condition $\mathbb{E}[\log|A_t|] < 0$ is the sufficient stationarity condition for the SRE. For $\delta > 0$, an application of Jensen's inequality gives that

$$\delta\mathbb{E}[\log|A_t|] = \mathbb{E}[\log|A_t|^\delta] \leq \log\mathbb{E}[|A_t|^\delta],$$

and we have that $\mathbb{E}[\log|A_t|] < 0$, provided that $\mathbb{E}[|A_t|^\delta] < 1$, which is the condition that we heuristically derived in the previous section.

Example V.4.1 (AR (1) ctd.) *We have that* $A_t = \rho$ *and* $\{B_t\}_{t\in\mathbb{Z}}$ *an i.i.d. process with* $\mathbb{E}[B_t] = 0$ *and* $\mathbb{E}[B_t^2] = \sigma^2 < \infty$. *Note that* $\mathbb{E}[\log|A_t|] = \log(|\rho|) < 0$, *if* $|\rho| < 1$, *and* $\mathbb{E}[\log^+|B_t|] \leq \mathbb{E}[|B_t|] \leq \sigma < \infty$. *Using Theorem V.4.1, we conclude that the AR(1) process is stationary and ergodic if* $|\rho| < 1$. *The causal solution is given by the infinite series*

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

identical to the one derived in (I.4) Likewise, we have that the process does not have a stationary causal solution, if $|\rho| \geq 1$.
Note that even if the AR(1) process has very heavy tailed errors ε_t, *it is still stationary and ergodic provided that* $|\rho| < 1$, *as we only require that*

$\mathbb{E}[\log^+ |\varepsilon_t|]$ *is finite. For instance, suppose that* ε_t *is Cauchy distributed such that* $\mathbb{E}[\varepsilon_t]$ *does not exist and* $\mathbb{E}[\varepsilon_t^2] = \infty$*. Then*

$$\begin{aligned}\mathbb{E}[\log^+ |\varepsilon_t|] &= \int_{-\infty}^{\infty} \log^+ |x| f_\varepsilon(x) dx \\ &= \int_{-\infty}^{\infty} \log^+ |x| \frac{1}{\pi(1+x^2)} dx \\ &= \frac{2}{\pi} \int_1^\infty \frac{\log(x)}{1+x^2} dx \approx 0.58 < \infty.\end{aligned}$$

Example V.4.2 (ARCH(1) ctd.) *If* $\alpha = 0$*, such that* $\mathbb{P}(A_t = 0) = 1$*, we have by Theorem V.4.1 that the ARCH(1) process is stationary and ergodic with* $X_t = B_t$ *(almost surely). Consider the case* $\alpha > 0$*. Let* $z \overset{D}{=} N(0,1)$*, and note that* $\mathbb{E}[\log^+ |B_t|] \leq \mathbb{E}[|B_t|] = \mathbb{E}[|\sqrt{\omega}z|] = \sqrt{\omega}\mathbb{E}[|z|] = \sqrt{\omega}\sqrt{2/\pi} < \infty$*. Moreover,* $\mathbb{E}[\log |A_t|] = \log(\alpha)/2 + \mathbb{E}[\log |z|]$*. We then have that* $\{x_t\}_{t\in\mathbb{Z}}$ *is stationary and ergodic if* $\log(\alpha)/2 + \mathbb{E}[\log |z|] < 0$*, or equivalently, if* $\alpha < \exp(-2\mathbb{E}[\log |z|]) = 3.56\ldots$*. In contrast, there exists no stationary causal solution to the SRE if* $\alpha \geq 3.56\ldots$*. In fact, in this case, if we consider (V.1) for only* $t \geq 1$ *with* x_0 *fixed, it can be shown that* $\sigma_t^2 = \omega + \alpha x_{t-1}^2 \to \infty$ *with probability tending to one as* $t \to \infty$*, and one may say that the conditional variance explodes; see Pedersen and Rahbek (2016, Supplementary Material) for more details.*

Alternatively, for the ARCH(1) process, we have that $x_t^2 = \sigma_t^2 z_t^2 = \omega z_t^2 + \alpha z_t^2 x_{t-1}^2$*. Hence, with* $X_t := x_t^2$*, we have the SRE in (V.1) with* $(A_t, B_t) = (\alpha z_t^2, \omega z_t^2)$*. Here* $\{X_t\}_{t\in\mathbb{Z}}$ *has a stationary and ergodic causal solution provided that* $E[\log |A_t|] = E[\log(\alpha z_t^2)] < 0$*. From (V.3) we have that*

$$x_t^2 = \sum_{i=0}^{\infty} \prod_{j=0}^{i-1} A_{t-j} B_{t-i} = \sum_{i=0}^{\infty} \prod_{j=0}^{i-1} (\alpha z_{t-j}^2) \left(\omega z_{t-i}^2\right) = \omega \sum_{i=0}^{\infty} \alpha^i \prod_{j=0}^{i} z_{t-i}^2,$$

identical to the expression for x_t^2 *found in (I.8).*

V.4.2 Moments and tail shape

Recall that the drift criterion considered in Part I provided sufficient conditions for finite moments of a particular order of the stationary solution to a Markov chain. Likewise, in terms of SREs we have the following result containing easy-to-verify conditions ensuring finite moments of the distribution of X_t.

Theorem V.4.2 (BDM, Lemmas 2.3.1-2.3.2) *Suppose that for $d = 1$, the SRE in (V.1) has a causal stationary solution. Moreover, assume that $\mathbb{P}(A_t x + B_t = x) < 1$ for all $x \in \mathbb{R}$.*

1. *If $\mathbb{P}(A_t = 0) = \mathbb{P}(|A_t| = 1) = 0$ and $\mathbb{P}(B_t = 0) < 1$, then for any $s > 0$,*
$$\mathbb{E}[e^{s|X_t|}] < \infty \textit{ if and only if } \mathbb{P}(|A_t| < 1) = 1 \textit{ and } \mathbb{E}[e^{s|B_t|}] < \infty.$$

2. *If $\mathbb{P}(A_t = 0) = 0$ and $\mathbb{P}(B_t = 0) < 1$, then*
$$\mathbb{E}[|X_t|^p] < \infty \textit{ if and only if } \mathbb{E}[|A_t|^p] < 1 \textit{ and } \mathbb{E}[|B_t|^p] < \infty.$$

Theorem V.4.2.2 provides conditions that are necessary and sufficient for X_t having a finite absolute moment of a given order. In contrast, the drift criterion in Part I only gave us sufficient conditions. For instance, in terms of and ARCH(1) process, in Example I.4.10, it was shown that $\mathbb{E}[X_t^2] < \infty$ if $\alpha < 1$ and $\mathbb{E}[X_t^4] < \infty$ if $\alpha < \sqrt{1/3}$. Below we show that these restrictions on α are sharp, in the sense that they are also necessary.

Example V.4.3 (AR(1) ctd.) *Recall that $A_t = \rho$. Let $\rho \neq 0$ with $|\rho| < 1$ (such that the AR(1) process is stationary and ergodic). Clearly for any fixed $x \in \mathbb{R}$, $\mathbb{P}(A_t x + B_t = x) = \mathbb{P}(B_t = x(1 - \rho)) < 1$, since $\mathbb{V}(B_t) > 0$. Likewise, $\mathbb{P}(B_t = 0) < 1$. We conclude from Theorem V.4.2.1 that for any $s > 0$, $E[e^{s|X_t|}] < \infty$ if and only if $\mathbb{E}[e^{s|B_t|}] < \infty$. Likewise, from Theorem V.4.2.2, $\mathbb{E}[|X_t|^p] < \infty$ if and only if $\mathbb{E}[|B_t|^p] < \infty$.*

Example V.4.4 (ARCH(1) ctd.) *We focus on the stationary region $0 < \alpha < 3.56\ldots$. Since (A_t, B_t) are jointly normal, for any fixed $x \in \mathbb{R}$, $A_t x + B_t \overset{D}{=} N(0, \alpha x^2 + \omega)$ with $\omega > 0$. Consequently, $\mathbb{P}(A_t x + B_t = x) < 1$ for all $x \in \mathbb{R}$. Since $A_t \overset{D}{=} N(0, \alpha)$ with $\alpha > 0$, $\mathbb{P}(|A_t| < 1) < 1$. Hence, from Theorem V.4.2.1 there exists no $s > 0$ such that $\mathbb{E}[e^{s|X_t|}] < \infty$. On the other hand, by Gaussianity $\mathbb{P}(A_t = 0) = \mathbb{P}(B_t = 0) = 0$ and $\mathbb{E}[|B_t|^p] < \infty$ for any finite p. Consequently, by Theorem V.4.2.2, $\mathbb{E}[|X_t|^p] < \infty$ if and only if $\mathbb{E}[|A_t|^p] < 1$. For example, $\mathbb{E}[X_t^2] < \infty$ if and only if $\alpha < 1$, and $\mathbb{E}[X_t^4] < \infty$ if and only if $\alpha < \sqrt{1/3}$.*

The second part of Theorem V.4.2 suggests that X_t may be heavy-tailed in the sense that X_t may have infinite moments. We have the following result about the tail-shape of X_t, typically referred to as the Kesten-Goldie Theorem and refer to BDM (Section 2.4) for additional details (including the precise definitions of the constants) and arguments. In the following we use the notation that for positive functions f and g, $f(x) \sim g(x)$ as $x \to \infty$ if $\lim_{x\to\infty} f(x)/g(x) = 1$.

Theorem V.4.3 (BDM, Theorems 2.4.4 and 2.4.7) *Suppose that for $d = 1$, the SRE in (V.1) has a causal stationary solution. Moreover, assume that $\mathbb{P}(A_t x + B_t = x) < 1$ for all $x \in \mathbb{R}$.*

1. *If $\mathbb{P}(A_t \geq 0) = 1$ and the distribution of $\log(A_t)$ conditional on $\{A_t > 0\}$ is non-arithmetic*[4]*, and there exists a $\kappa > 0$ such that $\mathbb{E}[A_t^\kappa] = 1$, $\mathbb{E}[|B_t|^\kappa] < \infty$ and $\mathbb{E}[A_t^\kappa \log_+(A_t)] < \infty$, then there exist constants $c_1, c_2 \geq 0$ such that $c_1 + c_2 > 0$ and*

$$\mathbb{P}(X_t > x) \sim c_1 x^{-\kappa} \text{ and } \mathbb{P}(X_t < -x) \sim c_2 x^{-\kappa} \quad \text{as } x \to \infty. \tag{V.4}$$

2. *If $\mathbb{P}(A_t < 0) > 0$ and the distribution of $\log|A_t|$ conditional on $\{A_t \neq 0\}$ is non-arithmetic, and there exists a $\kappa > 0$ such that $\mathbb{E}[|A_t|^\kappa] = 1$, $\mathbb{E}[|B_t|^\kappa] < \infty$ and $\mathbb{E}[|A_t|^\kappa \log_+ |A_t|] < \infty$, then there exists a constant $c_3 > 0$ such that*

$$\mathbb{P}(X_t > x) \sim c_3 x^{-\kappa} \text{ and } \mathbb{P}(X_t < -x) \sim c_3 x^{-\kappa} \quad \text{as } x \to \infty. \tag{V.5}$$

Theorem V.4.3 describes the tail-shape of X_t. The result may be useful for quantifying the probabilities of extreme losses in actuarial sciences or risk management. The properties (V.4)-(V.5) mean that the distribution of X_t has power law, or so-called *regularly varying*, tails with $\kappa > 0$ denoting the *tail index*. Note that (V.4) or (V.5) imply that $\mathbb{E}[|X_t|^s] < \infty$ for $s < \kappa$, and $\mathbb{E}[|X_t|^s] = \infty$ for $s \geq \kappa$. This is important in terms of econometric theory, as finite moments are needed in order to apply the LLN (Theorem V.2.2) and CLT (V.2.3) when deriving limit results for estimators.

Example V.4.5 (AR(1) ctd.) *For the stationary case, $|\rho| < 1$, we note that there exists no $\kappa > 0$ such that $\mathbb{E}[A_t^\kappa] = |\rho|^\kappa = 1$. Hence, the results in Theorem V.4.3 do not apply to AR(1) processes. In fact, it can be shown that X_t can only have regularly varying tails if B_t has regularly varying tails.*

Example V.4.6 (ARCH(1) ctd.) *We focus again on the stationary region $0 < \alpha < 3.56\ldots$. We previously argued that $\mathbb{P}(A_t x + B_t = x) < 1$ for all $x \in \mathbb{R}$. Since A_t is Gaussian, it holds that $\mathbb{P}(A_t < 0) > 0$, and we hence seek to apply the second part of Theorem V.4.3. Clearly, since $|A_t|$ has a Lebesgue density, the distribution of $\log|A_t|$ conditional on $\{A_t \neq 0\}$ is non-arithmetic. Let $A_t = \sqrt{\alpha} z_t$ with $z_t \overset{D}{=} N(0,1)$, and note that for any*

[4] A distribution is said to be non-arithmetic if its support is not given by a set of the type $a\mathbb{Z}$, for some $a \geq 0$. Note that this condition is mild and holds, for instance, if A_t has a Lebesgue density.

$\kappa > 0$, $\mathbb{E}[|B_t|^\kappa] < \infty$ and $\mathbb{E}[|A_t|^\kappa \log_+ |A_t|] \leq \mathbb{E}[|A_t|^{1+\kappa}] < \infty$. Moreover, $\mathbb{E}[|A_t|^\kappa] = 1$ implies that $\alpha = (1/\mathbb{E}[|z_t|^\kappa])^{2/\kappa}$. Hence for such combinations of α and κ, X_t has regularly varying tails, that is, (V.5) holds. For instance, we have the pairs

$$(\kappa, \alpha) \in \{(1, \sqrt{\pi/2}), (2, 1), (4, 1/3^{1/2}), (6, 1/15^{1/3}), (8, 1/105^{1/4})\}.$$

In particular, the tails become thicker – that is, κ gets smaller – as α increases.

V.5 Properties of multivariate SREs

This section extends the previous results to the multivariate case. Inherently, the multivariate nature of the process makes the conditions more technical. Throughout, for a column vector $x \in \mathbb{R}^d$, let $|x|$ denote any norm. Moreover for any $d \times d$ matrix A, we consider the matrix norm induced by $|\cdot|$, $\|A\| = \sup_{x \in \mathbb{R}^d, |x|=1} |Ax|$ (the sup-norm).[5]

V.5.1 Stationarity and ergodicity

Recall from that for the univariate case in the previous section, we had (essentially) that the condition $\mathbb{E}[\log |A_t|] < 0$ implied stationarity of the SRE. In the multivariate setting ($d > 1$), the condition is different, reflecting that the matrix product $\prod_{i=1}^t A_i$ should vanish in order to ensure asymptotic negligibility. Here the relevant quantity of interest is the so-called top Lyapunov exponent associated with (V.1) as given by

$$\gamma = \inf_{t \geq 1} \frac{1}{t} \mathbb{E}[\log \| \prod_{i=1}^t A_i \|]. \tag{V.6}$$

We have the following result.

Theorem V.5.1 (BDM, Theorem 4.1.4) *Consider the i.i.d. process $\{(A_t, B_t)\}_{t \in \mathbb{Z}}$ with A_t a $d \times d$ real matrix and B_t $\mathbb{R}^d$-valued. Suppose that one of the following conditions hold:*

1. *$\mathbb{P}(A_t = 0) > 0$.*
2. *$\mathbb{E}[\log^+ \|A_t\|] < \infty$, $\mathbb{E}[\log^+ |B_t|] < \infty$ and the top Lyapunov exponent γ in (V.6) is strictly negative.*

[5] As an example, let $|x|$ denote the ℓ_2-norm, that is, $|x| = \sqrt{x'x}$, such that the induced matrix norm is the spectral norm, that is, the square-root of the largest eigenvalue of $A'A$.

Then there exists a stationary and ergodic causal solution to the SRE in (V.1). The solution is given by (V.3), and this infinite series converges almost surely.

Parallel to the univariate case, we may say that $\gamma < 0$ ensures stationarity of the SRE. Note that the matrix sup-norm is multiplicative, such that $\| \prod_{i=1}^{t} A_i \| \leq \prod_{i=1}^{t} \|A_i\|$. Hence we have that,

$$\frac{1}{t}\mathbb{E}[\log \| \prod_{i=1}^{t} A_i \|] \leq \frac{1}{t} \sum_{i=1}^{t} \mathbb{E}[\log \|A_i\|] = \mathbb{E}[\log \|A_t\|],$$

such that $\gamma < 0$ is implied by $\mathbb{E}[\log \|A_t\|] < 0$, which again (by Jensen's inequality) is implied by $\mathbb{E}[\|A_t\|] < 1$.

Example V.5.1 (VAR(1) ctd.) *Recall that $A_t = A$ is a constant matrix. If A is a zero matrix, then $X_t = B_t$ which clearly yields a stationary and ergodic process. We turn our attention to the non-trivial case where A is non-zero, and seek to apply the second part of Theorem V.5.1. Clearly $\mathbb{E}[\log^+ \|A_t\|] = \log^+ \|A\| < \infty$ and $\mathbb{E}[\log^+ |B_t|] \leq E[|B_t|] < \infty$, since B_t has a finite covariance matrix. It remains to find a condition such that the top Lyapunov exponent is strictly negative. Let $\rho(A)$ denote the spectral radius of A, and recall from Lemma A.3 in Part I that $\lim_{t\to\infty} \|A^t\|^{\frac{1}{t}} = \rho(A)$. We then have that $\gamma = \inf_{t\geq 1} \frac{1}{t}\mathbb{E}[\log \| \prod_{i=1}^{t} A_i \|] = \inf_{t\geq 1} \log \|A^t\|^{\frac{1}{t}} \leq \log \rho\,(A)$. Consequently $\gamma < 0$, if $\rho(A) < 1$. This condition is identical to the one derived in Section I.5, and we have the solution*

$$X_t = \sum_{i=0}^{\infty} A^i \varepsilon_{t-i}.$$

Furthermore, considering the VAR(k) process in Section I.5.1, identical arguments yield the stationarity condition $\rho(A) < 1$, with A denoting the companion matrix of the VAR(k) process.

Example V.5.2 (BEKK ctd.) *Recall that $A_t = m_t A$ with $m_t \overset{D}{=} N(0,1)$ and A a $(d \times d)$ matrix. Assume that A is non-zero. We then have that $P(A_t = 0) = 0$, and seek to apply the second part of Theorem V.5.1. Using that A_t and B_t have Gaussian entries, $\mathbb{E}[\log^+ \|A_t\|] < \infty$ and $\mathbb{E}[\log^+ |B_t|] \leq E[|B_t|] < \infty$, and it remains to find conditions ensuring that $\gamma < 0$. We have that $\mathbb{E}[\log \| \prod_{i=1}^{t} A_i \|] = \mathbb{E}[\log \|A^t \prod_{i=1}^{t} m_i\|] = \log \|A^t\| + t\mathbb{E}[\log(|m_t|)]$. Hence, $\gamma = \inf_{t\geq 1} \log \|A^t\|^{\frac{1}{t}} + \mathbb{E}[\log(|m_t|)] \leq \log \rho(A) + \mathbb{E}[\log(|m_t|)]$. We conclude that $\gamma < 0$ if $\rho(A) < \exp\left(-\mathbb{E}[\log(|m_t|)]\right) = \sqrt{3.56\ldots}$, where we recall the number $3.56\ldots$ from the stationarity condition in the univariate case in Example V.3.3.*

V.5.2 Moments and tail shape

Similar to the univariate case in Section V.4.2, we may consider conditions for finite moments of X_t as well as its tail shape. Inherently, this is a complicated task. In particular, the tail probabilities of a vector may be defined in multiple ways. For instance, each entry of X_t could have different tail index. This is for instance the case if A_t is a diagonal matrix such that X_t consists of d univariate SREs that may have different tail indexes. In such a case $\mathbb{E}[\|X_t\|^s] < \infty$ for any $s > 0$ smaller than the minimum of the tail indices. One way of defining tail probabilities of a vector is to consider the tail probability of linear combinations of X_t, and we refer BDM (Chapter 4) for many more details about so-called multivariate regular variation for multivariate SREs. For details about tail probabilities and finite moments of BEKK-ARCH processes, we refer to Matsui and Pedersen (2022). In terms of showing that $\|X_t\|$ has a a fininte moment of a given order, one may rely on the results in Tweedie (1988) which closely resembles the drift criterion but without requiring existence of a positive, continuous transition density.

V.6 Concluding remarks

The assumption that $\{(A_t, B_t)\}_{t\in\mathbb{Z}}$ being an i.i.d. process can be relaxed. In particular, in a univariate setting, Brandt (1986) considered conditions for stationarity and ergodicity for the case where $\{(A_t, B_t)\}_{t\in\mathbb{Z}}$ itself is stationary and ergodic. A multivariate extension of Brandt's result can be found in Bougerol and Picard (1992). The tail behavior of multivariate X_t and associated limit theory for partial sums of $\{X_t\}_{t\in\mathbb{Z}}$ is an active area of research. We refer to the recent textbook by Mikosch and Wintenberger (2024) for a comprehensive overview of results and examples.

References

Bougerol, P., & Picard, N., 1992, "Strict stationarity of generalized autoregressive processes", *The Annals of Statistics*, Vol. 20, pp. 1714–1730.

Brandt, A., 1986, "The stochastic equation $Y_{n+1} = A_n Y_n + B_n$ with stationary coefficients", *Advances in Applied Probability*, Vol. 18, pp. 211–220.

Buraczewski, D., Damek, E., & Mikosch, T., 2016, *Stochastic Models with Power-Law Tails: The Equation $X = AX + B$*, Springer.

Engle, R.F., 1982, "Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation", *Econometrica*, Vol. 50, pp. 987–1008.

Engle, R.F., & Kroner, K.F., 1995, "Multivariate simultaneous generalized ARCH", *Econometric Theory*, Vol. 11, pp. 122–150.

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

Hansen, A.L., 2021, "Modeling persistent interest rates with double-autoregressive processes", *Journal of Banking and Finance*, Vol. 133, 106302.

Ling, S., 2004, "Estimation and testing stationarity for double-autoregressive models", *Journal of the Royal Statistical Society: Series B*, Vol. 66, pp. 63–78.

Matsui, M., & Pedersen, R.S., 2022, "Characterization of the tail behavior of a class of BEKK processes: A stochastic recurrence equation approach", *Econometric Theory*, Vol. 38, pp. 1–34.

Mikosch, T., & Wintenberger, O., 2024, *Extreme Value Theory for Time Series: Models with Power-Law Tails*, Springer.

Nielsen, H.B., & Rahbek, A., 2014, "Unit root vector autoregression with volatility induced stationarity", *Journal of Empirical Finance*, Vol. 29, pp. 144–167.

Pedersen, R.S., & Rahbek, A., 2016, "Nonstationary GARCH with t-distributed innovations", *Economics Letters*, Vol. 138, pp. 19–21.

Tweedie, R.L., 1988, "Invariant measures for Markov chains with no irreducibility assumptions", *Journal of Applied Probability*, Vol. 25, pp. 275–285.

Anders Rahbek
Rasmus Søndergaard Pedersen
University of Copenhagen

October 2024

Part VI
Statistical inference for time series models

In Part II we considered the properties of the maximum likelihood (ML) estimators for AR and VAR models. A neat feature of these estimators is that they can be written in closed-form as a solution to the problem of maximizing the log-likelihood function. Given this closed form, limit theorems, such as a LLN and CLT, can be applied directly to show consistency and asymptotic normality of the estimators. In this chapter we consider a much more general setting, where an estimator is a maximizer of a criterion function. This class of estimators is labelled *extremum estimators*, and covers a wide range of estimators such as least squares and ML. We state conditions ensuring consistency and asymptotic normality of such estimators. The estimators may not be given in closed-form, as is the case for the ML estimators for ARCH models, and the properties of the estimator are derived from the properties of the criterion function and its derivatives. We consider the AR(1) and ARCH(1) models as running examples and prove that the estimators are consistent and asymptotically normal under the assumption that the data-generating process (DGP) is stationary and ergodic with certain moments finite. To do so, we rely on the limit theorems presented in Part V. *We emphasize that, alternatively, one may consider geometrically ergodic DGPs and apply the limit theory presented in Part I.* Lastly we discuss how the results can be used for hypothesis testing and present potential extensions. The general results for extremum estimators presented are based on Newey and McFadden (1994) [NM henceforth]. We emphasize that several of the assumptions and arguments can be relaxed and refined, and we refer to Amemiya (1985) for many more details as well as Francq and Zakoïan (2019) with particular emphasis on estimation of conditional heteroskedasticity models such as the ARCH(1).

VI.1 Extremum estimators

Suppose that we seek to estimate a vector of parameters in an econometric model. We assume that this parameter vector, θ, is of finite dimension and belongs to a parameter space $\Theta \subseteq \mathbb{R}^k$ for some $k \geq 1$. Given a data set $\{X_t\}_{t=1,\dots,T}$ of length $T \geq 1$, we consider an objective function $Q_T(\theta) := f(\{X_t\}_{t=1}^T;\theta)$, with f measurable. The extremum estimator of θ, labelled $\hat{\theta}_T$, is defined as any solution to

$$\hat{\theta}_T = \arg\max_{\theta\in\Theta} Q_T(\theta). \tag{VI.1}$$

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

$$x_t = \rho x_{t-1} + \varepsilon_t, \quad t \in \mathbb{Z},$$

with $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ is an i.i.d.$N(0,\sigma^2)$ process. As considered in Part II, for a set of observations $\{x_t\}_{t=0}^T$, the log-likelihood function (conditional on x_0) is given by

$$\frac{1}{T}\sum_{t=1}^{T}\log\left[\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x_t-\rho x_{t-1})^2}{2\sigma^2}\right)\right].$$

Treating σ^2 as known (fixed), the single parameter of interest is $\rho =: \theta$, and the parameter space is given by $\Theta = \mathbb{R}$. In this case, the objective function is given by

$$Q_T(\theta) = \frac{1}{T}\sum_{t=1}^{T} q_t(\theta), \quad q_t(\theta) = -\frac{1}{2}(x_t - \theta x_{t-1})^2. \tag{VI.2}$$

Recall also from Part II that the ML estimator is given by,

$$\hat{\theta}_T = \frac{T^{-1}\sum_{t=1}^T x_{t-1}x_t}{T^{-1}\sum_{t=1}^T x_{t-1}^2}. \tag{VI.3}$$

Although the estimator is given in closed form, for illustrative purposes, we consider the estimation of AR(1) models as an ongoing example and consider the stochastic properties of the objective function $Q_T(\theta)$ and its derivatives later on.

Example VI.1.2 (ARCH(1)) *Consider the ARCH(1) model given by*

$$\begin{aligned} x_t &= \sigma_t z_t, \quad t \in \mathbb{Z} \\ \sigma_t^2 &= \omega + \alpha x_{t-1}^2, \end{aligned}$$

with $\{z_t\}_{t\in\mathbb{Z}}$ an i.i.d. process with $z_t \overset{D}{=} N(0,1)$. Moreover, the parameter vector is given by $\theta = (\omega, \alpha)'$ with $\omega > 0$ and $\alpha \geq 0$. Recall from Part I that x_t has conditional density

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

Consequently, for a sample $\{x_t\}_{t=0}^T$, the log-likelihood function (conditional on x_0) is given by

$$Q_T(\theta) = \sum_{t=1}^T q_t(\theta), \quad q_t(\theta) = \log\left[\frac{1}{\sqrt{2\pi\sigma_t^2(\theta)}} \exp\left(-\frac{x_t^2}{2\sigma_t^2(\theta)}\right)\right], \tag{VI.4}$$

with $\sigma_t^2(\theta) = \omega + \alpha x_{t-1}^2$. The parameter space $\Theta \subseteq (0,\infty) \times [0,\infty)$ is considered in more detail later.

VI.2 Consistency

In this section, we consider conditions ensuring that the extremum estimator in (VI.1) has a non-stochastic probability limit, θ_0, as the sample size T diverges. Suppose that $Q_T(\theta)$ has a non-stochastic probability limit $Q(\theta)$ as $T \to \infty$, and suppose that $Q(\theta)$ is uniquely maximized at some point θ_0. Then we say that θ_0 is the true value of θ. This value, inherently, depends on (or is determined from) the underlying DGP. Likewise, $Q(\theta)$ depends on the DGP, as discussed in more detail in the next examples. As mentioned, throughout we consider only DGPs that are stationary and ergodic, which for notational convenience we label $\{x_t\}_{t\in\mathbb{Z}}$ (and not $\{x_t^*\}_{t\in\mathbb{Z}}$).

Example VI.2.1 (AR(1) ctd.) *Consider the objective function in (VI.2) with $\Theta = \mathbb{R}$. Suppose that the DGP is given in terms of $\theta_0 = \rho_0$ with $|\rho_0| < 1$ and that $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ is an i.i.d. process with $\mathbb{E}[\varepsilon_t] = 0$ and $\mathbb{V}[\varepsilon_t] = \sigma_0^2 < \infty$. From Example V.4.1, we know that the DGP is stationary and ergodic, and we have that $\mathbb{E}[x_t] = 0$, $\mathbb{V}[x_t] = \sigma_0^2/(1-\theta_0^2)$ and $\mathbb{E}[x_t x_{t-1}] = \theta_0\sigma_0^2/(1-\theta_0^2)$ for all t. Note that for any $\theta \in \Theta$,*

$$\begin{aligned}
\mathbb{E}[|q_t(\theta)|] &= \frac{1}{2}\mathbb{E}[|x_t^2 + \theta^2 x_{t-1}^2 - 2\theta x_t x_{t-1}|] \\
&\leq \frac{1}{2}\left(\mathbb{E}[x_t^2] + \theta^2\mathbb{E}[x_{t-1}^2] + 2|\theta|\mathbb{E}[|x_t x_{t-1}|]\right) \quad \textit{(triangle)} \\
&\leq \frac{1}{2}\left(\mathbb{E}[x_t^2] + \theta^2\mathbb{E}[x_{t-1}^2] + 2|\theta|\sqrt{\mathbb{E}[x_t^2]\mathbb{E}[x_{t-1}^2]}\right) \quad \textit{(Cauchy-Schwarz)} \\
&= \frac{1}{2}\left(1 + \theta^2 + 2|\theta|\right)\mathbb{E}[x_t^2] < \infty,
\end{aligned}$$

where the last equality follows by stationarity. Moreover,

$$\begin{aligned}\mathbb{E}\left[q_t(\theta)\right] &= -\frac{1}{2}\mathbb{E}[x_t^2 + \theta^2 x_{t-1}^2 - 2\theta x_t x_{t-1}] \\ &= -\frac{1}{2}\left(\mathbb{E}[x_t^2] + \theta^2\mathbb{E}[x_{t-1}^2] - 2\theta\mathbb{E}[x_t x_{t-1}]\right) \\ &= -\frac{1}{2}\left(\frac{\sigma_0^2}{1-\theta_0^2}\right)\left(1 + \theta^2 - 2\theta_0\theta\right) =: Q(\theta)\end{aligned}$$

which is maximized at $\theta = \theta_0$.
Using Theorem V.2.1, noting for a fixed θ, $q_t(\theta)$ is a function of the stationary and ergodic $\{x_t\}_{t\in\mathbb{Z}}$ with $\mathbb{E}[|q_t(\theta)|] < \infty$, we have that $\{q_t(\theta)\}_{t\in\mathbb{Z}}$ itself is stationary and ergodic. Lastly, by Theorem V.2.2, as $T \to \infty$,

$$Q_T(\theta) = \frac{1}{T}\sum_{t=1}^{T} q_t(\theta) \xrightarrow{p} \mathbb{E}\left[q_t(\theta)\right] = Q(\theta).$$

Example VI.2.2 (ARCH(1) ctd.) *Consider the log-likelihood function $Q_T(\theta)$ in (VI.4) with $\Theta \subseteq (0,\infty) \times [0,\infty)$ and*

$$\begin{aligned}q_t(\theta) &= -\frac{1}{2}\left[\log(2\pi) + \log(\sigma_t^2(\theta)) + \frac{x_t^2}{\sigma_t^2(\theta)}\right] \\ &= -\frac{1}{2}\left[\log(2\pi) + \log(\omega + \alpha x_{t-1}^2) + \frac{x_t^2}{\omega + \alpha x_{t-1}^2}\right]. \qquad \text{(VI.5)}\end{aligned}$$

Suppose that the DGP $\{x_t\}_{t\in\mathbb{Z}}$ has $\theta_0 = (\omega_0, \alpha_0)'$ with $\alpha_0 < 1$. Then from Examples V.4.2 and V.4.4, the DGP is stationary and ergodic with $\mathbb{E}[x_t^2] < \infty$. For any fixed $\theta \in \Theta$,

$$\mathbb{E}[|q_t(\theta)|] = \frac{1}{2}\mathbb{E}\left[\left|\log(2\pi) + \log(\omega + \alpha x_{t-1}^2) + \frac{x_t^2}{\omega + \alpha x_{t-1}^2}\right|\right].$$

Using that $\omega > 0$, $\alpha x_{t-1}^2 \geq 0$ and $\log(x) \leq x - 1$ for $x > 0$, we have that

$$\log(\omega) \leq \log\left(\omega + \alpha x_{t-1}^2\right) \leq \omega + \alpha x_{t-1}^2,$$

such that

$$|\log\left(\omega + \alpha x_{t-1}^2\right)| \leq |\log(\omega)| + \left(\omega + \alpha x_{t-1}^2\right).$$

Hence,

$$\begin{aligned}\mathbb{E}\left[|\log\left(\omega + \alpha x_{t-1}^2\right)|\right] &\leq |\log(\omega)| + \mathbb{E}[\omega + \alpha x_{t-1}^2] \\ &= |\log(\omega)| + \omega + \alpha\mathbb{E}[x_{t-1}^2] < \infty.\end{aligned}$$

Likewise, using that $\alpha x_{t-1}^2 \geq 0$

$$0 \leq \mathbb{E}\left[\frac{x_t^2}{\omega + \alpha x_{t-1}^2}\right] \leq \mathbb{E}\left[\frac{x_t^2}{\omega}\right] = \omega^{-1}\mathbb{E}[x_t^2] < \infty.$$

We then have, using the triangle inequality,

$$\mathbb{E}[|q_t(\theta)|] \leq \frac{1}{2}\left[\log(2\pi) + \mathbb{E}\left[|\log\left(\omega + \alpha x_{t-1}^2\right)|\right] + \mathbb{E}\left[\frac{x_t^2}{\omega + \alpha x_{t-1}^2}\right]\right] < \infty.$$

Using Theorem V.2.1, noting for any fixed θ*,* $q_t(\theta)$ *is a function of the stationary and ergodic* $\{x_t\}_{t\in\mathbb{Z}}$ *with* $\mathbb{E}[|q_t(\theta)|] < \infty$*, we have that* $\{q_t(\theta)\}_{t\in\mathbb{Z}}$ *itself is stationary and ergodic. Moreover, Theorem V.2.2 implies that*

$$Q_T(\theta) = \frac{1}{T}\sum_{t=1}^{T} q_t(\theta) \xrightarrow{p} \mathbb{E}\left[q_t(\theta)\right] =: Q(\theta).$$

Lastly, we seek to show that θ_0 *is the unique maximizer of* $Q(\theta)$*. To do so, note that*

$$\mathbb{E}\left[\frac{x_t^2}{\omega_0 + \alpha_0 x_{t-1}^2}\right] = \mathbb{E}\left[\frac{z_t^2(\omega_0 + \alpha_0 x_{t-1}^2)}{\omega_0 + \alpha_0 x_{t-1}^2}\right] = \mathbb{E}[z_t^2] = 1,$$

and

$$\mathbb{E}\left[\frac{x_t^2}{\omega + \alpha x_{t-1}^2}\right] = \mathbb{E}\left[\frac{z_t^2(\omega_0 + \alpha_0 x_{t-1}^2)}{\omega + \alpha x_{t-1}^2}\right] = \mathbb{E}\left[z_t^2\right]\mathbb{E}\left[\frac{\omega_0 + \alpha_0 x_{t-1}^2}{\omega + \alpha x_{t-1}^2}\right] = \mathbb{E}\left[\frac{\omega_0 + \alpha_0 x_{t-1}^2}{\omega + \alpha x_{t-1}^2}\right].$$

Then

$$\begin{aligned} Q(\theta_0) - Q(\theta) &= \mathbb{E}\left[q_t(\theta_0)\right] - \mathbb{E}\left[q_t(\theta)\right] \\ &= -\frac{1}{2}\mathbb{E}\left[\log\left(\omega_0 + \alpha_0 x_{t-1}^2\right) + 1 - \log\left(\omega + \alpha x_{t-1}^2\right) - \frac{\omega_0 + \alpha_0 x_{t-1}^2}{\omega + \alpha x_{t-1}^2}\right] \\ &= \frac{1}{2}\mathbb{E}\left[\log\left(\frac{\omega + \alpha x_{t-1}^2}{\omega_0 + \alpha_0 x_{t-1}^2}\right) + \frac{\omega_0 + \alpha_0 x_{t-1}^2}{\omega + \alpha x_{t-1}^2} - 1\right]. \end{aligned}$$

Note again that for $x > 0$, $\log(x) \leq x - 1$*, or equivalently,* $0 \leq \log(x^{-1}) + x - 1$*, with equality if and only if* $x = 1$*. We then have that* $Q(\theta_0) - Q(\theta) \geq 0$ *with equality if and only if* $\omega_0 + \alpha_0 x_{t-1}^2 = \omega + \alpha x_{t-1}^2$ *almost surely, or equivalently,* $(\alpha_0 - \alpha)x_{t-1}^2 = \omega - \omega_0$ *almost surely. This can only hold if* $\theta = \theta_0$ *or if* x_{t-1}^2 *has a degenerate distribution. The latter is ruled out by the assumption that* z_{t-1} *is Gaussian. Consequently,* $Q(\theta_0) - Q(\theta) \geq 0$ *with equality if and only if* $\theta = \theta_0$*. We conclude that* θ_0 *is the unique maximizer of* $Q(\theta)$*.*

Given that $\hat{\theta}_T$ maximizes $Q_T(\theta)$, that $Q_T(\theta)$ has limit $Q(\theta)$, and that θ_0 maximizes $Q(\theta)$, it seems natural that $\hat{\theta}_T$ converges to θ_0 as the sample size diverges. This is indeed correct under additional technical conditions, stated in the following theorem:

Theorem VI.2.1 (NM, Theorem 2.7) *Suppose that*

1. *Θ is a convex set,*
2. *there exists a function $Q(\theta)$ that has unique maximum at θ_0 on Θ,*
3. *θ_0 is an interior point of Θ,*
4. *$Q_T(\theta)$ is concave, and*
5. *$Q_T(\theta) \xrightarrow{p} Q(\theta)$ for all $\theta \in \Theta$ as $T \to \infty$.*

Then $\hat{\theta}_T$, given by (VI.1), exists with probability approaching one, and $\hat{\theta}_T \xrightarrow{p} \theta_0$ as $T \to \infty$.

Proof. See NM (pp. 2133–2134). ■

Example VI.2.3 (AR(1) ctd.) *Clearly $\Theta = \mathbb{R}$ is convex. Moreover, $Q(\theta)$ is uniquely maximized at any θ_0 with $|\theta_0| < 1$ which is an interior point of Θ. We have that*

$$\frac{\partial^2 Q_T(\theta)}{\partial\theta\partial\theta} = -\frac{1}{T}\sum_{t=1}^{T} x_{t-1}^2 \leq 0, \qquad \text{(VI.6)}$$

such that $Q_T(\theta)$ is concave. Lastly, as showed earlier, it holds that $Q_T(\theta) \xrightarrow{p} Q(\theta)$ for all $\theta \in \Theta$. By Theorem VI.2.1, we conclude that $\hat{\theta}_T \xrightarrow{p} \theta_0$. Note that this conclusion was reached in a much more direct way in Part II. In particular, recall that (VI.1) has solution (VI.3). Noting that $x_t = \theta_0 x_{t-1} + \varepsilon_t$, we have that

$$\hat{\theta}_T = \theta_0 + \frac{T^{-1}\sum_{t=1}^{T} x_{t-1}\varepsilon_t}{T^{-1}\sum_{t=1}^{T} x_{t-1}^2}.$$

Applying Theorem V.2.2 to both the numerator and denominator, as $T \to \infty$,

$$\left(T^{-1}\sum_{t=1}^{T} x_{t-1}\varepsilon_t, T^{-1}\sum_{t=1}^{T} x_{t-1}^2\right) \xrightarrow{p} \left(\mathbb{E}[x_{t-1}\varepsilon_t], \mathbb{E}[x_t^2]\right) = \left(0, \frac{\sigma_0^2}{1-\theta_0^2}\right).$$

Consequently, $\hat{\theta}_T \xrightarrow{p} \theta_0 + 0 = \theta_0$.

When considering non-linear models where the parameters might be constrained, one or more of the conditions in Theorem VI.2.1 may be violated. For instance, the objective function may not be concave, or θ_0 may not be an interior point. Instead, it is customary to assume that the parameter space Θ is compact, which under continuity of the objective function, automatically ensures existence of a maximizer. In this case, we have the following result.

Theorem VI.2.2 (NM, Theorem 2.1) *Suppose that*

1. *Θ is compact,*
2. *there exists a function $Q(\theta)$ that has unique maximum at θ_0 on Θ,*
3. *$Q(\theta)$ is continuous in θ, and*
4. *$\sup_{\theta\in\Theta} |Q_T(\theta) - Q(\theta)| \xrightarrow{p} 0$ as $T \to \infty$.*

Then with $\hat{\theta}_T$ given by (VI.1), $\hat{\theta}_T \xrightarrow{p} \theta_0$ as $T \to \infty$.

The first three conditions in Theorem VI.2.2 are typically straightforward to verify. The fourth condition states that the objective function $Q_T(\theta)$ converges uniformly in probability to $Q(\theta)$. Essentially, this condition ensures that the maximizer of $Q_T(\theta)$ should lie close to the the maximizer of $Q(\theta)$ as the sample size increases, as demonstrated in equation (VI.22) in the proof provided in the Appendix. The uniform convergence of the objective function can typically be shown to hold by applying the Uniform Law of Large Numbers (ULLN) for stationary and ergodic processes stated in Theorem A.1 in the Appendix, or – in the case of geometrically processes – Lemma A.2

Example VI.2.4 (AR(1) ctd.) *Consider the compact parameter space $\Theta = [-1, 1]$ and consider the stationary and ergodic setting with the true value $|\theta_0| < 1$. From earlier, we have that*

$$\begin{aligned}
Q(\theta) &= \mathbb{E}[q_t(\theta)] \\
&= \mathbb{E}\left[-\frac{1}{2}(x_t - \theta x_{t-1})^2\right] \\
&= -\frac{1}{2}\left(\frac{\sigma_0^2}{1-\theta_0^2}\right)\left(1 + \theta^2 - 2\theta_0\theta\right),
\end{aligned}$$

which is continuous in θ and uniquely maximized at $\theta = \theta_0$. Clearly, $q_t(\theta)$ is measurable and continuous in θ. Furthermore, for any $\theta \in \Theta = [-1, 1]$,

$$\begin{aligned}
|q_t(\theta)| &\leq \frac{1}{2}|x_t^2 + \theta^2 x_{t-1}^2 - 2\theta x_t x_{t-1}| \\
&\leq \frac{1}{2}\left(x_t^2 + x_{t-1}^2 + 2|x_t x_{t-1}|\right),
\end{aligned}$$

such that

$$\mathbb{E}[\sup_{\theta\in\Theta}|q_t(\theta)|] \leq \frac{1}{2}\mathbb{E}\left[x_t^2 + x_{t-1}^2 + 2|x_t x_{t-1}|\right] < \infty,$$

as $\mathbb{E}[x_t^2] < \infty$*. By Theorem A.1,* $\sup_{\theta\in\Theta}|Q_T(\theta) - Q(\theta)| \xrightarrow{p} 0$ *as* $T \to \infty$*, and consequently, by Theorem VI.2.2, we have that* $\hat{\theta}_T \xrightarrow{p} \theta_0$ *as* $T \to \infty$*.*

Example VI.2.5 (ARCH(1) ctd.) *We seek to find assumptions such that Conditions 1-4 of Theorem VI.2.2 hold. We initially elaborate on the parameter space* Θ *for the estimation of the parameters,* $\theta = (\omega, \alpha) \in (0,\infty)\times[0,\infty)$*. Note that Condition 4 requires that the log-likelihood function converges uniformly in probability on* Θ*, and we seek to show this by applying Theorem A.1. To do so we need that* $\mathbb{E}[\sup_{\theta\in\Theta}|q_t(\theta)|] < \infty$*, with* $q_t(\theta)$ *the log-likelihood contribution given in (VI.5). Note initially that for any* $x, y \in \mathbb{R}$*,*

$$\sup_{\theta\in(0,\infty)\times[0,\infty)}\left|\log(2\pi) + \log(\omega + \alpha x^2) + \frac{x^2}{\omega + \alpha y^2}\right| = \infty.$$

In particular, this suggests that $\omega > 0$ *should be bounded away from zero and bounded from above, whereas* $\alpha \geq 0$ *should be bounded from above. Consequently, we consider the compact parameter space (Condition 1)*

$$\Theta = [\omega_L, \omega_U] \times [0, \alpha_U], \tag{VI.7}$$

with $0 < \omega_L < \omega_U < \infty$ *and* $0 < \alpha_U < \infty$*. Then for any* $x, y \in \mathbb{R}$*, using the same type of bounds as in Example VI.2.2,*

$$\begin{aligned}
&\sup_{\theta\in\Theta}\left|\log(2\pi) + \log(\omega + \alpha y^2) + \frac{x^2}{\omega + \alpha y^2}\right| \\
&\leq \sup_{\theta\in[\omega_L,\omega_U]\times[0,\alpha_U]}\left(\log(2\pi) + |\log(\omega)| + \left(\omega + \alpha y^2\right) + \frac{x^2}{\omega + \alpha y^2}\right) \\
&\leq \log(2\pi) + |\log(\omega_L)| + |\log(\omega_U)| + \left(\omega_U + \alpha_U y^2\right) + \frac{x^2}{\omega_L}.
\end{aligned}$$

Maintaining the assumption that $\alpha_0 < 1$ *such that* $\{x_t\}_{t\in\mathbb{Z}}$ *is stationary and ergodic with* $\mathbb{E}[x_t^2] < \infty$*, we have that*

$$\mathbb{E}[\sup_{\theta\in\Theta}|q_t(\theta)|] \leq \log(2\pi) + |\log(\omega_L)| + |\log(\omega_U)| + \omega_U + \alpha_U\mathbb{E}(x_{t-1}^2) + \frac{\mathbb{E}[x_t^2]}{\omega_L} < \infty.$$

Clearly, $q_t(\theta)$ *is finite almost surely for any* $\theta \in \Theta$ *and continuous in* θ*. Consequently, by Theorem A.1,*

$$\sup_{\theta\in\Theta}|Q_T(\theta) - \mathbb{E}\left[q_t(\theta)\right]| \xrightarrow{p} 0 \quad \text{as } T \to \infty,$$

such that Condition 4 holds. Moreover, by dominated convergence $Q(\theta) = \mathbb{E}\left[q_t(\theta)\right]$ is continuous in θ (Condition 3). Lastly, assuming that $\theta_0 \in \Theta$, from Example VI.2.2, it holds that $Q(\theta)$ is uniquely maximized at θ_0 (Condition 2). We conclude that all Conditions 1-4 of Theorem VI.2.2 hold with Θ given by (VI.7) provided that $\theta_0 \in \Theta$ and $\alpha_0 < 1$. Under these conditions, with $\hat{\theta}_T$ given by (VI.1), $\hat{\theta}_T \xrightarrow{p} \theta_0$ as $T \to \infty$.

Remark VI.2.1 *Example VI.2.5 showed that for a compact parameter space, the MLE $\hat{\theta}_T$ converges to θ_0 provided that the DGP is stationary and ergodic with $\mathbb{E}[x_t^2] < \infty$. The assumption about finite second moment can be relaxed such that $\mathbb{E}[|x_t|^\delta] < \infty$ for some $\delta > 0$; see e.g. Francq and Zakoïan (2019, Chapter 7). One can also relax the assumption about the compactness of the parameter space, as done in Kristensen and Rahbek (2005), who assume that the DGP $\{x_t\}_{t=0,1,\ldots}$ is geometrically ergodic (but not necessarily stationary). Lastly, it is possible to show that the MLE for α is consistent, even if $\mathbb{E}[\log(\alpha_0 z_t^2)] > 0$, such that no stationary solution to the ARCH(1) process exists (Example V.4.2). This relies on rather technical arguments, and we refer to Jensen and Rahbek (2004a) and Francq and Zakoïan (2012) for details.*

VI.3 Asymptotic normality

In this section, we present conditions such that the estimator $\hat{\theta}_T$ is asymptotically normal. Let $\|\cdot\|$ denote any matrix norm; cf. the matrix results provided in the appendix to Part I.

We have the following result.

Theorem VI.3.1 *Suppose that with $\hat{\theta}_T$ given by (VI.1),*

1. *$\hat{\theta}_T \xrightarrow{p} \theta_0$ as $T \to \infty$,*

2. *the true value θ_0 is an interior point of the parameter space Θ,*

3. *$Q_T(\theta_0)$ is twice continuously differentiable in a neighborhood $\mathcal{N}(\theta_0)$ of θ_0 almost surely,*

4. *the score,*

$$\sqrt{T}\frac{\partial Q_T(\theta_0)}{\partial \theta} := \sqrt{T}\left.\frac{\partial Q_T(\theta)}{\partial \theta}\right|_{\theta=\theta_0} \xrightarrow{D} N(0, \Omega_0) \quad \text{as } T \to \infty,$$

with Ω_0 a positive definite $(k \times k)$ matrix, and

5. *there exists a matrix-valued function $\Sigma(\theta)$ that is continuous at θ_0, such that*
$$\sup_{\theta \in \mathcal{N}(\theta_0)} \left\| \frac{\partial^2 Q_T(\theta)}{\partial\theta\partial\theta'} - \Sigma(\theta) \right\| \xrightarrow{p} 0 \quad \text{as } T \to \infty,$$
and $\Sigma_0 := \Sigma(\theta_0)$ invertible.

Then as $T \to \infty$,

$$\sqrt{T}\left(\hat{\theta}_T - \theta_0\right) \xrightarrow{D} N(0, \Sigma_0^{-1}\Omega_0\Sigma_0^{-1}).$$

Proof. Note that Condition 1 implies that $\hat{\theta}_T \in \mathcal{N}(\theta_0)$ with probability approaching one. This combined with Conditions 2 and 3 yield that, with probability approaching one,

$$\frac{\partial Q_T(\hat{\theta}_T)}{\partial\theta} = 0_{k\times 1}.$$

Mean value expansions (element-by-element) of $\partial Q_T(\hat{\theta}_T)/\partial\theta$ at θ_0 yield (with probability approaching one)

$$0_{k\times 1} = \frac{\partial Q_T(\theta_0)}{\partial\theta} + \frac{\partial^2 Q_T(\theta_T^*)}{\partial\theta\partial\theta'}(\hat{\theta}_T - \theta_0), \tag{VI.8}$$

where the mean value θ_T^* lies between[1] $\hat{\theta}_T$ and θ_0. Note that Condition 1 implies that $\theta_T^* \xrightarrow{p} \theta_0$ as $T \to \infty$. This combined with Condition 5 and Lemma A.1 yield that

$$\frac{\partial^2 Q_T(\theta_T^*)}{\partial\theta\partial\theta'} = \Sigma_0 + o_p(1). \tag{VI.9}$$

Combining (VI.8) and (VI.9) yield that

$$0_{k\times 1} = \sqrt{T}\frac{\partial Q_T(\theta_0)}{\partial\theta} + (\Sigma_0 + o_p(1))\sqrt{T}(\hat{\theta}_T - \theta_0).$$

Since Σ_0 is invertible, re-arranging yields

$$\sqrt{T}(\hat{\theta}_T - \theta_0) = -\left(\Sigma_0 + o_p(1)\right)^{-1}\sqrt{T}\frac{\partial Q_T(\theta_0)}{\partial\theta} \xrightarrow{D} N(0, \Sigma_0^{-1}\Omega_0\Sigma_0^{-1}),$$

where the convergence follows by Condition 4 and Slutsky's lemma. ■

[1] Note that θ_T^* may vary across the rows of $\partial^2 Q_T(\theta_T^*)/\partial\theta\partial\theta'$; see, e.g., Jensen and Rahbek (2004b) for the precise details.

Example VI.3.1 (AR(1) ctd.) *Consider Example VI.2.4 with the compact parameter space $\Theta = [-1, 1]$, and recall that $|\theta_0| < 1$. We seek to verify Conditions 1–5 of Theorem VI.3.1. The estimator is consistent as shown in Example VI.2.4 (Condition 1). Since $|\theta_0| < 1$, θ_0 is an interior point of the parameter space (Condition 2). From VI.6,*

$$\frac{\partial^2 Q_T(\theta)}{\partial \theta^2} = -\frac{1}{T}\sum_{t=1}^{T} x_{t-1}^2,$$

which is clearly continuous in θ (Condition 3), and clearly by Theorem V.2.2,

$$\frac{\partial^2 Q_T(\theta)}{\partial \theta^2} \xrightarrow{p} -\mathbb{E}[x_t^2] =: \Sigma_0 < 0$$

as $T \to \infty$ uniformly in θ (Condition 5). It remains to show Condition 4. We have that

$$\begin{aligned}\frac{\partial Q_T(\theta)}{\partial \theta} &= \frac{1}{T}\sum_{t=1}^{T} \frac{\partial q_t(\theta)}{\partial \theta} \\ &= \frac{1}{T}\sum_{t=1}^{T} \frac{\partial}{\partial \theta}\left(-\frac{1}{2}(x_t - \theta x_{t-1})^2\right) \\ &= \frac{1}{T}\sum_{t=1}^{T}(x_t - \theta x_{t-1})x_{t-1},\end{aligned}$$

such that

$$\sqrt{T}\frac{\partial Q_T(\theta_0)}{\partial \theta} = \frac{1}{\sqrt{T}}\sum_{t=1}^{T} \varepsilon_t x_{t-1}.$$

Clearly $\varepsilon_t x_{t-1}$ forms a martingale difference with $\mathbb{E}[\varepsilon_t^2 x_{t-1}^2] = \mathbb{E}[\varepsilon_t^2]\mathbb{E}[x_{t-1}^2] < \infty$. By Theorem V.2.3 we have that

$$\sqrt{T}\frac{\partial Q_T(\theta_0)}{\partial \theta} \xrightarrow{D} N(0, \Omega_0),$$

with $\Omega_0 = \mathbb{E}[\varepsilon_t^2]\mathbb{E}[x_{t-1}^2]$. By Theorem VI.3.1 we have that

$$\sqrt{T}(\hat{\theta}_T - \theta_0) \xrightarrow{D} N(0, \Sigma_0^{-1}\Omega_0\Sigma_0^{-1}).$$

Noting that $\Sigma_0^{-1}\Omega_0\Sigma_0^{-1} = \mathbb{E}[\varepsilon_t^2]/\mathbb{E}[x_t^2] = (1 - \theta_0^2)$

$$\sqrt{T}(\hat{\theta}_T - \theta_0) \xrightarrow{D} N(0, (1 - \theta_0^2)),$$

identical to the conclusion for the ML estimator in (II.6).

VI.3.1 Asymptotic normality of the ML estimator for the ARCH(1) model

As in Example VI.2.5, consider the parameter space Θ in (VI.7), and assume throughout that the true $\alpha_0 < 1$ such that the DGP is stationary and ergodic with $\mathbb{E}[x_t^2] < \infty$. We seek to show that Conditions 1–5 of Theorem VI.3.1 hold.

Condition 1:

From Example VI.2.5 we have that $\hat{\theta}_T$ is consistent, that is, Condition 1 holds.

Condition 2:

We assume that θ_0 is an interior point of Θ, so that Condition 2 holds.

Condition 3:

Recall that $Q_T(\theta) = T^{-1}\sum_{t=1}^{T} q_t(\theta)$ with $q_t(\theta)$ given by (VI.5). Clearly $q_t(\theta)$, and hence $Q_T(\theta)$, is twice continuously differentiable (almost surely), and we have that Condition 3 holds.

Condition 4:

We have that

$$\begin{aligned}\frac{\partial Q_T(\theta)}{\partial \theta} &= \frac{1}{T}\sum_{t=1}^{T}\frac{\partial q_t(\theta)}{\partial \theta},\\ &= -\frac{1}{2T}\sum_{t=1}^{T}\frac{\partial}{\partial \theta}\left[\log(2\pi) + \log(\omega + \alpha x_{t-1}^2) + \frac{x_t^2}{\omega + \alpha x_{t-1}^2}\right]\end{aligned}$$

with

$$\begin{aligned}\frac{\partial q_t(\theta)}{\partial \theta} &= -\frac{1}{2}\frac{\partial}{\partial \theta}\left[\log(2\pi) + \log(\omega + \alpha x_{t-1}^2) + \frac{x_t^2}{\omega + \alpha x_{t-1}^2}\right]\\ &= -\frac{1}{2}\frac{1}{\omega + \alpha x_{t-1}^2}\left(1 - \frac{x_t^2}{\omega + \alpha x_{t-1}^2}\right) w_t.\end{aligned}$$

with

$$w_t := \begin{bmatrix} 1 \\ x_{t-1}^2 \end{bmatrix}.$$

Use that $x_t^2 = (\omega_0 + \alpha_0 x_{t-1}^2) z_t^2$, such that

$$\frac{\partial q_t(\theta_0)}{\partial \theta} = \frac{1}{2} \frac{1}{\omega_0 + \alpha_0 x_{t-1}^2} \left(z_t^2 - 1\right) w_t.$$

With $\mathcal{F}_t = \sigma(x_t, x_{t-1}, \dots)$, note that (provided that $\mathbb{E}[\|\partial q_t(\theta_0)/\partial\theta\|] < \infty$, which is shown below), almost surely

$$\mathbb{E}\left[\frac{\partial q_t(\theta_0)}{\partial \theta} | \mathcal{F}_{t-1}\right] = 0_{2\times 1}.$$

Hence $\partial q_t(\theta_0)/\partial\theta$ is a martingale difference, and we seek to prove Condition 4 by applying the CLT in Theorem V.2.3. To do so, we show that the CLT applies to any linear combination of $\partial q_t(\theta_0)/\partial\theta$. For any fixed non-zero $\lambda := (\lambda_1, \lambda_2)' \in \mathbb{R}^2$, let

$$\begin{aligned} Y_t^{(\lambda)} &:= \lambda' \frac{\partial q_t(\theta_0)}{\partial \theta} \\ &= \frac{1}{2}\left(z_t^2 - 1\right) \frac{\lambda' w_t}{\omega_0 + \alpha_0 x_{t-1}^2} \\ &= \frac{1}{2}\left(z_t^2 - 1\right) \frac{\lambda_1 + \lambda_2 x_{t-1}^2}{\omega_0 + \alpha_0 x_{t-1}^2}. \end{aligned} \tag{VI.10}$$

Clearly, $\mathbb{E}[Y_t^{(\lambda)} | \mathcal{F}_{t-1}] = 0$ almost surely. Moreover, by independence,

$$\begin{aligned} \mathbb{E}\left[\left(Y_t^{(\lambda)}\right)^2\right] &= \mathbb{E}\left[\left(\frac{1}{2}\left(z_t^2 - 1\right) \frac{\lambda_1 + \lambda_2 x_{t-1}^2}{\omega_0 + \alpha_0 x_{t-1}^2}\right)^2\right] \\ &= \frac{1}{4}\mathbb{E}\left[\left(z_t^2 - 1\right)^2\right] \mathbb{E}\left[\left(\frac{\lambda_1 + \lambda_2 x_{t-1}^2}{\omega_0 + \alpha_0 x_{t-1}^2}\right)^2\right], \\ &= \frac{1}{2}\mathbb{E}\left[\left(\frac{\lambda_1 + \lambda_2 x_{t-1}^2}{\omega_0 + \alpha_0 x_{t-1}^2}\right)^2\right], \end{aligned}$$

where we have used that $\mathbb{E}\left[\left(z_t^2 - 1\right)^2\right] = \mathbb{E}[z_t^4 + 1 - 2z_t^2] = 3 + 1 - 2 = 2$ since $z_t \overset{D}{=} N(0,1)$. Since θ_0 is an interior point of Θ_0, we have that both $\omega_0, \alpha_0 > 0$, so that

$$\left|\frac{\lambda_1}{\omega_0 + \alpha_0 x_{t-1}^2}\right| = \frac{|\lambda_1|}{\omega_0 + \alpha_0 x_{t-1}^2} \leq \frac{|\lambda_1|}{\omega_0} < \infty,$$

and

$$\left|\frac{\lambda_2 x_{t-1}^2}{\omega_0 + \alpha_0 x_{t-1}^2}\right| = \frac{|\lambda_2| x_{t-1}^2}{\omega_0 + \alpha_0 x_{t-1}^2} \leq \frac{|\lambda_2|}{\alpha_0} < \infty.$$

These bounds yield that

$$\mathbb{E}\left[\left(\frac{\lambda_1+\lambda_2 x_{t-1}^2}{\omega_0+\alpha_0 x_{t-1}^2}\right)^2\right] \leq \left(\frac{|\lambda_1|}{\omega_0}\right)^2 + \left(\frac{|\lambda_2|}{\alpha_0}\right)^2 + 2\left(\frac{|\lambda_1|}{\omega_0}\right)\left(\frac{|\lambda_2|}{\alpha_0}\right) < \infty,$$

and we have that

$$\mathbb{E}\left[\left(Y_t^{(\lambda)}\right)^2\right] < \infty.$$

By Theorem V.2.3, we then have that for any non-zero λ,

$$\frac{1}{\sqrt{T}}\lambda'\frac{\partial Q_T(\theta_0)}{\partial\theta} = \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\lambda'\frac{\partial q_t(\theta_0)}{\partial\theta} \xrightarrow{D} N(0,\gamma_\lambda),$$

with $\gamma_\lambda = \mathbb{E}[(Y_t^{(\lambda)})^2] < \infty$ provided that $\gamma_\lambda > 0$. Note that

$$\begin{aligned}\gamma_\lambda &= \mathbb{E}\left[\left(\lambda'\frac{\partial q_t(\theta_0)}{\partial\theta}\right)^2\right]\\ &= \lambda'\mathbb{E}\left[\frac{\partial q_t(\theta_0)}{\partial\theta}\frac{\partial q_t(\theta_0)}{\partial\theta'}\right]\lambda\\ &= \lambda'\Omega_0\lambda,\end{aligned}$$

with

$$\begin{aligned}\Omega_0 &:= \mathbb{E}\left[\frac{\partial q_t(\theta_0)}{\partial\theta}\frac{\partial q_t(\theta_0)}{\partial\theta'}\right]\\ &= \mathbb{E}\left[\frac{1}{4}\frac{1}{(\omega_0+\alpha_0 x_{t-1}^2)^2}\left(z_t^2-1\right)^2 w_t w_t'\right]\\ &= \frac{1}{2}\mathbb{E}\left[\frac{1}{(\omega_0+\alpha_0 x_{t-1}^2)^2}w_t w_t'\right].\end{aligned}$$

Clearly, $\gamma_\lambda > 0$ for all non-zero λ, if and only if Ω_0 is positive definite. We prove the latter by contradiction. Note that $\lambda'\Omega_0\lambda \geq 0$ for all λ. Suppose that Ω_0 is not positive definite. Then there exists a non-zero λ such that

$$\lambda'\Omega_0\lambda = \mathbb{E}\left[\left(\frac{\lambda' w_t}{\omega_0+\alpha_0 x_{t-1}^2}\right)^2\right] = 0.$$

Since $\left(\lambda' w_t/(\omega_0+\alpha_0 x_{t-1}^2)\right)^2$ is non-negative and $\omega_0+\alpha_0 x_{t-1}^2 > 0$ it holds that $\lambda' w_t = 0$ almost surely, that is, $\mathbb{P}\left(\lambda_1+\lambda_2 x_{t-1}^2 = 0\right) = 1$. Since z_{t-1} is Gaussian, we rule out $\mathbb{P}\left(\lambda_1+\lambda_2 x_{t-1}^2 = 0\right) = 1$ unless $(\lambda_1,\lambda_2) = (0,0)$, that is, λ is the zero vector. We conclude that Ω_0 is positive definite, and we have that

$$\frac{1}{\sqrt{T}}\lambda'\frac{\partial Q_T(\theta_0)}{\partial\theta} \xrightarrow{D} N(0, \lambda'\Omega_0\lambda),$$

for all non-zero λ. The Cramér-Wold theorem yields that

$$\frac{1}{\sqrt{T}}\frac{\partial Q_T(\theta_0)}{\partial\theta} \xrightarrow{D} N(0, \Omega_0),$$

and we conclude that Condition 4 holds.

Condition 5:

We set up for an application of the ULLN in Theorem A.1 to each entry of $\partial^2 Q_T(\theta)/\partial\theta\partial\theta'$. Since θ_0 is an interior point of Θ, $\alpha_0 > 0$, and we have that there exists a α_L satisfying $0 < \alpha_L < \alpha_0 < \alpha_U$, such that $\theta_0 \in [\omega_L, \omega_U] \times [\alpha_L, \alpha_U] =: \mathcal{N}(\theta_0)$. We have that

$$\frac{\partial^2 Q_T(\theta)}{\partial\theta\partial\theta'} = \frac{1}{T}\sum_{t=1}^{T}\frac{\partial^2 q_t(\theta)}{\partial\theta\partial\theta'},$$

with

$$\begin{aligned}\frac{\partial^2 q_t(\theta)}{\partial\theta\partial\theta'} &= \begin{bmatrix} \frac{\partial^2}{\partial\omega^2}q_t(\theta) & \frac{\partial^2}{\partial\omega\partial\alpha}q_t(\theta) \\ \frac{\partial^2}{\partial\omega\partial\alpha}q_t(\theta) & \frac{\partial^2}{\partial\alpha^2}q_t(\theta)\end{bmatrix} \\ &= \frac{1}{2}\frac{1}{\sigma_t^4(\theta)}\left(1 - \frac{2x_t^2}{\sigma_t^2(\theta)}\right)w_t w_t'.\end{aligned}$$

It suffices to show that

$$\mathbb{E}\left[\sup_{\theta\in\mathcal{N}(\theta_0)}\left\|\frac{\partial^2 q_t(\theta)}{\partial\theta\partial\theta'}\right\|\right] < \infty. \qquad \text{(VI.11)}$$

Note that for some constant $C > 0$,

$$\begin{aligned}\left\|\frac{\partial^2 q_t(\theta)}{\partial\theta\partial\theta'}\right\| &= \left\|\frac{1}{2}\frac{1}{\sigma_t^4(\theta)}\left(1-\frac{2x_t^2}{\sigma_t^2(\theta)}\right)w_t w_t'\right\| \\ &\leq \left(\frac{1}{2}\frac{1}{\sigma_t^4(\theta)}\left(1+\frac{2x_t^2}{\sigma_t^2(\theta)}\right)\right)\|w_t w_t'\| \\ &\leq C\left(\frac{1}{2}\frac{1}{\sigma_t^4(\theta)}\left(1+\frac{2x_t^2}{\sigma_t^2(\theta)}\right)\right)w_t' w_t \\ &= \frac{C}{2}\left(\frac{1}{\sigma_t^4(\theta)}\left(1+\frac{2x_t^2}{\sigma_t^2(\theta)}\right)\right)(1+x_{t-1}^4) \\ &= \frac{C}{2}\frac{(1+x_{t-1}^4)}{\sigma_t^4(\theta)}+\frac{C}{2}\frac{(1+x_{t-1}^4)2x_t^2}{\sigma_t^6(\theta)} \\ &= \frac{C}{2}\frac{(1+x_{t-1}^4)}{\sigma_t^4(\theta)}+C\frac{z_t^2(1+x_{t-1}^4)(\omega_0+\alpha_0 x_{t-1}^2)}{\sigma_t^6(\theta)},\end{aligned}$$

so that (VI.11) holds provided that

$$\mathbb{E}\left[\sup_{\theta\in\mathcal{N}(\theta_0)}\left|\frac{C}{2}\frac{(1+x_{t-1}^4)}{\sigma_t^4(\theta)}\right|\right]+\mathbb{E}\left[\sup_{\theta\in\mathcal{N}(\theta_0)}\left|C\frac{z_t^2(1+x_{t-1}^4)(\omega_0+\alpha_0 x_{t-1}^2)}{\sigma_t^6(\theta)}\right|\right]<\infty.$$

We restrict our attention to the second term, and note that the finiteness of the first term follows by similar arguments. We have (ignoring the constant $C > 0$)

$$\begin{aligned}&\mathbb{E}\left[\sup_{\theta\in\mathcal{N}(\theta_0)}\left|\frac{z_t^2(1+x_{t-1}^4)(\omega_0+\alpha_0 x_{t-1}^2)}{\sigma_t^6(\theta)}\right|\right] \\ &= \mathbb{E}\left[\sup_{\theta\in\mathcal{N}(\theta_0)}\frac{(1+x_{t-1}^4)(\omega_0+\alpha_0 x_{t-1}^2)}{\sigma_t^6(\theta)}\right] \qquad \text{(independence)} \\ &\leq \mathbb{E}\left[\frac{(1+x_{t-1}^4)(\omega_0+\alpha_0 x_{t-1}^2)}{(\omega_L+\alpha_L x_{t-1}^2)^3}\right] \\ &= \mathbb{E}\left[\frac{(\omega_0+\alpha_0 x_{t-1}^2)}{(\omega_L+\alpha_L x_{t-1}^2)^3}+\frac{x_{t-1}^4(\omega_0+\alpha_0 x_{t-1}^2)}{(\omega_L+\alpha_L x_{t-1}^2)^3}\right] \\ &\leq \frac{\omega_0}{\omega_L^3}+\frac{\alpha_0}{\omega_L^2\alpha_L}+\frac{\omega_0}{\omega_L\alpha_L^2}+\frac{\alpha_0}{\alpha_L^3}<\infty.\end{aligned}$$

We conclude that (VI.11) holds such that, by Theorem A.1,

$$\sup_{\theta\in\mathcal{N}(\theta_0)}\left\|T^{-1}\sum_{t=1}^{T}\frac{\partial^2 q_t(\theta)}{\partial\theta\partial\theta'}-\Sigma(\theta)\right\|\xrightarrow{p}0 \quad \text{as } T\to\infty,$$

with

$$\Sigma(\theta) = \mathbb{E}\left[\frac{\partial^2 q_t(\theta)}{\partial\theta\partial\theta'}\right].$$

Clearly, $\Sigma(\theta)$ is continuous at θ_0, and it remains to show that $\Sigma_0 = \Sigma(\theta_0)$ is invertible. But this is immediate, noting that

$$\begin{aligned}\Sigma_0 &= \mathbb{E}\left[\frac{1}{2}\frac{1}{\sigma_t^4(\theta_0)}\left(1 - \frac{2x_t^2}{\sigma_t^2(\theta_0)}\right)w_t w_t'\right] \\ &= \mathbb{E}\left[\frac{1}{2}\frac{1}{\sigma_t^4(\theta_0)}\left(1 - 2z_t^2\right)w_t w_t'\right] \\ &= -\frac{1}{2}\mathbb{E}\left[\frac{1}{\sigma_t^4(\theta_0)}w_t w_t'\right] \\ &= -\Omega_0.\end{aligned}$$

We conclude that Condition 5 holds.

To sum up:

If $\alpha_0 \in (0,1)$, it holds that the ML estimator $\hat{\theta}_T$ satisfies

$$\sqrt{T}(\hat{\theta}_T - \theta_0) \overset{D}{\to} N(0, -\Sigma_0^{-1}),$$

using that $\Omega_0 = -\Sigma_0$. Note that the latter equality is the so-called information equality that holds for ($\sqrt{T}$-consistent) ML estimators.

VI.4 Hypothesis testing

Similar to Section II.4.4, we may consider testing hypotheses about the parameter vector θ. We restrict our attention to linear hypotheses of the form

$$H : R'\theta = r, \tag{VI.12}$$

with R a known $(k \times l)$ matrix with rank $0 < l \leq k$, and r a known l-dimensional vector.[2]

Example VI.4.1 (ARCH(1) ctd) *Following Example V.4.6 we may test the hypothesis*

$$H : \alpha = \frac{1}{\sqrt{3}}, \tag{VI.13}$$

such that the unconditional distribution of x_t has tail index $\kappa = 4$. In this case, recalling that $\theta = (\omega, \alpha)'$, we have that $R = (0,1)'$ and $r = 1/\sqrt{3}$.

[2] The following considerations may be extended to more general hypotheses of the form $F(\theta) = r$ for certain known functions F; see e.g. Section 9 of NM for additional details.

VI.4.1 Likelihood Ratio (LR) tests

Similar to Part II, we consider a likelihood ratio (LR) test for the hypothesis. Specifically, let $\tilde{\theta}_T$ denote the maximizer of $Q_T(\theta)$ over Θ subject to the constraint (VI.12). Then the LR statistic for the hypothesis in (VI.12) is given by

$$LR_T(H) := 2T[Q_T(\hat{\theta}_T) - Q_T(\tilde{\theta}_T)]. \tag{VI.14}$$

In the special case of the hypothesis $\theta = \theta_0$ (that is, $l = k$, $R = I_k$ and $r = \theta_0$), $\tilde{\theta}_T = \theta_0$. In this case, under the assumptions of Theorem VI.3.1, a second-order mean-value expansion of $Q_T(\theta_0)$ around $\hat{\theta}_T$ yields that

$$Q_T(\theta_0) = Q_T(\hat{\theta}_T) + \frac{\partial Q_T(\hat{\theta}_T)}{\partial \theta'}(\theta_0 - \hat{\theta}_T) + \frac{1}{2}(\hat{\theta}_T - \theta_0)'\hat{\Sigma}_T(\hat{\theta}_T - \theta_0) + o_p(T^{-1}),$$

with

$$\hat{\Sigma}_T := \frac{\partial^2 Q_T(\hat{\theta}_T)}{\partial\theta\partial\theta'}.$$

Hence, using the first-order condition $\partial Q_T(\hat{\theta}_T)/\partial\theta = 0_{k\times 1}$ (with probability approaching one), we have that the LR statistic satisfies

$$LR_T\,(\theta = \theta_0) = 2T[Q_T(\hat{\theta}_T) - Q_T(\theta_0)] = W + o_p(1), \tag{VI.15}$$

where $W := T(\hat{\theta}_T - \theta_0)'(-\hat{\Sigma}_T)(\hat{\theta}_T - \theta_0)$ corresponds to the Wald statistic derived in Part II. By Lemma VI.4.1 provided below, $\hat{\Sigma}_T \xrightarrow{p} \Sigma_0$ as $T \to \infty$. In the case where the information equality holds, that is,

$$\Omega_0 = -\Sigma_0, \tag{VI.16}$$

such that $\sqrt{T}(\hat{\theta} - \theta_0)$ has asymptotic covariance $-\Sigma_0^{-1}$, we then have that

$$LR_T\,(\theta = \theta_0) \xrightarrow{D} \chi_k^2 \quad \text{as } T \to \infty.$$

Under a more general hypothesis of the form (VI.12), we have the following result, where the consistency of $\tilde{\theta}_T$ may be shown to hold under the assumptions of Theorem VI.2.2.

Theorem VI.4.1 *Suppose that the assumptions of Theorem VI.3.1 hold. Moreover, assume that the information equality (VI.16) holds, and that under the hypothesis H in (VI.12), $\tilde{\theta}_T \xrightarrow{p} \theta_0$ as $T \to \infty$. Then under the hypothesis H in (VI.12), the LR statistic in (VI.14) satisfies*

$$LR_T(H) \xrightarrow{D} \chi_l^2, \quad \textit{as } T \to \infty.$$

Proof. See Appendix. ■

VI.4.2 Wald tests

If the information equality does not hold – which is for instance the case for so-called *quasi* ML estimators for ARCH models, as considered later – $LR_T(H)$ does not have a standard chi-squared distribution. In such a case, one may alternatively make use of a (robustified) Wald test, provided that one has a consistent estimator for the asymptotic covariance matrix of $\sqrt{T}(\hat{\theta}_T - \theta_0)$. Specifically, suppose that there exists a matrix $\hat{\Psi}_T$, such that

$$\hat{\Psi}_T \xrightarrow{p} \Sigma_0^{-1}\Omega_0\Sigma_0^{-1} \quad \text{as } T \to \infty. \tag{VI.17}$$

Then a Wald statistic for the hypothesis (VI.12) is given by

$$W_{\text{rob}} = T\left(R'\hat{\theta}_T - r\right)'\left(R'\hat{\Psi}_T R\right)^{-1}\left(R'\hat{\theta}_T - r\right). \tag{VI.18}$$

We have the following result.

Theorem VI.4.2 *Under the assumptions of Theorem VI.3.1, suppose that (VI.17) holds. Then under the hypothesis H in (VI.12), the Wald statistic W_{rob} in (VI.18) satisfies*

$$W_{\text{rob}} \xrightarrow{D} \chi^2(l).$$

Proof. Under the hypothesis H in (VI.12), we have that

$$R'\theta_0 = r,$$

such that by Theorem VI.3.1 and the continuous mapping theorem

$$R'\hat{\theta}_T - r = R'(\hat{\theta}_T - \theta_0) \xrightarrow{D} N(0, R'\Sigma_0^{-1}\Omega_0\Sigma_0^{-1}R). \tag{VI.19}$$

The $l \times l$ matrix $R'\Sigma_0^{-1}\Omega_0\Sigma_0^{-1}R$ is positive definite, since $\Sigma_0^{-1}\Omega_0\Sigma_0^{-1}$ and R have full rank. Likewise, from (VI.17),

$$R'\hat{\Psi}_T R \xrightarrow{p} R'\Sigma_0^{-1}\Omega_0\Sigma_0^{-1}R. \tag{VI.20}$$

Combining (VI.19) and (VI.20) together with the continuous mapping theorem yields the desired result. ■

Remark VI.4.1 *By similar arguments, if one is interested in testing a hypothesis about the jth entry of θ, that is,*

$$H : \theta_j = r,$$

then one may use that t-statistic

$$\tau_{\theta_j=r} = \frac{\hat{\theta}_{T,j} - r}{\sqrt{(\hat{\Psi}_T)_{jj}/T}},$$

with $\hat{\theta}_{T,j}$ and $(\hat{\Psi}_T)_{jj}$ the j entry of $\hat{\theta}_T$ and the diagonal of $\hat{\Psi}_T$, respectively. Under the assumptions of Theorem VI.4.2,

$$\tau_{\theta_j=r} \xrightarrow{D} N(0,1).$$

For instance in terms of the AR(1) model, testing the hypothesis $\theta = \theta_0$ with $|\theta_0| < 1$, we may from Example VI.3.1 apply the t-statistic

$$\tau_{\theta=\theta_0} = \frac{\hat{\theta}_T - \theta_0}{\sqrt{(1-\hat{\theta}_T^2)/T}} \xrightarrow{D} N(0,1).$$

Example VI.4.2 (ARCH(1) ctd.) *We seek to apply Theorem VI.4.2 for constructing a Wald test for the hypothesis in (VI.13) based on the ML estimator $\hat{\theta}_T = (\hat{\omega}_T, \hat{\alpha}_T)'$. Note that under the hypothesis, $\alpha_0 < 1$, and as carefully argued in Section VI.3.1, all the conditions of Theorem VI.3.1 hold under this condition. It remains to find a consistent estimator of the asymptotic covariance of the ML estimator. Recall from Section VI.3.1,*

$$\sqrt{T}(\hat{\theta}_T - \theta_0) \xrightarrow{D} N(0, -\Sigma_0^{-1}),$$

with $\Sigma_0 = \Sigma(\theta_0)$ and

$$\Sigma(\theta) = \mathbb{E}\left[\frac{\partial^2 q_t(\theta)}{\partial\theta\partial\theta'}\right],$$

$$\frac{\partial^2 q_t(\theta)}{\partial\theta\partial\theta'} = \frac{1}{2}\frac{1}{\omega + \alpha x_{t-1}^2}\left(1 - \frac{2x_t^2}{\omega + \alpha x_{t-1}^2}\right) w_t w_t', \quad w_t = (1, x_{t-1}^2)'.$$

From Condition 5 of Theorem VI.3.1,

$$\sup_{\theta\in\mathcal{N}(\theta_0)} \left\| T^{-1}\sum_{t=1}^T \frac{\partial^2 q_t(\theta)}{\partial\theta\partial\theta'} - \Sigma(\theta)\right\| \xrightarrow{p} 0,$$

and this combined with Condition 1 and Lemma A.1, gives that

$$T^{-1}\sum_{t=1}^T \frac{\partial^2 q_t(\hat{\theta}_T)}{\partial\theta\partial\theta'} \xrightarrow{p} \Sigma(\theta_0) = \Sigma_0.$$

Hence a consistent estimator of the asymptotic covariance is

$$\begin{aligned}\hat{\Psi}_T &= \left(-T^{-1} \sum_{t=1}^{T} \frac{\partial^2 q_t(\hat{\theta}_T)}{\partial\theta\partial\theta'} \right)^{-1} \\ &= \left(T^{-1} \sum_{t=1}^{T} \frac{1}{2} \frac{1}{\hat{\omega}_T + \hat{\alpha}_T x_{t-1}^2} \left(1 - \frac{2x_t^2}{\hat{\omega}_T + \hat{\alpha}_T x_{t-1}^2} \right) w_t w_t' \right)^{-1},\end{aligned}$$

that satisfies

$$\hat{\Psi}_T \xrightarrow{p} -\Sigma_0^{-1}.$$

For this choice of estimator, under the hypothesis in (VI.13), the Wald statistic satisfies

$$W_{\text{rob}} = \frac{T(\hat{\alpha}_T - 1/\sqrt{3})^2}{(\hat{\Psi}_T)_{22}} \xrightarrow{D} \chi^2(1).$$

We end this section by considering and estimator for $\hat{\Psi}_T$. As demonstrated in the previous Example VI.4.2, in the case where $\sqrt{T}(\hat{\theta} - \theta_0)$ has asymptotic covariance given by $-\Sigma_0^{-1}$, we use

$$\hat{\Psi}_T = \left(-\frac{\partial^2 Q_T(\hat{\theta}_T)}{\partial\theta\partial\theta'} \right)^{-1},$$

which, as already argued, converges in probability to $-\Sigma_0^{-1}$. In the more general case where the asymptotic covariance is given by $\Sigma_0^{-1}\Omega_0\Sigma_0^{-1}$ one would in addition need a consistent estimator for Ω_0. We restrict here our attention to estimators given as maximizers of criterion functions of the form

$$Q_T(\theta) = \frac{1}{T} \sum_{t=1}^{T} q_t(\theta). \tag{VI.21}$$

This class of estimators includes the estimators for the parameters in AR and ARCH models already considered. We have the following result, which follows directly from applications of Lemma A.1.

Lemma VI.4.1 *Under the assumptions of Theorem VI.3.1,*

$$\hat{\Sigma}_T := \frac{\partial^2 Q_T(\hat{\theta}_T)}{\partial\theta\partial\theta'} \xrightarrow{p} \Sigma_0, \quad \text{as } T \to \infty.$$

Suppose in addition that the criterion function is given by (VI.21), and that there exists a matrix-valued function $\Omega(\theta)$ that is continuous at θ_0, such that

$$\sup_{\theta \in \mathcal{N}(\theta_0)} \left\| T^{-1} \sum_{t=1}^{T} \frac{\partial q_t(\theta)}{\partial\theta} \frac{\partial q_t(\theta)}{\partial\theta'} - \Omega(\theta) \right\| \xrightarrow{p} 0, \quad \text{as } T \to \infty,$$

and $\Omega(\theta_0) = \Omega_0$. *Then*

$$\hat{\Omega}_T := T^{-1} \sum_{t=1}^{T} \frac{\partial q_t(\hat{\theta}_T)}{\partial \theta} \frac{\partial q_t(\hat{\theta}_T)}{\partial \theta'} \xrightarrow{p} \Omega_0, \quad \text{as } T \to \infty,$$

and, consequently,

$$\hat{\Psi}_T := \hat{\Sigma}_T^{-1} \hat{\Omega}_T \hat{\Sigma}_T^{-1}$$

satisfies (VI.17).

Remark VI.4.2 *Note that the above lemma essentially restricts the criterion function to settings where the score contributions* $\partial q_t(\theta_0)/\partial\theta$ *are uncorrelated in time – or martingale differences – as considered for the AR and ARCH models. This is true, as the CLT for martingale differences, e.g., Theorem V.2.3, is compatible with a limiting covariance matrix given by* $\mathbb{E}[(\partial q_t(\theta_0)/\partial\theta)(\partial q_t(\theta_0)/\partial\theta')] = \Omega_0$. *If* $\partial q_t(\theta_0)/\partial\theta$ *is autocorrelated, the structure of the asymptotic covariance looks different; see, e.g.,Theorem I.4.3 in the context of geometrically ergodic processes. In such a case alternative covariance estimators are needed, see e.g. Francq and Zakoïan (2019, Chapter 5).*

VI.5 Concluding remarks

We conclude this note by emphasizing that the results for consistency and asymptotic normality applies to a general range of estimators. As demonstrated, the high-level conditions of Theorems VI.2.2 and VI.3.1 are typically manageable to verify in cases where the criterion function is of the form (VI.21) and the DGP is stationary and ergodic by relying on appropriate LLNs and CLTs.

An important assumption made in Theorem VI.3.1 is that the true parameter value θ_0 is an interior point of the parameter space. For instance, this rules out the case where $\alpha_0 = 0$ in the ARCH(1) model, as the parameter $\alpha \geq 0$. In such a case one can show that the limiting distribution of $\sqrt{T}(\hat{\alpha}_T - \alpha_0) = \sqrt{T}\hat{\alpha}_T \geq 0$ is non-Gaussian, and alternative arguments are needed. We refer to Cavaliere et al. (2022) for additional details.

References

Amemiya, T., 1985, *Advanced Econometrics*, Harvard University Press, 1st edition.

Cavaliere, G., Nielsen, H.B., Pedersen, R.S., & Rahbek, A., 2022, ”Bootstrap inference on the boundary of the parameter space with application to conditional volatility models“, *Journal of Econometrics*, Vol. 227, pp. 241–263.

Francq, C., & Zakoïan, J.-M., 2012, ”Strict stationarity testing and estimation of explosive and stationary generalized autoregressive conditional heteroscedasticity models“, *Econometrica*, Vol. 80, pp. 821–861.

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

Jensen, S.T., & Rahbek, A., 2004a, ”Asymptotic normality of the QMLE estimator of ARCH in the nonstationary case“, *Econometrica*, Vol. 72, pp. 641–646.

Jensen, S.T., & Rahbek, A., 2004b, ”Asymptotic inference for nonstationary GARCH“, *Econometric Theory*, Vol. 20, pp. 1203–1226.

Kristensen, D., & Rahbek, A., 2005, ”Asymptotics of the QMLE for a class of ARCH(q) models“, *Econometric Theory*, Vol. 21, pp. 946–961.

Lange, T., Rahbek, A., & Jensen, S.T., 2011, ”Estimation and asymptotic inference in the AR-ARCH model“, *Econometric Reviews*, Vol. 30, pp. 129–153.

Newey, W.K., & McFadden, D., 1994, ”Large sample estimation and hypothesis testing“, in R.F. Engle and D. McFadden (Eds.), *Handbook of Econometrics*, Vol. 4, pp.2111-2245.

Ranga Rao, R., 1962, ”Relations between weak and uniform convergence of measures with applications“, *The Annals of Mathematical Statistics*, Vol. 33, 659–680.

Straumann, D., & Mikosch, T., 2006, ”Quasi-maximum-likelihood estimation in conditionally heteroscedastic time series: A stochastic recurrence equations approach“, *The Annals of Statistics*, Vol. 34, pp. 2449–2495.

A Additional limit results

Theorem A.1 *Let $\{X_t\}_{t\in\mathbb{Z}}$ be stationary and ergodic with $X_t \in \mathbb{R}^d$. Let $\Theta \subset \mathbb{R}^k$ be compact, and let f be a measurable function from $(\mathbb{R}^d)^\infty \times \Theta$ to $\mathbb{R}$. Define*

$$q_t(\theta) = f(\ldots, X_{t-1}, X_t, X_{t+1}, \ldots; \theta).$$

If $q_t(\theta)$ is finite almost surely, then $\{q_t(\theta)\}_{t\in\mathbb{Z}}$ is stationary and ergodic. Furthermore, assume that $q_t(\theta)$ is continuous in θ almost surely and that $\mathbb{E}[\sup_{\theta\in\Theta} |q_t(\theta)|] < \infty$. Then

$$\sup_{\theta\in\Theta} \left| T^{-1} \sum_{t=1}^{T} q_t(\theta) - \mathbb{E}\left[q_t(\theta)\right] \right| \xrightarrow{p} 0 \quad \text{as } T \to \infty.$$

Proof. The stationarity and ergodicity of $\{q_t(\theta)\}_{t\in\mathbb{Z}}$ follows by Theorem V.2.1. The uniform convergence follows by Ranga Rao (1962); see also Mikosch and Straumann (2006, Section 2.3) for additional details. ■

Lemma A.1 *Suppose that*

1. $X_T \xrightarrow{p} x_0 \in \mathbb{R}^k$,
2. $\sup_{x\in B(x_0,\epsilon)} |g_T(\theta) - g(\theta)| \xrightarrow{p} 0$ *with* $B(x_0,\epsilon) := \{x : \|x - x_0\| < \epsilon\}$ *for some* $\epsilon > 0$, *and*
3. *the non-stochastic function* $g(x)$ *is continuous at* x_0.

Then $g_T(X_T) \xrightarrow{p} Q(x_0)$.

Proof. We have with probability approaching one,

$$\begin{aligned} |g_T(X_T) - g(x_0)| &\leq |g_T(X_T) - g(X_T)| + |g(X_T) - g(x_0)| \\ &\leq \sup_{x\in B(x_0,\epsilon)} |g_T(x) - g(x)| + |g(X_T) - g(x_0)|, \end{aligned}$$

where the first inequality follows by the triangle inequality, and second equality follows from the fact that $X_T \in B(x_0,\epsilon)$ with probability approaching one. Conditions 1 and 3 imply that $|g(X_T) - g(x_0)| \xrightarrow{p} 0$. This combined with Condition 2 yield that $|g_T(X_T) - g(x_0)| \xrightarrow{p} 0$. ■

Lemma A.2 (Lange et al., 2011, Lemma 3) *Let $\{x_t\}_{t=-k,-k+1,\ldots}$ be a geometrically ergodic Markov chain for some finite $k \geq 0$. Let $w_t := (x_{t-k}, \ldots, x_t)$ and let $q(w,\theta)$ be a real-valued measurable function which is continuous in θ*

for all w. Let Θ be a compact set of the same dimension as θ. Assume that $E[q(w_t^*, \theta)] = 0$ for all $\theta \in \Theta$, and that $E[\sup_{\theta\in\Theta} |q(w_t^*, \theta)|] < \infty$. Then

$$\sup_{\theta\in\Theta} \left| T^{-1} \sum_{t=1}^{T} q(w_t, \theta) \right| \xrightarrow{p} 0 \quad \textit{as } T \to \infty.$$

B Proofs

Proof of Theorem VI.2.2

The proof follows NM (pp.2121–2122). Let $Q(\hat\theta_T) := E[q_t(\theta)]|_{\theta=\hat\theta_T}$. For any $\epsilon > 0$ we have with probability approaching one that

$$\begin{aligned} Q_T(\hat\theta_T) &> Q_T(\theta_0) - \epsilon/3, \\ Q(\hat\theta_T) &> Q_T(\hat\theta_T) - \epsilon/3, \\ Q_T(\theta_0) &> Q(\theta_0) - \epsilon/3, \end{aligned}$$

where the first inequality follows by the fact that $\hat\theta_T$ is a maximizer of $Q_T(\theta)$ on Θ, and the second and third follow by Condition 4. Combining these inequalities, we have with probability approaching one,

$$Q(\hat\theta_T) > Q_T(\hat\theta_T) - \epsilon/3 > Q_T(\theta_0) - 2\epsilon/3 > Q(\theta_0) - \epsilon.$$

Hence, for any $\epsilon > 0$, with probability approaching one,

$$Q(\hat\theta_T) > Q(\theta_0) - \epsilon. \tag{VI.22}$$

Let $\mathcal{N}(\theta_0)$ be *any* open subset of Θ containing θ_0, and note that $\Theta \cap \mathcal{N}(\theta_0)^\complement$ is compact. Hence, using continuity of $Q(\theta)$, there exists a $\theta^* \in \Theta \cap \mathcal{N}(\theta_0)^\complement$ such that

$$Q(\theta^*) = \sup_{\theta\in\Theta\cap\mathcal{N}(\theta_0)^\complement} Q(\theta).$$

Since θ_0 is the unique maximizer of $Q(\theta)$ on Θ, and $\theta_0 \notin \Theta \cap \mathcal{N}(\theta_0)^\complement$, it must hold that

$$Q(\theta^*) < Q(\theta_0).$$

Choosing $\epsilon = Q(\theta_0) - Q(\theta^*) > 0$, it follows that with probability approaching one,

$$Q(\hat\theta_T) > Q(\theta_0) - \epsilon > Q(\theta^*) = \sup_{\theta\in\Theta\cap\mathcal{N}(\theta_0)^\complement} Q(\theta).$$

Hence, $\hat\theta_T \in \mathcal{N}(\theta_0)$ with probability approaching one, that is, $\mathbb{P}(\hat\theta_T \in \mathcal{N}(\theta_0)) \to 1$. Note that this holds for any $\mathcal{N}(\theta_0)$ (being an open subset of Θ containing

θ_0). For any $\epsilon > 0$, there exists an open subset $\mathcal{N}_\epsilon(\theta_0)$ such that the event $\hat{\theta}_T \in \mathcal{N}_\epsilon(\theta_0) \implies \|\hat{\theta}_T - \theta_0\| < \epsilon$. Hence, for any $\epsilon > 0$,

$$\mathbb{P}\left(\|\hat{\theta}_T - \theta_0\| < \epsilon\right) \geq \mathbb{P}(\hat{\theta}_T \in \mathcal{N}_\epsilon(\theta_0)) \to 1,$$

and we conclude that $\hat{\theta}_T \xrightarrow{p} \theta_0$ as $T \to \infty$.

Proof of Theorem VI.4.1

The proof follows the arguments given in the proofs of Theorems 9.1 and 9.2 in NM. Recall that.

$$LR_T(H) = 2T[Q_T(\hat{\theta}_T) - Q_T(\tilde{\theta}_T)].$$

A second order mean-value expansion of $Q_T(\tilde{\theta}_T)$ around $\hat{\theta}_T$ gives that

$$Q_T(\tilde{\theta}_T) = Q_T(\hat{\theta}_T) + \frac{\partial Q_T(\hat{\theta}_T)}{\partial \theta'}(\tilde{\theta}_T - \hat{\theta}_T) + \frac{1}{2}(\tilde{\theta}_T - \hat{\theta}_T)' \frac{\partial^2 Q_T(\theta_T^*)}{\partial \theta \partial \theta'}(\tilde{\theta}_T - \hat{\theta}_T),$$

with θ_T^* between $\tilde{\theta}_T$ and $\hat{\theta}_T$, and $\theta_T^* \xrightarrow{p} \theta_0$. Recall that with probability approaching one $\partial Q_T(\hat{\theta}_T)/\partial \theta' = 0_{k\times 1}$, so that

$$LR_T(H) = T(\hat{\theta}_T - \tilde{\theta}_T)' \left[-\frac{\partial^2 Q_T(\theta_T^*)}{\partial \theta \partial \theta'}\right](\hat{\theta}_T - \tilde{\theta}_T). \tag{VI.23}$$

Note that the right-hand side of (VI.23) is a Wald-type statistic analogous to (VI.15). From Lemma VI.4.1 $\partial^2 Q_T(\theta_T^*)/\partial\theta\partial\theta' \xrightarrow{p} \Sigma_0$, so it remains to find the limiting distribution of

$$\sqrt{T}(\hat{\theta}_T - \tilde{\theta}_T) = \sqrt{T}(\hat{\theta}_T - \theta_0) - \sqrt{T}(\tilde{\theta}_T - \theta_0).$$

To do so, we derive expressions for $\sqrt{T}(\hat{\theta}_T - \theta_0)$ [Step 1] and $\sqrt{T}(\tilde{\theta}_T - \theta_0)$ [Step 2], and combine [Step 3], and lastly plug into (VI.23) [Step 4].

Step 1: From the arguments given in the proof of Theorem VI.3.1, we have that

$$\sqrt{T}(\hat{\theta}_T - \theta_0) = -\Sigma_0^{-1}\sqrt{T}\frac{\partial Q_T(\theta_0)}{\partial \theta} + o_p(1). \tag{VI.24}$$

Step 2: Note that $\tilde{\theta}_T$ solves a maximization problem with Lagrangian

$$\mathcal{L}_T(\theta) = Q_T(\theta) - [R'\theta - r]'\lambda_T,$$

where λ_T is a $(l \times 1)$ vector of Lagrange multipliers. The first-order condition to this maximization problem is given by

$$0_{k\times 1} = \sqrt{T}\frac{\partial Q_T(\tilde{\theta}_T)}{\partial \theta} - \sqrt{T}R\lambda_T. \tag{VI.25}$$

A mean-value expansion of $\partial Q_T(\tilde{\theta}_T)/\partial\theta$ around θ_0 gives

$$\frac{\partial Q_T(\tilde{\theta}_T)}{\partial \theta} = \frac{\partial Q_T(\theta_0)}{\partial \theta} + \frac{\partial^2 Q_T(\theta_T^{**})}{\partial\theta\partial\theta'}(\tilde{\theta}_T - \theta_0), \tag{VI.26}$$

with θ_T^{**} between $\tilde{\theta}_T$ and θ_0, and $\theta_T^{**} \xrightarrow{p} \theta_0$. Note that by Lemma VI.4.1 $\partial^2 Q_T(\theta_T^{**})/\partial\theta\partial\theta' \xrightarrow{p} \Sigma_0$, so $\partial^2 Q_T(\theta_T^{**})/\partial\theta\partial\theta'$ is invertible with probability approaching one. Under the hypothesis H it holds that $r = R'\theta_0$, such that

$$0_{l\times 1} = [R'\tilde{\theta}_T - r] = R'(\tilde{\theta}_T - \theta_0). \tag{VI.27}$$

Combining (VI.25)-(VI.27), gives that

$$\begin{bmatrix} 0_{k\times 1} \\ 0_{l\times 1} \end{bmatrix} = \begin{bmatrix} \frac{\partial Q_T(\theta_0)}{\partial \theta} \\ 0 \end{bmatrix} - \begin{bmatrix} -\frac{\partial^2 Q_T(\theta_T^{**})}{\partial\theta\partial\theta'} & R \\ R' & 0 \end{bmatrix} \begin{bmatrix} \sqrt{T}(\tilde{\theta}_T - \theta_0) \\ \sqrt{T}\lambda_T \end{bmatrix},$$

and we seek to solve for $\sqrt{T}(\tilde{\theta}_T - \theta_0)$. We have the following result from linear algebra: Suppose that the $(l \times k)$ matrix A with $l \leq k$ has rank l, and that the $(k \times k)$ matrix B has full rank. Then

$$\begin{bmatrix} B & A' \\ A & 0 \end{bmatrix}^{-1} = \begin{bmatrix} B^{-1/2}MB^{-1/2} & B^{-1}A'(AB^{-1}A')^{-1} \\ (AB^{-1}A')^{-1}AB^{-1} & -(AB^{-1}A')^{-1} \end{bmatrix},$$

with $M = I_k - B^{-1/2}A'(AB^{-1}A')^{-1}AB^{-1/2}$. Using this result, we find that

$$\sqrt{T}(\tilde{\theta}_T - \theta_0) = \left(-\frac{\partial^2 Q_T(\theta_T^{**})}{\partial\theta\partial\theta'}\right)^{-1/2} M_T \left(-\frac{\partial^2 Q_T(\theta_T^{**})}{\partial\theta\partial\theta'}\right)^{-1/2} \sqrt{T}\frac{\partial Q_T(\theta_0)}{\partial\theta},$$

with

$$\begin{aligned} M_T &:= I_k - \left(-\frac{\partial^2 Q_T(\theta_T^{**})}{\partial\theta\partial\theta'}\right)^{-1/2} R \left(R' \left(-\frac{\partial^2 Q_T(\theta_T^{**})}{\partial\theta\partial\theta'}\right)^{-1} R\right)^{-1} R' \left(-\frac{\partial^2 Q_T(\theta_T^{**})}{\partial\theta\partial\theta'}\right)^{-1/2} \\ &\xrightarrow{p} I_k - (-\Sigma_0)^{-1/2}R\left[R'(-\Sigma_0)^{-1}R\right]^{-1}R'(-\Sigma_0)^{-1/2}. \end{aligned}$$

Since $\sqrt{T}\partial Q_T(\theta_0)/\partial\theta = O_p(1)$,

$$\begin{aligned}
&\sqrt{T}(\tilde{\theta}_T - \theta_0) \\
&= (-\Sigma_0)^{-1} - (-\Sigma_0)^{-1}R\left[R'(-\Sigma_0)^{-1}R\right]^{-1}R'(-\Sigma_0)^{-1}\sqrt{T}\frac{\partial Q_T(\theta_0)}{\partial\theta} + o_p(1).
\end{aligned} \tag{VI.28}$$

Step 3: Combining (VI.24) and (VI.28) gives that

$$\begin{aligned}
\sqrt{T}(\hat{\theta}_T - \tilde{\theta}_T) &= \sqrt{T}(\hat{\theta}_T - \theta_0) - \sqrt{T}(\tilde{\theta}_T - \theta_0) \\
&= \left[(-\Sigma_0)^{-1}R\left[R'(-\Sigma_0)^{-1}R\right]^{-1}R'(-\Sigma_0)^{-1}\right]\sqrt{T}\frac{\partial Q_T(\theta_0)}{\partial\theta} + o_p(1) \\
&\overset{D}{\to} Z,
\end{aligned} \tag{VI.29}$$

with $Z \overset{D}{=} N(0, B)$ and, using that $\Omega_0 = -\Sigma_0$,

$$\begin{aligned}
B &= \left[(-\Sigma_0)^{-1}R\left[R'(-\Sigma_0)^{-1}R\right]^{-1}R'(-\Sigma_0)^{-1}\right]\Omega_0\left[(-\Sigma_0)^{-1}R\left[R'(-\Sigma_0)^{-1}R\right]^{-1}R'(-\Sigma_0)^{-1}\right]' \\
&= \left[(-\Sigma_0)^{-1}R\left[R'(-\Sigma_0)^{-1}R\right]^{-1}R'(-\Sigma_0)^{-1}\right](-\Sigma_0)\left[(-\Sigma_0)^{-1}R\left[R'(-\Sigma_0)^{-1}R\right]^{-1}R'(-\Sigma_0)^{-1}\right]' \\
&= \left[(-\Sigma_0)^{-1}R\left[R'(-\Sigma_0)^{-1}R\right]^{-1}R'(-\Sigma_0)^{-1}\right].
\end{aligned}$$

Step 4: From (VI.23), using that $\sqrt{T}(\hat{\theta}_T - \tilde{\theta}_T) = O_p(1)$,

$$LR_T(H) = T(\hat{\theta}_T - \tilde{\theta}_T)'(-\Sigma_0)(\hat{\theta}_T - \tilde{\theta}_T) + o_p(1).$$

Noting that B has rank l, and that $B(-\Sigma_0)B = B$, it follows from Lemma 9.7 in NM and (VI.29) that

$$LR_T(H) \overset{D}{\to} Z'(-\Sigma_0)Z \overset{D}{=} \chi^2_l.$$

Anders Rahbek
Rasmus Søndergaard Pedersen
University of Copenhagen

October 2024

Part VII
Models for time-varying volatility and their applications

This chapter provides an overview of ARCH and Generalized ARCH (GARCH) models. We state conditions for stationarity and ergodicity of ARCH and GARCH processes, relying on the results for SREs considered in Chapter V, and emphasize that similar conditions can be derived by relying on the drift criterion considered in Chapter I. Moreover, we present limit theory for the ML (and quasi-ML) estimators for ARCH(1) models based on the theory for extremum estimators considered in Chapter VI. As emphasized in Chapter VI, the limit results for the estimators can be derived using either limit theorems for stationary and ergodic processes or for geometrically ergodic Markov chains. As possible extensions of the models, we consider non-linear models, multivariate (MGARCH) models as well as alternative error distributions. We conclude the chapter by presenting two applications of the models in relation to portfolio choice (Section VII.4) and risk quantification (Section VII.5). We emphasize that ARCH models and their extensions have several other applications within finance such as option pricing (Chorro et al., 2015) and empirical asset pricing (e.g., Engle, 2016 and Blasques et al., 2024).

VII.1 ARCH (1)

Consider the ARCH(1) process given by

$$x_t = \sigma_t z_t, \quad t \in \mathbb{Z}, \tag{VII.1}$$

$$\sigma_t^2 = \omega + \alpha x_{t-1}^2, \quad \omega > 0, \alpha \geq 0, \tag{VII.2}$$

where

$$\{z_t\}_{t\in\mathbb{Z}} \text{ is an i.i.d. process with } z_t \overset{D}{=} N(0,1), \tag{VII.3}$$

and z_t and x_{t-1} are independent for all t. From Chapter I we have that (almost surely)

$$\mathbb{E}[x_t|x_{t-1}] = 0 \quad \text{and} \quad \mathbb{V}[x_t|x_{t-1}] = \mathbb{E}[x_t^2|x_{t-1}] = \sigma_t^2,$$

that is, the process has zero conditional mean for all t but time-varying conditional variance. In particular, since $z_t \overset{D}{=} N(0,1)$ and σ_t are independent

$$x_t|x_{t-1} \overset{D}{=} N(0,\sigma_t^2),$$

such that x_t has conditional density

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

Hence the *conditional* distribution of x_t is Gaussian, while as you may recall from Example V.4.6 in Chapter V, under certain assumptions, the unconditional distribution of x_t has so-called power law tails, meaning that some moments of the distribution are infinite. The remainder of this section summarizes the probabilistic properties of the ARCH(1) found in Chapter V and states the properties of the ML estimator for the model parameters found in Chapter VI. Lastly, we consider the notion of quasi-maximum likelihood estimation.

VII.1.1 Stochastic properties

From Examples V.4.2 and V.4.4 in Chapter V we have the following result.

Corollary VII.1.1 *Consider the ARCH(1) process $\{x_t\}_{t\in\mathbb{Z}}$ given by (VII.1)-(VII.3). If $\alpha < 3.56\ldots$, then $\{x_t\}_{t\in\mathbb{Z}}$ is a stationary and ergodic process. If in addition $\alpha > 0$, then*

$$\mathbb{E}[|x_t|^p] < \infty \quad \textit{if and only if} \quad \mathbb{E}[|\sqrt{\alpha}z_t|^p] < 1,$$

for $p \in [0,\infty)$. If $\alpha = 0$, $\mathbb{E}[|x_t|^p] = \omega^{p/2}\mathbb{E}[|z_t|^p] < \infty$ for all $p \in [0,\infty)$.

The corollary allows us to derive conditions for finite moments and well-defined autocovariances of x_t, as considered in the next example.

Example VII.1.1 *Based on Corollary VII.1.1, if $\alpha < 1$ then $\mathbb{E}[x_t^2] < \infty$ and hence $\mathbb{E}[\sigma_t^2] < \infty$. We then have that*

$$\mathbb{E}[x_t^2] = \mathbb{E}[\sigma_t^2 z_t^2] = \mathbb{E}[\sigma_t^2]\mathbb{E}[z_t^2] = \mathbb{E}[\sigma_t^2] = \mathbb{E}[\omega + \alpha x_{t-1}^2] = \omega + \alpha\mathbb{E}[x_{t-1}^2],$$

where we have used that z_t and σ_t are independent. By stationarity, $\mathbb{E}[x_t^2] = \mathbb{E}[x_{t-1}^2]$, such that

$$\mathbb{E}[x_t^2] = \frac{\omega}{1-\alpha}.$$

In this case, we also have that for any $k > 0$,

$$\mathbb{E}[x_t x_{t-k}] = \mathbb{E}[\mathbb{E}[x_t | x_{t-k}] x_{t-k}] = 0,$$

such that the autocovariance function of x_t is zero at any lag order $k > 0$. Likewise if $\alpha < 1/\sqrt{3}$, then $\mathbb{E}[x_t^4] < \infty$. It holds that

$$\mathbb{E}[x_t^4] = \mathbb{E}[\sigma_t^4 z_t^4] = \mathbb{E}[\sigma_t^4]\mathbb{E}[z_t^4] = 3\mathbb{E}[\sigma_t^4],$$

with

$$\begin{aligned}\mathbb{E}[\sigma_t^4] &= \mathbb{E}[(\omega + \alpha x_{t-1}^2)^2] \\ &= \mathbb{E}[\omega^2 + \alpha^2 x_{t-4}^4 + 2\omega\alpha x_{t-1}^2] \\ &= \omega^2 + \alpha^2 \mathbb{E}[x_t^4] + 2\omega\alpha\mathbb{E}[x_t^2] \\ &= \omega^2 + \alpha^2 \mathbb{E}[x_t^4] + \frac{2\omega^2\alpha}{1-\alpha}.\end{aligned}$$

Hence,

$$\mathbb{E}[x_t^4] = 3\left(\omega^2 + \alpha^2\mathbb{E}[x_t^4] + \frac{2\omega^2\alpha}{1-\alpha}\right),$$

such that

$$\mathbb{E}[x_t^4] = 3\frac{\omega^2 + \frac{2\omega^2\alpha}{1-\alpha}}{1-3\alpha^2} = 3\left(\frac{1-\alpha^2}{1-3\alpha^2}\right)\left(\frac{\omega}{1-\alpha}\right)^2.$$

Moreover, under the same assumption,

$$\begin{aligned}\mathbb{E}[x_t^2 x_{t-1}^2] &= \mathbb{E}[z_t^2 \sigma_t^2 x_{t-1}^2] \\ &= \mathbb{E}[\sigma_t^2 x_{t-1}^2] \\ &= \mathbb{E}[(\omega + \alpha x_{t-1}^2) x_{t-1}^2] \\ &= \omega\mathbb{E}[x_t^2] + \alpha\mathbb{E}[x_t^4] \\ &= \frac{\omega^2}{1-\alpha} + 3\alpha\left(\frac{1-\alpha^2}{1-3\alpha^2}\right)\left(\frac{\omega}{1-\alpha}\right)^2 \\ &= \left[3\alpha\left(\frac{1-\alpha^2}{1-3\alpha^2}\right) + (1-\alpha)\right]\left(\frac{\omega}{1-\alpha}\right)^2,\end{aligned}$$

such that autocovariance of x_t^2 of order one is given by

$$\begin{aligned}Cov(x_t^2, x_{t-1}^2) &= \mathbb{E}[x_t^2 x_{t-1}^2] - \mathbb{E}[x_t^2]\mathbb{E}[x_{t-1}^2] \\ &= \left[3\alpha\left(\frac{1-\alpha^2}{1-3\alpha^2}\right) + (1-\alpha)\right]\left(\frac{\omega}{1-\alpha}\right)^2 - \left(\frac{\omega}{1-\alpha}\right)^2 \\ &= \left[3\alpha\left(\frac{1-\alpha^2}{1-3\alpha^2}\right) - \alpha\right]\left(\frac{\omega}{1-\alpha}\right)^2 \\ &= 2\alpha\left(\frac{\omega}{1-\alpha}\right)^2,\end{aligned}$$

and the autocorrelation of x_t^2 of order one is

$$\begin{aligned} Corr(x_t^2, x_{t-1}^2) &= \frac{Cov(x_t^2, x_{t-1}^2)}{\sqrt{\mathbb{E}[x_t^4]\mathbb{E}[x_{t-1}^4]}} \\ &= \frac{Cov(x_t^2, x_{t-1}^2)}{\mathbb{E}[x_t^4]} \\ &= \frac{2\alpha}{3}\left(\frac{1-3\alpha^2}{1-\alpha^2}\right) \geq 0. \end{aligned}$$

VII.1.2 ML Estimation

In applications, such as the ones described in Sections VII.4 and VII.5 below, one typically needs estimates of the ARCH model parameters. The estimation is typically carried out by ML, as detailed in this section. Given a sample $\{x_t\}_{t=0}^T$ generated by the ARCH(1) process in (VII.1)-(VII.3), we seek to estimate the parameter vector $\theta = (\omega, \alpha)'$. Recall from Chapter VI that the log-likelihood function is given by

$$\begin{aligned} Q_T(\theta) &= \frac{1}{T}\sum_{t=1}^T q_t(\theta), \qquad \text{(VII.4)} \\ q_t(\theta) &= \log f_\theta(x_t|x_{t-1}), \\ f_\theta(x_t|x_{t-1}) &= \frac{1}{\sqrt{2\pi\sigma_t^2(\theta)}} \exp\left\{-\frac{x_t^2}{2\sigma_t^2(\theta)}\right\}, \\ \sigma_t^2(\theta) &= \omega + \alpha x_{t-1}^2. \end{aligned}$$

We consider estimation over the compact parameter space

$$\Theta = [\omega_L, \omega_U] \times [0, \alpha_U] \qquad \text{(VII.5)}$$

with $0 < \omega_L < \omega_U < \infty$ and $0 < \alpha_U < \infty$, and let $\hat{\theta}_T$ denote the ML estimator given by

$$\hat{\theta}_T = \arg\max_{\theta\in\Theta} Q_T(\theta), \qquad \text{(VII.6)}$$

with $Q_T(\theta)$ given by (VII.4) and Θ given by (VII.5). Moreover, the true value of θ (that is, the value of θ for the DGP) is labelled θ_0. From Examples VI.2.5 and VI.4.2 and Section VI.3.1 we have the following result.

Corollary VII.1.2 *Let $\{x_t\}_{t=0}^T$ follow the ARCH(1) process in (VII.1)-(VII.3) with true value $\theta_0 = (\omega_0, \alpha_0)' \in \Theta$ and $\alpha_0 < 1$. Then the ML estimator $\hat{\theta}_T$ in (VII.6) is consistent for θ_0, that is*

$$\hat{\theta}_T \xrightarrow{p} \theta_0 \quad \textit{as } T \to \infty.$$

Suppose in addition that θ_0 lies in the interior of Θ. Then

$$\sqrt{T}(\hat{\theta}_T - \theta_0) \xrightarrow{D} N(0, -\Sigma_0^{-1}),$$

with $-\Sigma_0^{-1}$ a positive definite matrix with

$$\Sigma_0 = \mathbb{E}\left[\frac{\partial^2 q_t(\theta_0)}{\partial\theta\partial\theta'}\right]. \tag{VII.7}$$

A consistent estimator for Σ_0 is given by

$$\hat{\Sigma}_T = \frac{1}{T}\sum_{t=1}^{T}\frac{\partial^2 q_t(\hat{\theta}_T)}{\partial\theta\partial\theta'}.$$

VII.1.3 Quasi-maximum likelihood

Often in applied work, upon estimation, the ARCH(1) model (or its extension as considered below) is found to have non-Gaussian innovations, that is, estimated innovations $\hat{z}_t = x_t/\sigma_t(\hat{\theta}_T)$ do not appear (approximately) standard normal as the theory would suggest; see e.g. Tsay (2010, Chapter 3) and Mikosch and Starica (2000, Section 6) for empirical examples. Fortunately, it shows up that reliable estimation and inference can still be done using the estimator $\hat{\theta}_T$ given by (VII.6). Suppose that the DGP is given by the ARCH(1) process in (VII.1)-(VII.2) under the assumption that

$$\{z_t\}_{t\in\mathbb{Z}} \text{ is an i.i.d. process with } \mathbb{E}[z_t] = 0 \text{ and } \mathbb{E}[z_t^2] = 1. \tag{VII.8}$$

Then x_t^2 satisfies the SRE

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

and $\{x_t^2\}_{t\in\mathbb{Z}}$ (and hence $\{x_t\}_{t\in\mathbb{Z}}$; see Exercises) is strictly stationary and ergodic provided that $\mathbb{E}[\log(\alpha z_t^2)] < 0$ (Theorem V.4.1). Moreover, from Theorem V.4.2 $\mathbb{E}[x_t^2] < \infty$ if $\mathbb{P}(z_t^2 = 1) < 1$ and $\alpha < 1$. In terms of the estimator $\hat{\theta}_T$ we have the following result that follows by arguments given in Example VI.2.5 and Section VI.3.1.

Corollary VII.1.3 *Suppose that the DGP $\{x_t\}_{t\in\mathbb{Z}}$ is given by the ARCH(1) process in (VII.1)-(VII.2) and (VII.8) with true value $\theta_0 = (\omega_0, \alpha_0)'$ and $\alpha_0 < 1$. Assume that $\theta_0 \in \Theta$ with Θ given by (VII.5) and $\mathbb{P}(z_t^2 = 1) < 1$. Then the estimator $\hat{\theta}_T$ given by (VII.6) satisfies*

$$\hat{\theta}_T \xrightarrow{p} \theta_0, \quad \text{as } T \to \infty.$$

Suppose in addition that $\mathbb{E}[z_t^4] < \infty$ *and* θ_0 *lies in the interior of* Θ. *Then*

$$\sqrt{T}(\hat{\theta}_T - \theta_0) \overset{D}{\to} N(0, \Sigma_0^{-1}\Omega_0\Sigma_0^{-1}),$$

with Σ_0 *invertible given by (VII.7) and* Ω_0 *positive definite given by*

$$\Omega_0 = \mathbb{E}\left[\frac{\partial q_t(\theta_0)}{\partial \theta}\frac{\partial q_t(\theta_0)}{\partial \theta'}\right].$$

The above result is powerful in the sense that the estimator $\hat{\theta}_T$ is useful (under mild conditions) even if one is not willing to specify a particular distribution for z_t. In this setting the objective function $Q_T(\theta)$ in (VII.4) is not necessarily the true log-likelihood function, and $Q_T(\theta)$ is referred to as the *quasi*-log-likelihood function, and $\hat{\theta}_T$ is the quasi-ML (QML) estimator. The estimator has another limiting variance $(\Sigma_0^{-1}\Omega_0\Sigma_0^{-1})$ than the ML estimator $(-\Sigma_0^{-1})$. It can be shown that $\Omega_0 = \mathbb{E}[(z_t^4 - 1)/2](-\Sigma_0)$, such that $\Omega_0 = -\Sigma_0$ in the case where $z_t \overset{D}{=} N(0,1)$. The matrix Σ_0 can be estimated by $\hat{\Sigma}_T$ provided in Corollary VII.1.2, whereas the quantity $\mathbb{E}[(z_t^4 - 1)/2]$ may be estimated from the standardized residuals, that is, by $T^{-1}\sum_{t=1}^T(\hat{z}_t^4 - 1)/2$.

Note that asymptotic normality of the QML estimator relies on the assumption that $\mathbb{E}[z_t^4] < \infty$, and the asymptotic covariance matrix is increasing (entry-wise) in $\mathbb{E}[z_t^4]$. Hence, the QML estimator is potentially imprecise when $\mathbb{E}[z_t^4]$ is large. Moreover, in the case where $\mathbb{E}[z_t^4] = \infty$, the rate of convergence of the QML estimator is slower than $\sqrt{T}$ and the limiting distribution is given in terms of a random vector with a non-Gaussian stable distribution; see Mikosch and Straumann (2006) for technical details and precise assumptions. Consequently, if one expects (or the estimation results suggests) that z_t is heavy-tailed, in the sense that $\mathbb{E}[z_t^4]$ is very large or infinite, it might be desirable to consider a model that directly addresses this feature, such as the model with (scaled) Student's t innovations described in Section VII.2.4 below.

We emphasize that the above considerations about quasi-ML estimation based on the Gaussian pdf applies to any extension, including the multivariate setting, of the ARCH model considered in the following sections.

VII.2 Extensions

The ARCH(1) model can be extended in multiple ways. A general structure is of the form

$$x_t = \mu_t + \sigma_t z_t, \quad t \in \mathbb{Z}, \tag{VII.9}$$

with $\{z_t\}_{t\in\mathbb{Z}}$ an i.i.d. process with $\mathbb{E}[z_t]=0$ and $\mathbb{E}[z_t^2]=1$, and z_t independent of $\mathcal{F}_{t-1}$ for all t. Here $\mu_t, \sigma_t \in \mathcal{F}_{t-1}$ with $\mathcal{F}_t$ the natural filtration generated by $\{x_s\}_{s\leq t}$ and potentially other (observable) processes. The conditional mean μ_t may be constant or contain lagged values of x_t, e.g. $\mu_t = \delta + \rho x_{t-1}$ for constants δ and ρ. Likewise, the conditional variance σ_t^2 could include several lags of x_t (ARCH(q)), lags of σ_t^2 (GARCH), non-linear terms (e.g., GJR-GARCH), or explanatory covariates (GARCH-X), all considered in the following sections. For most practical purposes a minimal requirement is that $\mathbb{P}(\sigma_t^2 > 0) = 1$ for all t. We refer to Bollerslev's (2009) "Glossary to ARCH (GARCH“ for an overview of potential specifications for σ_t^2.

VII.2.1 ARCH(q)

Similar to the AR(k) extension of the AR(1) model in Chapter II, we have ARCH(q) models of the form

$$\begin{aligned} x_t &= \sigma_t z_t, \quad t \in \mathbb{Z}, \\ \sigma_t^2 &= \omega + \alpha_1 x_{t-1}^2 + \cdots + \alpha_q x_{t-q}^2, \end{aligned}$$

with the innovations z_t satisfying (VII.3), and $\omega > 0$, $\alpha_1, \ldots, \alpha_q \geq 0$. With $X_t := (x_t^2, x_{t-1}^2, \ldots, x_{t-q+1}^2)'$ we have the SRE

$$X_t = A_t X_{t-1} + B_t,$$

with

$$A_t = \begin{pmatrix} z_t^2\alpha_1 & z_t^2\alpha_2 & \cdots & \cdots & z_t^2\alpha_q \\ 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 & 0 \end{pmatrix}, \quad B_t = \begin{pmatrix} z_t^2\omega \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

and conditions for stationarity and ergodicity can be derived from Theorem V.5.1. Estimation can be carried out by means of ML or quasi-ML as above. The motivation for letting $q > 1$ is to have a richer model that provides a better specification of the conditional variance of x_t and that takes into account the fact that squared (or absolute) returns are typically found to have a large order of non-zero autocorrelation; see, e.g., the recent work by Nielsen and Rahbek (2024) that consider ARCH models with $q > 100$.

VII.2.2 GARCH

The most prominent extension of the ARCH(1) model is the Generalized ARCH (GARCH) of Bollerslev (1986) and Taylor (1986) given by

$$\begin{aligned} x_t &= \sigma_t z_t, \quad t \in \mathbb{Z}, \\ \sigma_t^2 &= \omega + \alpha x_{t-1}^2 + \beta \sigma_{t-1}^2, \end{aligned}$$

with the innovations z_t satisfying (VII.3), $\omega > 0$ and $\alpha, \beta \geq 0$. The GARCH model is the workhorse model in classical volatility modelling and is widely used in empirical work.

The conditional variance σ_t^2 obeys the SRE

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

such that $\{\sigma_t^2\}_{t\in\mathbb{Z}}$ is stationary and ergodic if

$$\mathbb{E}[\log(\alpha z_t^2 + \beta)] < 0.$$

To conclude that $\{x_t\}_{t\in\mathbb{Z}}$ is stationary and ergodic under the same condition, note that

$$\begin{pmatrix} z_{t-1} \\ \sigma_t^2 \end{pmatrix} = \underbrace{\begin{pmatrix} 0 & 0 \\ 0 & \alpha z_{t-1}^2 + \beta \end{pmatrix}}_{A_t} \begin{pmatrix} z_{t-2} \\ \sigma_{t-1}^2 \end{pmatrix} + \begin{pmatrix} z_{t-1} \\ \omega \end{pmatrix},$$

is an SRE. By Theorem V.5.1 the process $\{(z_{t-1}, \sigma_t^2)'\}_{t\in\mathbb{Z}}$ is strictly stationary and ergodic, if the top Lyapunov exponent, γ, associated with the SRE is negative, which is immediate, as

$$\begin{aligned} \gamma &= \inf_{t\geq 1} \frac{1}{t}\mathbb{E}[\log \| \prod_{i=1}^{t} A_i \|] \\ &= \inf_{t\geq 1} \frac{1}{t}\mathbb{E}\left[\log \left\| \prod_{i=1}^{t} \begin{pmatrix} 0 & 0 \\ 0 & \alpha z_i^2 + \beta \end{pmatrix} \right\|\right] \\ &= \inf_{t\geq 1} \frac{1}{t}\mathbb{E}\left[\log \left\| \begin{pmatrix} 0 & 0 \\ 0 & \prod_{i=1}^{t}(\alpha z_i^2 + \beta) \end{pmatrix} \right\|\right] \\ &= \inf_{t\geq 1} \frac{1}{t}\mathbb{E}[\log \prod_{i=1}^{t}(\alpha z_i^2 + \beta)] \\ &= \inf_{t\geq 1} \frac{1}{t}\sum_{t=1}^{t}\mathbb{E}[\log(\alpha z_t^2 + \beta)] \\ &= \mathbb{E}[\log(\alpha z_t^2 + \beta)] < 0. \end{aligned}$$

Since, $x_t = z_t\sqrt{\sigma_t^2}$ is a measurable function of $\{(z_{t-1}, \sigma_t^2)'\}_{t\in\mathbb{Z}}$ with $P(|x_t| < \infty) = 1$, we have by Theorem V.2.1 that $\{x_t\}_{t\in\mathbb{Z}}$ is stationary and ergodic. Theorem V.4.2 can be used to find conditions for finite moments of x_t. For instance $\mathbb{E}[\sigma_t^2] = \mathbb{E}[x_t^2] < \infty$ if and only if $\alpha + \beta < 1$.

Maximum likelihood estimation is more involved than for ARCH(1) as σ_t^2 is unobservable even if the true values of the model parameters were known. To see this, under the stationarity condition above,

$$\sigma_t^2 = \sum_{i=0}^{\infty} \beta^i(\omega + \alpha x_{t-1-i}^2),$$

such that the conditional variance σ_t^2 depends on returns from the infinite past. In practice, one only has the observations $\{x_t\}_{t=0}^T$ at hand. Consequently, it is customary to let the initial conditional variance for the log-likelihood function be fixed, that is $\sigma_0^2(\theta) := c > 0$ is constant. Then the log-likelihood function $Q_T(\theta)$ is given as in (VII.4) with

$$\sigma_t^2(\theta) = \omega + \alpha x_{t-1}^2 + \beta\sigma_{t-1}^2(\theta), \quad t = 1, \ldots, T.$$

The asymptotic theory for the ML estimator becomes additionally tedious due to (i) the introduction of the fixed initial value c in the log-likelihood function, and (ii) the recursive structure of $\sigma_t^2(\theta)$. To see the latter, note that the first derivative (say, with respect to α) of the log-likelihood contribution involves the derivative

$$\frac{\partial\sigma_t^2(\theta)}{\partial\theta} = x_{t-1}^2 + \beta\frac{\partial\sigma_{t-1}^2(\theta)}{\partial\theta}, \quad t \geq 0.$$

We refer to Francq and Zakoïan (2019, Chapter 7) for detailed arguments.

VII.2.3 Nonlinear models and explanatory covariates

Another strand of extensions of the ARCH and GARCH models consider non-linear dynamics for the conditional variance σ_t^2. The most prominent example is the Glosten-Jagannathan-Runkle (GJR) GARCH model given by

$$\begin{aligned}
&x_t = \sigma_t z_t, \quad t \in \mathbb{Z},\\
&\sigma_t^2 = \omega + \alpha x_{t-1}^2 + \gamma\mathbb{I}(x_{t-1} < 0)x_{t-1}^2 + \beta\sigma_{t-1}^2,
\end{aligned}$$

with $\omega > 0$, $\alpha, \gamma, \beta \geq 0$, and

$$\mathbb{I}(x_{t-1} < 0) = \begin{cases} 1 & \text{if } x_{t-1} < 0, \\ 0 & \text{if } x_{t-1} \geq 0. \end{cases}$$

The term $\gamma\mathbb{I}(x_{t-1} < 0)x_{t-1}^2$ allows for the possibility that a negative shock/return at time $t-1$ has a larger impact on the conditional variance at time t than a positive return, which is referred to as a so-called leverage effect; see Glosten et al. (1993, Section II.B) for further economic motivation.

Noting that $\mathbb{I}(x_{t-1} < 0) = \mathbb{I}(z_{t-1} < 0)$, we have that σ_t^2 obeys the SRE

$$\sigma_t^2 = \omega + [\alpha z_{t-1}^2 + \gamma\mathbb{I}(z_{t-1} < 0)z_{t-1}^2 + \beta]\sigma_{t-1}^2,$$

and conditions for stationarity, ergodicity and finite moments can be derived from Theorem V.4.1 and V.4.2.

A last notable extension of the univariate ARCH and GARCH models is the inclusion of explanatory covariates in the conditional variance equation,

$$\sigma_t^2 = \sigma_t^2 = \omega + \alpha x_{t-1}^2 + \beta\sigma_{t-1}^2 + \tau y_{t-1},$$

where $\tau \geq 0$ and y_{t-1} is a non-negative random variable. This class of GARCH models is referred to as GARCH-X. Examples of y_{t-1} include credit spreads, macroeconomic variables, unexpected shocks from other assets or markets, as well as alternative volatility measures. Since the dynamics of y_{t-1} are not directly specified, its inclusion in the model requires careful considerations about the dependence structure between y_{t-1} and (x_{t-1}, z_t). We refer to Han and Kristensen (2014), Francq and Thieu (2019) and Pedersen and Rahbek (2019) for many more details.

VII.2.4 Alternative distributions for z_t

For the ARCH(1) in (VII.1)-(VII.2), instead of having a standard normal distribution, z_t could be assumed to have a more heavy tailed distribution, which may for instance be desirable (and empirically relevant) in terms of risk quantification; see Section VII.5 below. Specifically, assume that $\{z_t\}_{t\in\mathbb{Z}}$ is i.i.d. with z_t scaled Student's t-distributed with $\nu > 2$ degrees of freedom, as considered in Example I.4.3 in Part I. It holds that $\mathbb{E}[z_t] = 0$ and $\mathbb{E}[z_t^2] = 1$, and z_t has pdf given by

$$f(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)/\Gamma\left(\frac{\nu}{2}\right)}{\sqrt{(\nu-2)\,\pi}}\left(1 + \frac{x^2}{(\nu-2)}\right)^{-\left(\frac{\nu+1}{2}\right)}, \quad x \in \mathbb{R}.$$

In this case, consider the SRE for x_t^2,

$$x_t^2 = \sigma_t^2 z_t^2 = \underbrace{\alpha z_t^2}_{A_t} x_{t-1}^2 + \underbrace{\alpha z_t^2}_{B_t},$$

and conditions stationarity and ergodicity are derived along the arguments in Section VII.1.3. In particular, $\{x_t\}_{t\in\mathbb{Z}}$ is stationary and ergodic, if $\mathbb{E}[\log(A_t)] < \infty$, or equivalently

$$\alpha < \exp\left(-\mathbb{E}[\log(z_t^2)]\right).$$

Here the quantity

$$\mathbb{E}[\log(z_t^2)] = \frac{\Gamma\left(\frac{\nu+1}{2}\right)/\Gamma\left(\frac{\nu}{2}\right)}{\sqrt{(\nu-2)\,\pi}} \int_{-\infty}^{\infty} \log(x^2)\left(1+\frac{x^2}{(\nu-2)}\right)^{-\left(\frac{\nu+1}{2}\right)} dx$$

depends on the degrees of freedom $\nu > 2$. For instance, $\mathbb{E}[\log(z_t^2)] = -2$ for $\nu = 3$, such that the stationarity condition becomes $\alpha < \exp(2) = 7.38\ldots$, which is milder than the condition stated in Corollary VII.1.1 for the case of $z_t \overset{D}{=} N(0,1)$.

Assuming that z_t is scaled Student's t-distributed, introduces the additional parameter ν that one would typically estimate, such that the parameter vector is given by $\theta = (\omega, \alpha, \nu)'$. From Example I.4.3, we note that x_t has conditional density given by

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

and the log-likelihood function is given by

$$Q_T(\theta) = \frac{1}{T}\sum_{t=1}^{T}\log\left\{\frac{\Gamma\left(\frac{\nu+1}{2}\right)/\Gamma\left(\frac{\nu}{2}\right)}{\sqrt{(\omega+\alpha x_{t-1}^2)\,(\nu-2)\,\pi}}\left(1+\frac{x_t^2}{(\omega+\alpha x_{t-1}^2)\,(\nu-2)}\right)^{-\left(\frac{\nu+1}{2}\right)}\right\}.$$

The ML estimator is found by maximizing $Q_T(\theta)$ over the parameter space $\Theta \subset (0,\infty)\times[0,\infty)\times(2,\infty)$. Limit theory for the ML estimator may be derived from Theorems VI.2.2 and VI.3.1 in Chapter VI. We refer to Tsay (2010, Chapter 3) for alternative distributions for z_t, including the so-called skewed scaled Student's t-distribution that allows for the possibility that z_t has an asymmetric heavy-tailed distribution.

VII.3 Multivariate GARCH

In Chapter II we considered the VAR models as multivariate extensions of the AR models. Likewise, we consider here multivariate extensions of the ARCH and GARCH which we label Multivariate GARCH (MGARCH). Let $X_t = (X_{t,1},\ldots,X_{t,d})' \in \mathbb{R}^d$ be a vector that (for instance) contains returns of $d \geq 1$ different assets (e.g., different stocks, stock indices, currencies, commodities etc.). A simple MGARCH process – corresponding to a d-dimensional ARCH(1) – is given by

$$X_t = \Omega_t^{1/2} Z_t, \quad t \in \mathbb{Z}, \tag{VII.10}$$

$$\Omega_t^{1/2} (\Omega_t^{1/2})' = \Omega_t = g(X_{t-1}), \tag{VII.11}$$

for some measurable matrix function g, satisfying that $g(x)$ is positive definite for all $x \in \mathbb{R}^d$. Moreover

$$\{Z_t\}_{t \in \mathbb{Z}} \text{ is an } i.i.d. \text{ process with } Z_t \overset{D}{=} N(0, I_d), \tag{VII.12}$$

and Z_t and X_{t-1} are independent for all t. Parallel to the univariate ARCH(1), we have that

$$X_t | X_{t-1} \overset{D}{=} N(0, \Omega_t),$$

such that X_t has conditional density given by

$$f(X_t | X_{t-1}) = \frac{1}{\sqrt{(2\pi)^d \det(\Omega_t)}} \exp\left(-\frac{1}{2} X_t' \Omega_t^{-1} X_t \right).$$

The positive definiteness of Ω_t is (for most practical purposes) a minimal requirement for a covariance matrix, parallel to the positivity of σ_t required for univariate models. The matrix square-root $\Omega_t^{1/2}$ of Ω_t is not unique. For instance, $\Omega_t^{1/2}$ could be lower triangular stemming from a Cholesky decomposition of Ω_t, or $\Omega_t^{1/2}$ could be symmetric, obtained via an eigendecomposition of Ω_t. In the following sections we provide examples of specifications of Ω_t.

In terms of estimation, the conditional covariance matrix is assumed to be parametrized by a vector $\theta \in \mathbb{R}^k$ such that $\Omega_t(\theta) = g(X_{t-1}; \theta)$ with g known. The log-likelihood function is then given via the conditional density of X_t, and the properties of the ML estimator may be derived from the results i Chapter VI. Typically, due to the multivariate nature of MGARCH, derivations are more cumbersome compared to the univariate case. We refer to Francq and Zakoïan (2019, Chapter 10.4) for general (high-level) conditions ensuring consistency and asymptotic normality of the ML estimator for MGARCH models. Moreover, as will be clear from the following examples, the MGARCH models may contain many parameters such that k is large. In addition, the parametrization of the MGARCH model should ensure that Ω_t is positive definite. In practice these challenges imply that numerical maximization of the log-likelihood with respect to all parameters simultaneously is tedious (if not infeasible). This has given rise to alternative estimation methods, where estimation is carried out in multiple steps; see e.g. Noureldin et al. (2014), Pedersen and Rahbek (2014) and Francq and Zakoïan (2016).

In the following sections we present two classical MGARCH specifications, namely the so-called Baba-Engle-Kraft-Kroner (BEKK) and the constant conditional correlation (CCC) model. MGARCH models are widely used in empirical work, and we emphasize that several other relevant specifications exist, including dynamic conditional correlation (DCC) models (Engle, 2002), stochastic correlation models (Pelletier, 2006), orthogonal models (see Hetland et al., 2023, and the references therein) and so-called score-driven models (see D'Innocenzo and Lucas, 2024, and the references therein).

VII.3.1 BEKK

As already considered in Example V.3.4 in Chapter V, with $\{X_t\}_{t\in\mathbb{Z}}$ given by (VII.10)-(VII.12), Engle and Kroner (1995) considered the so-called BEKK specification for Ω_t given by

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

for some positive definite matrix Ω and square matrix A. By construction, this specification ensures that Ω_t is positive definite.

Example VII.3.1 *Let $d = 2$ with*

$$\Omega = \begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{12} & \Omega_{22} \end{pmatrix}, \quad \textit{and} \quad A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}.$$

Then the conditional covariance matrix is given by

$$\begin{aligned}\Omega_t &= \begin{pmatrix} \Omega_{t,11} & \Omega_{t,12} \\ \Omega_{t,12} & \Omega_{t,22} \end{pmatrix} \\ &= \begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{12} & \Omega_{22} \end{pmatrix} + \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} X_{t,1}^2 & X_{t,1}X_{t,2} \\ X_{t,1}X_{t,2} & X_{t,1}^2 \end{pmatrix} \begin{pmatrix} A_{11} & A_{21} \\ A_{12} & A_{22} \end{pmatrix}.\end{aligned}$$

with

$$\begin{aligned}\Omega_{t,11} &= \Omega_{11} + A_{11}^2X_{t,1}^2 + 2X_{t,2}A_{11}A_{12}X_{t,1} + A_{12}^2X_{t,1}^2, \\ \Omega_{t,22} &= \Omega_{11} + A_{21}^2X_{t,1}^2 + 2X_{t,2}A_{21}A_{22}X_{t,1} + A_{22}^2X_{t,1}^2, \\ \Omega_{t,12} &= \Omega_{12} + A_{11}A_{21}X_{t,1}^2 + A_{12}A_{22}X_{t,1}^2 + A_{11}A_{22}X_{t,1}X_{t,2} + A_{12}A_{21}X_{t,1}X_{t,2}.\end{aligned}$$

Recall from Example V.3.4 that X_t has an SRE representation given by

$$X_t = m_tAX_{t-1} + B_t,$$

with $\{(m_t, B_t')\}_{t\in\mathbb{Z}}$ an i.i.d. process with $(m_t, B_t')'$ a $(d+1)$-dimensional random vector satisfying

$$\begin{pmatrix} m_t \\ B_t \end{pmatrix} \overset{D}{=} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & \Omega \end{bmatrix}\right).$$

From Example V.5.2 we have that $\{X_t\}_{t\in\mathbb{Z}}$ is strictly stationary and ergodic if

$$\rho(A) < \sqrt{3.56\ldots},$$

with $\rho(A)$ denoting the spectral radius of A. Furthermore, applications of the drift criterion (see Pedersen and Rahbek, 2014, Appendix C) give that

$$\begin{aligned}
\mathbb{E}[\|X_t\|^2] &< \infty \quad \text{if } \rho(A) < 1, \\
\mathbb{E}[\|X_t\|^4] &< \infty \quad \text{if } \rho(A) < 1/3^{1/4} = 0.75\ldots, \\
\mathbb{E}[\|X_t\|^6] &< \infty \quad \text{if } \rho(A) < 1/15^{1/6} = 0.63\ldots, \\
\mathbb{E}[\|X_t\|^8] &< \infty \quad \text{if } \rho(A) < 1/105^{1/8} = 0.55\ldots,
\end{aligned}$$

parallel to the results for the tail index for univariate ARCH(1) processes in Example V.4.6. We refer to Matsui and Pedersen (2022) for additional results for BEKK processes.

The BEKK process considered above can be extended to a GARCH version given by

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

for a square matrix B. The properties of the resulting BEKK GARCH process $\{X_t\}_{t\in\mathbb{Z}}$ are complicated to derive, and the results for SREs are not directly applicable. The process $\{(X_t, \Omega_t)\}_{t\in\mathbb{Z}}$ is a Markov chain, and by carefully taking into account that Ω_t belongs to the space of positive definite matrices, Boussama et al. (2011) used an extended version of the drift criterion to prove that $\{X_t\}_{t\in\mathbb{Z}}$ is stationary and ergodic with $\mathbb{E}[\|X_t\|^2] < \infty$ provided that $\rho(A\otimes A + B\otimes B) < 1$, where $\otimes$ denotes the Kronecker product. This condition is analogous to the condition $\alpha + \beta < 1$ in Section VII.2.2 that ensures stationarity, ergodicity and finite second moment of x_t, when x_t is given by a univariate GARCH process.

ML estimation of BEKK models is considered in the works by Hafner and Preminger (2009) and Avarucci et al. (2012) and references therein.

VII.3.2 CCC

Another much applied specification for Ω_t is the so-called constant conditional correlation (CCC) model of Bollerslev (1990) and Jeantheau (1998).

Here, with $\{X_t\}_{t\in\mathbb{Z}}$ given by (VII.10)-(VII.12),

$$\Omega_t^{1/2} = D_t R^{1/2},$$

where $R^{1/2}$ is a lower triangular matrix arising from a Cholesky decomposition of a correlation matrix R, and D_t is a diagonal matrix with *positive* diagonal elements given by $\sqrt{h_{t,1}}, \ldots, \sqrt{h_{t,d}}$ with

$$h_t = (h_{t,1}, \ldots, h_{t,d})' = \omega + A\,(X_{t-1} \odot X_{t-1})\,. \tag{VII.13}$$

Here ω is a d-dimensional vector with strictly positive entries, A is a $(d \times d)$ matrix, and $\odot$ denotes element-wise multiplication of vectors or matrices of same dimensions, that is,

$$(X_t \odot X_t) = (X_{t,1}^2, \ldots, X_{t,d}^2)'.$$

We note that for $R = I_d$ and A diagonal with non-negative entries, $\{X_t\}_{t\in\mathbb{Z}}$ simply stacks d independent univariate ARCH(1) processes.

Example VII.3.2 *Let $d = 2$ such that the correlation matrix*

$$R = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix},$$

with correlation $\rho \in (-1, 1)$ and

$$R^{1/2} = \begin{pmatrix} 1 & 0 \\ \rho & \sqrt{1-\rho^2} \end{pmatrix}.$$

With $Z_t = (Z_{t,1}, Z_{t,2})'$, we have that

$$\begin{pmatrix} X_{t,1} \\ X_{t,2} \end{pmatrix} = \begin{pmatrix} \sqrt{h_{t,1}} Z_{t,1} \\ \sqrt{h_{t,1}}\rho Z_{t,1} + \sqrt{h_{t,1}}\sqrt{1-\rho^2} Z_{t,2} \end{pmatrix},$$

with conditional covariance matrix

$$\Omega_t = \begin{pmatrix} h_{t,1} & \rho\sqrt{h_{t,1}h_{t,2}} \\ \rho\sqrt{h_{t,1}h_{t,2}} & h_{t,2} \end{pmatrix}.$$

With

$$\omega = \begin{pmatrix} \omega_1 \\ \omega_2 \end{pmatrix} \quad \textit{and} \quad A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix},$$

it holds that

$$h_t = \begin{pmatrix} h_{t,1} \\ h_{t,2} \end{pmatrix} = \begin{pmatrix} \omega_1 + A_{11} X_{t-1,1}^2 + A_{12} X_{t-1,2}^2 \\ \omega_2 + A_{21} X_{t-1,1}^2 + A_{22} X_{t-1,2}^2 \end{pmatrix}.$$

It holds that X_t has conditional covariance matrix (almost surely)

$$\begin{aligned}\mathbb{V}[X_t|X_{t-1}] &= \mathbb{E}[D_t R^{1/2} Z_t Z_t'(D_t R^{1/2})'|X_{t-1}] \\ &= D_t R^{1/2}\mathbb{E}[Z_t Z_t'|X_{t-1}](R^{1/2})' D_t \\ &= D_t R^{1/2}\mathbb{E}[Z_t Z_t'](R^{1/2})' D_t \\ &= D_t R^{1/2} I_d (R^{1/2})' D_t \\ &= D_t R D_t \\ &= \Omega_t.\end{aligned}$$

In particular, (almost surely)

$$\mathbb{V}(X_{t,i}|X_{t-1}) = h_{t,i}, \quad i = 1, \dots, d.$$

To ensure that all conditional variances $h_{t,i}$ are positive, all entries of the matrix A are typically assumed to be non-negative (see Conrad and Karanasos, 2010, for additional considerations).

Since R is a correlation matrix, the diagonal of D_t^2 is the diagonal of Ω_t, such that the *conditional* correlation matrix of X_t (given X_{t-1}) is

$$D_t^{-1}\Omega_t D_t^{-1} = R.$$

In terms of stationarity and ergodicity, let

$$\tilde{Z}_t = (\tilde{Z}_{t,1}, \dots, \tilde{Z}_{t,d})' := R^{1/2} Z_t,$$

such that $\{\tilde{Z}_t\}_{t\in\mathbb{Z}}$ is an i.i.d. process with $\tilde{Z}_t \overset{D}{=} N(0, R)$. Then, using that D_t^2 is a diagonal matrix with diagonal given by h_t,

$$X_t \odot X_t = (D_t\tilde{Z}_t) \odot (D_t\tilde{Z}_t) = \begin{pmatrix} \tilde{Z}_{t,1}^2 h_{t,1} \\ \vdots \\ \tilde{Z}_{t,d}^2 h_{t,d} \end{pmatrix} = K_t h_t.$$

with K_t a diagonal matrix with diagonal given by $(\tilde{Z}_t \odot \tilde{Z}_t)$. Using (VII.13), we have that $(X_t \odot X_t)$ obeys the SRE

$$(X_t \odot X_t) = \underbrace{K_t\omega}_{=B_t} + \underbrace{K_t A}_{=A_t}(X_{t-1} \odot X_{t-1}),$$

and conditions for stationarity and ergodicity can be derived using Theorem V.5.1. Sufficient conditions for finite moments of $\|X_t\|$ are given in Pedersen

(2017). We refer to Pedersen (2016) and Damek et al. (2019) for considerations about the tail index of the unconditional distributions of $\|X_t\|$ and $|X_{t,i}|$.

Similar to the BEKK model above, a GARCH extension of the CCC model is given by

$$h_t = \omega + A\left(X_{t-1} \odot X_{t-1}\right) + Bh_{t-1}.$$

Asymptotic theory for the ML estimator for CCC GARCH models is provided in Francq and Zakoïan (2012) and Pedersen (2017).

VII.4 (*) Application: Portfolio Choice

Consider an investor who can invest in a *risky asset* and a *risk-free asset* with returns x_{t+1} and $x_{t+1}^{(f)}$, respectively, from time t to $t+1$. The risk-free asset is risk-free in the sense that, given some information set $\mathcal{F}_t$, $x_{t+1}^{(f)} \in \mathcal{F}_t$. The objective of the investor at time t is to decide how much to invest in the risky asset relative to the risk-free asset. Specifically, let $w_t \in \mathcal{F}_t$ denote the weight put on the risky asset and $(1-w_t)$ the weight on the risk-free asset, that is, at time t $(w_t \times 100)\%$ of the money is invested in the risky asset. Then the one-period return of the portfolio, $x_{t+1}^{(p)}$, is given by

$$\begin{aligned} x_{t+1}^{(p)} &= w_t x_{t+1} + (1-w_t)x_{t+1}^{(f)} \\ &= w_t \tilde{x}_{t+1} + x_{t+1}^{(f)}, \end{aligned}$$

with $\tilde{x}_{t+1} = x_{t+1} - x_{t+1}^{(f)}$ denoting the *excess return* of the risky asset. The choice of ω_t depends on the investor's objective. For instance, $w_t = 0$ yields a deterministic return equal to the risk-free rate. This is desirable, if the investor is completely risk averse. In a more general setting, the investor may seek to balance (expected) reward and risk. The reward may be measured in terms of the expected portfolio return, $\mathbb{E}[x_{t+1}^{(p)}|\mathcal{F}_t]$, and the risk may be measured in terms of the conditional variance, $\mathbb{V}[x_{t+1}^{(p)}|\mathcal{F}_t]$ (or some other risk measure, such as Value-at-Risk or Expected Shortfall considered in Section VII.5 below). Consequently, and following e.g. Neely et al. (2014, Section 4), we suppose that the investor seeks to maximize the following utility function at time t,

$$\mathrm{U}_t = \mathbb{E}[x_{t+1}^{(p)}|\mathcal{F}_t] - \frac{\gamma}{2}\mathbb{V}[x_{t+1}^{(p)}|\mathcal{F}_t],$$

where $\gamma > 0$ is a risk aversion parameter. Note that a higher risk aversion γ, the lower utility for a given level of risk.

Using that $w_t, x_{t+1}^{(f)} \in \mathcal{F}_t$

$$\mathbb{E}[x_{t+1}^{(p)}|\mathcal{F}_t] = \mathbb{E}[w_t\tilde{x}_{t+1} + x_{t+1}^{(f)}|\mathcal{F}_t] = w_t\mathbb{E}[\tilde{x}_{t+1}|\mathcal{F}_t] + x_{t+1}^{(f)},$$

and

$$\mathbb{V}[x_{t+1}^{(p)}|\mathcal{F}_t] = w_t^2\mathbb{V}[\tilde{x}_{t+1}|\mathcal{F}_t],$$

so we have that

$$\mathrm{U}_t = w_t\mathbb{E}[\tilde{x}_{t+1}|\mathcal{F}_t] + x_{t+1}^{(f)} - \frac{\gamma}{2}w_t^2\mathbb{V}[\tilde{x}_{t+1}|\mathcal{F}_t].$$

Maximizing U_t with respect to w_t gives the first-order condition

$$\frac{\partial \mathrm{U}_t}{\partial w_t} = \mathbb{E}[\tilde{x}_{t+1}|\mathcal{F}_t] - \gamma w_t \mathbb{V}[\tilde{x}_{t+1}|\mathcal{F}_t] = 0,$$

with solution

$$\begin{aligned} w_t^* &= \frac{1}{\gamma}\left(\frac{\mathbb{E}[\tilde{x}_{t+1}|\mathcal{F}_t]}{\mathbb{V}[\tilde{x}_{t+1}|\mathcal{F}_t]}\right) \\ &= \frac{1}{\gamma}\left(\frac{\mathbb{E}[x_{t+1}|\mathcal{F}_t] - x_{t+1}^{(f)}}{\mathbb{V}[x_{t+1}|\mathcal{F}_t]}\right). \end{aligned}$$

The optimal weight w_t^* depends on the conditional mean and conditional variance of the risky asset. These two quantities may be specified in terms of econometric models of the type (VII.9).

Note that the above considerations can be extended to a multivariate setting where the investor can invest in d different risky assets (e.g., different stock indices, sectors, different asset classes), with returns given by the $(d \times 1)$ vector X_{t+1}. In this case, the weights on the risky assets are given by the $(d \times 1)$ vector w_t and the risk-free asset is given weight $1 - w_t'\iota_d$, where ι_d is a $(d \times 1)$ vector of ones. Then with $\tilde{X}_{t+1} = X_{t+1} - \iota_d x_{t+1}^{(f)}$ the $(d \times 1)$ vector of excess returns of the risky assets, the utility is given by

$$\mathrm{U}_t = w_t'\mathbb{E}[\tilde{X}_{t+1}|\mathcal{F}_t] + x_{t+1}^{(f)} - \frac{\gamma}{2} w_t'\mathbb{V}[\tilde{X}_{t+1}|\mathcal{F}_t] w_t,$$

where $\mathbb{E}[\tilde{X}_{t+1}|\mathcal{F}_t]$ is the vector of conditional mean excess returns and $\mathbb{V}[\tilde{X}_{t+1}|\mathcal{F}_t]$ the conditional covariance matrix of the excess returns. The vector of optimal weights are then given by

$$w_t^* = \frac{1}{\gamma}\left(\mathbb{V}[\tilde{X}_{t+1}|\mathcal{F}_t]\right)^{-1}\mathbb{E}[\tilde{X}_{t+1}|\mathcal{F}_t] = \frac{1}{\gamma}\left(\mathbb{V}[X_{t+1}|\mathcal{F}_t]\right)^{-1}\left(\mathbb{E}[X_{t+1}|\mathcal{F}_t] - \iota_d x_{t+1}^{(f)}\right),$$

provided that $\mathbb{V}[X_{t+1}|\mathcal{F}_t] = \Omega_{t+1}$ is invertible. Determining the optimal weight typically requires modelling the joint dynamics of the risky asset returns, which may be done in terms of VAR and/or multivariate GARCH models.

VII.5 (*) Application: Value-at-Risk (VaR) and beyond

An important measure for quantifying risk is the so-called *Value-at-Risk* (VaR) risk measure. VaR is important, for instance, from a regulatory perspective, as many financial institutions are obliged to disclose their estimates of such risk measures in relation to their holdings of risky assets (see, e.g., the Basel Committee on Banking Supervision). Importantly, a bank may be forced to set aside additional capital if their actual losses exceed their estimated VaR. We here discuss how VaR can be computed if the returns (or losses) are determined from an ARCH process. We emphasize that the computation of VaR boils down to computing a quantile of some (conditional) distribution. For the ARCH processes, 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 VaR is addressed.

VII.5.1 VaR for ARCH processes

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

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

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

$$\mathbb{P}(x_{t+1} < -\mathrm{VaR}_t^\kappa | \mathcal{F}_t) = \kappa, \quad \mathrm{VaR}_t^\kappa \in \mathcal{F}_t,$$

where $\mathcal{F}_t$ denotes some information set available at time t (e.g. the series of previous returns). Note that $\mathbb{P}(-x_{t+1} \leq \mathrm{VaR}_t^\kappa | \mathcal{F}_t) = 1-\kappa$, 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

[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 $\mathrm{VaR}_t^\kappa = \inf\{y \in \mathbb{R} : P(-x_{t+1} \leq y | \mathcal{I}_t) \geq 1-\kappa\}$. 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 $\max[0, \inf\{y \in \mathbb{R} : P(-x_{t+1} \leq y | \mathcal{I}_t) \geq 1-\kappa\}]$.

information set $\mathcal{F}_t$, as well as the *confidence level* $1-\kappa$. Typical values of κ in applications are 1%, 2.5%, and 5%. Throughout, we assume that the information set contains only past values of the returns, i.e. $\mathcal{F}_t = \{x_i : i \leq t\}$. 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.

Example VII.5.1 (Gaussian returns) *Suppose that* $\{x_t\}_{t\in\mathbb{Z}}$ *is an i.i.d. process with* $x_t \overset{D}{=} N(0,1)$. *Then, with* $\Phi(\cdot)$ *the cdf of the* $N(0,1)$ *distribution and using that* x_{t+1} *is independent of* $\mathcal{F}_t$,

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

Hence,

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

i.e. the VaR *is (negative) the* κ *percentile of the standard normal distribution. Likewise, if instead* $x_t \overset{D}{=} N(\mu, \sigma^2)$,

$$\mathrm{VaR}_t^\kappa = -\mu - \sigma\Phi^{-1}(\kappa).$$

Example VII.5.2 (ARCH returns) *Suppose that the returns are given by the stationary ARCH(1) process*

$$\begin{aligned} x_t &= \sigma_t z_t, \quad t \in \mathbb{Z}, \\ \sigma_t^2 &= \omega + \alpha x_{t-1}^2, \end{aligned}$$

with $\{z_t\}_{t\in\mathbb{Z}}$ *an i.i.d. process with* $z_t \overset{D}{=} N(0,1)$ *and* $\omega > 0$, $\alpha \geq 0$. *Then using that* $\sigma_{t+1}^2 \in \mathcal{F}_t$ *and* z_{t+1} *and* $\mathcal{F}_t$ *are independent,*

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

Hence,

$$\mathrm{VaR}_t^\kappa = -\sigma_{t+1}\Phi^{-1}(\kappa). \tag{VII.14}$$

Note that the above considerations apply to any $0 < \sigma_{t+1} \in \mathcal{F}_t$, *e.g. GARCH(1,1) processes.*

The above definition of 1-period VaR easily extends to the h-period VaR, $\mathrm{VaR}_{t,h}^\kappa$, given by

$$\mathbb{P}\left(x_{t+1,h} < -\mathrm{VaR}_{t,h}^\kappa|\mathcal{F}_t\right) = \kappa. \tag{VII.15}$$

Example VII.5.3 (Gaussian returns, ctd.) *Proceeding with the case of i.i.d. returns with* $x_t \overset{D}{=} N(\mu, \sigma^2)$, *it follows that* $x_{t+1,h} = \sum_{i=1}^{h} x_{t+i} \overset{D}{=} h\mu + \sigma\sqrt{h}z$ *where* $z \sim N(0,1)$ *and independent of* $\mathcal{F}_t$. *Hence,*

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

such that

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

Example VII.5.4 (ARCH returns, ctd.) *For the ARCH(1) we have* $x_{t+1} = \sqrt{\omega + \alpha x_t^2} z_{t+1}$, *with* $z_{t+1} \overset{D}{=} N(0,1)$. *Since the factor* $\sqrt{\omega + \alpha x_t^2} \in \mathcal{F}_t$, *and* z_{t+1} *is independent of* $\mathcal{F}_t$, *it holds that* $x_{t+1}|\mathcal{F}_t \overset{D}{=} N(0, \omega + \alpha x_t^2)$ *which we exploited in Example VII.5.2 to find the 1-period* VaR. *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* $\mathcal{F}_t$ *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, $x_{t+2}|\mathcal{F}_t \overset{D}{=} N(0, \omega + \alpha x_{t+1}^2)$, *but the conditional distribution of* x_{t+2} *(given* $\mathcal{F}_t$*) is non-Gaussian, since the factor* $\sqrt{\omega + \alpha(\omega + \alpha x_t^2) z_{t+1}^2}$ *does not belong to the information set* $\mathcal{F}_t$. *In fact, it can be shown that conditional distribution of* x_{t+2} *is a so-called normal variance mixture whenever* $\alpha > 0$ *(see, e.g., the recent work of Abadir et al., 2023). Consequently, the conditional distribution of* $x_{t+1,2}$ *is a sum of dependent normal variance mixture random variables, and may be viewed as effectively intractable. In such a case it is customary to quantify (or approximate) the* VaR *using the algorithm stated below.*

Algorithm VII.5.1 *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 (VII.15), obtain the approximate* VaR *as the* $(1-\kappa)$ *empirical quantile of the simulated losses, i.e.*

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

where $\lfloor y \rfloor$ *denotes the integer part of* $y \in \mathbb{R}$.

VII.5.2 VaR **Inference**

In practice, the VaR depends on unknown parameters, e.g., $\theta = (\mu, \sigma^2)'$ in Example VII.5.1 and $\theta = (\omega, \alpha)'$ in Example VII.5.2. As already considered, these parameters may be estimated by ML (or other estimation methods), leading to estimators of VaR. Provided that the ML estimators for θ are consistent and asymptotically normal (cf., Theorems VI.2.2 and VI.3.1), we likewise have that the VaR estimators are consistent and asymptotically normal.

Example VII.5.5 (Gaussian returns, ctd.) *Consider Example VII.5.1 with* $x_t \overset{D}{=} N(0, \sigma^2)$. *Given a set of observations* $\{x_t\}_{t=1,\ldots,T}$, *the ML estimator for* σ^2 *is given by*

$$\hat{\sigma}_T^2 = \frac{1}{T} \sum_{t=1}^{T} x_t^2,$$

and an estimator for the h-period VaR *is given by*

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

Trivially, the i.i.d. DGP $\{x_t\}_{t\in\mathbb{Z}}$ *is stationary and ergodic with* $\mathbb{E}[x_t^2] < \infty$. *By Theorem V.2.2* $\hat{\sigma}_T^2 \xrightarrow{p} \sigma_0^2$ *as* $T \to \infty$, *and hence* $\widehat{\mathrm{VaR}_{t,h}^{\kappa}}$ *is consistent for* $\mathrm{VaR}_{t,h}^{\kappa}$, *that is*

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

Likewise, noting that $\mathbb{E}[x_t^2 - \sigma_0^2 | \mathcal{F}_t] = \mathbb{E}[x_t^2 - \sigma_0^2] = 0$ *almost surely, and* $\Sigma := \mathbb{E}[(x_t^2 - \sigma_0^2)^2] = 2\sigma_0^4 < \infty$, *we have by Theorem V.2.3,* $\sqrt{T}(\hat{\sigma}_T^2 - \sigma_0^2) \xrightarrow{D} N(0, \Sigma)$. *Moreover, since* $x \mapsto \sqrt{x}$ *is continuously differentiable on* *the positive real axis, we have that* $\sqrt{T}(\hat{\sigma}_T - \sigma_0) \overset{D}{\rightarrow} N(0, \sigma_0^2/2)$ *[by the* Δ*-method]. Hence,*

$$\sqrt{T}\left(\widehat{\mathrm{VaR}_{t,h}^{\kappa}} - \mathrm{VaR}_{t,h}^{\kappa}\right) = -\sqrt{h}\Phi^{-1}(\kappa)\sqrt{T}(\hat{\sigma}_T - \sigma_0) \overset{D}{\rightarrow} N(0, h\sigma_0^2\Phi^{-1}(\kappa)^2/2).$$

Hence, we may apply the approximation

$$\widehat{\mathrm{VaR}_{t,h}^{\kappa}} \overset{D}{\approx} N(\mathrm{VaR}_{t,h}^{\kappa}, h\sigma_0^2\Phi^{-1}(\kappa)^2/(2T)),$$

and one may report, for instance, 95% error bands of the VaR *as* $\widehat{\mathrm{VaR}_{t,h}^{\kappa}} \pm \Phi^{-1}(0.975)\sqrt{h/(2T)}|\Phi^{-1}(\kappa)|\hat{\sigma}_T$*. 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".*

Example VII.5.6 (ARCH returns, ctd.) *From Part VI, we can estimate the parameters* $\theta = (\omega, \alpha)'$ *by ML estimation* $\hat{\theta}_T = (\hat{\omega}_T, \hat{\alpha}_T)'$*. Assume that the DGP satisfies* $\alpha_0 \in (0, 1)$*. Then from Example VI.2.5 and Section VI.3.1, respectively,*

$$\hat{\theta}_T \overset{p}{\rightarrow} \theta_0 \quad \textit{and} \quad \sqrt{T}(\hat{\theta}_T - \theta_0) \overset{D}{\rightarrow} N(0, -\Sigma_0^{-1}), \tag{VII.16}$$

for some positive definite matrix $-\Sigma_0$*. Based on the estimator* $\hat{\theta}_T$*, we obtain an estimator for the conditional volatility, given by*

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

and, using (VII.14), we have the VaR *estimator,*

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

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 $\hat{\sigma}_{t+1}^2 = \hat{\omega}_T + \hat{\alpha}_T x^2$ *and* $\sigma_{t+1}^2 = \omega + \alpha x^2$*. By (VII.16), a first-order Taylor expansion (up to a negligible remainder term),*

$$\hat{\sigma}_{t+1} - \sigma_{t+1} = \frac{1}{2\sqrt{\theta_0' w}} w'(\hat{\theta}_T - \theta_0),$$

with $w = (1, x^2)'$, and we conclude that

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

and

$$\sqrt{T}(\widehat{\mathrm{VaR}}_t^\kappa - \mathrm{VaR}_t^\kappa) \xrightarrow{D} N\left(0, \frac{\Phi^{-1}(\kappa)^2}{4\theta_0' w} w'(-\Sigma_0^{-1})w\right).$$

Similar to Example VII.5.5, one may use this result to construct error bands for the estimated VaR*, for instance by relying on the estimator $\hat{\Psi}_T$ for $(-\Sigma_0^{-1})$ provided in Example VI.4.2.*

VII.5.3 (*) Extensions and alternative risk measures

We end this section by providing some directions for potential extensions and additional details.

VII.5.3.1 Extending Algorithm VII.5.1

The Algorithm VII.5.1 easily extends to any finite horizon $h \geq 2$, other specifications for σ_t, as well as other distributions for z_t. Moreover, even if the distribution of z_t is assumed to be unknown one may incorporate draws from the empirical distribution of the standardized returns $\hat{z}_t = x_t/\hat{\sigma}_t$. This approach is widely used in applications, as is typically referred to as so-called *filtered historical simulation* (Barone-Adesi et al., 1999).

One may also extend the algorithm to incorporate estimation uncertainty. This may for instance be done by making draws of θ from the limiting distribution of the ML estimator in (VII.16), and compute the VaR for each draw (e.g., Blasques et al., 2016). Alternatively, one may instead draw θ from bootstrapped distributions of the ML estimator (see e.g. Beutner et al., 2024 and Cavaliere et al., 2018).

VII.5.3.2 VaR **Backtesting**

In practice, one may typically want to evaluate if a given (estimated) econometric model, such as the ARCH(1), does well in terms of quantifying VaR. One way of doing so is to test for so-called *unconditional correct coverage*, which can be viewed as a misspecification test. Given a set of observations $\{x_t\}_{t=1}^T$ and their associated VaR, $\{\mathrm{VaR}_{t-1,1}^\kappa\}_{t=1}^T$, define the *hit* sequence

$$\mathrm{Hit}_t = \begin{cases} 1 & \text{if } -x_t > \mathrm{VaR}_{t-1,1}^\kappa \\ 0 & \text{otherwise} \end{cases}, \quad t = 1, 2, \ldots, T.$$

Given correct model specification, it holds that the process $\{\mathrm{Hit}_t\}_{t=1}^T$ is i.i.d. Bernoulli(κ), such that

$$E[\mathrm{Hit}_t] = \kappa.$$

Kupiec suggested to model $\{\mathrm{Hit}_t\}_{t=1}^T$ as an *i.i.d.* Bernoulli(p) sequence, and test the hypothesis $p = \kappa$ against $p \neq \kappa$. The ML estimator for p is given by

$$\hat{p} = T^{-1} \sum_{t=1}^{T} \mathrm{Hit}_t,$$

and the LR statistic for the hypothesis is given by

$$LR_T(p = \kappa) = -2 \log \left[\frac{(1-\kappa)^{T_0} \kappa^{T_1}}{(1-\hat{p})^{T_0} \hat{p}^{T_1}} \right],$$

with $T_1 = \sum_{t=1}^T \mathrm{Hit}_t$ and $T_0 = T - T_1$. Under the hypothesis, it holds that

$$LR_T(p = \kappa) \xrightarrow{D} \chi_1^2, \quad \text{as } T \to \infty,$$

which can be proved by verifying the conditions of Theorem VI.4.1

Note that as in Example VII.5.6, the VaR is typically estimated, based on estimated parameters. Hence, in practice one has the sequence $\{\widehat{\mathrm{Hit}}_t\}_{t=1}^T$ based on estimates of $\mathrm{VaR}_{t-1,1}^{\kappa}$. Inherently, the LR statistic $LR_T(p = \kappa)$ depends on the ML estimator for the model parameters $\hat{\theta}_T = (\hat{\omega}_T, \hat{\alpha}_T)'$, and the χ_1^2 limiting distribution is potentially unreliable. This issue has been addressed by Escanciano and Olmo (2010).

Lastly, one may note that there exist various extensions of Kupiec's test, for instance allowing for alternatives where violations of the value-at-risk, that is, the events $\mathrm{Hit}_t = 1$ appear in consecutive time periods; see e.g. the much applied test by Christoffersen (1998).

VII.5.3.3 Expected Shortfall (ES)

Although VaR is widely used in practice, some researchers and practitioners has raised their concern that VaR is unreliable risk measure in certain applications. Specifically, it can be shown that VaR is a so-called incoherent risk measure. Specifically, under certain (heavy tailed) asset return distributions, it can be shown that VaR discourages risk diversification and hence may be an undesirable measure with respect to managing risk; 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 (which happens with probability κ). These concerns have 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

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

and clearly $\mathrm{ES}_t^{\kappa} \geq \mathrm{VaR}_t^{\kappa}$. If x_t follows an ARCH(1) process as in Example VII.5.2, it can be shown that

$$\mathrm{ES}_{t,1}^{\kappa} = \kappa^{-1}\sigma_{t+1}\phi(-\Phi^{-1}(\kappa)),$$

where ϕ is the pdf of the $N(0,1)$ distribution. Similar to VaR, one may address the estimation uncertainty when estimating ES. Likewise, parallel to Algorithm VII.5.1, one may have to compute multiple period ES by means of simulations.

VII.5.3.4 Alternative risk measures

Recently, risk quantification has been given wide attention in relation to so-called systemic risk. For instance Adrian and Brunnermeier (2016), see also Banulescu-Radu et al. (2021), consider VaR **co**nditional on a particular event, labelled CoVaR. As an example, let $-x_{t+1}$ denote the loss of a share of Company 1 and $-y_{t+1}$ the loss of a share of Company 2. Then the CoVaR is given by

$$\mathbb{P}(-x_{t+1} > \mathrm{CoVaR}_{t,1}^{\kappa}| - y_{t+1} > \mathrm{VaR}_t^{Y,\kappa}, \mathcal{F}_t) = \kappa,$$

where $\mathrm{VaR}_t^{Y,\kappa}$ is the VaR for Company 2. Likewise, Brownlees and Engle (2017) have considered the notion of so-called long-run marginal expected shortfall, LRMES: Let $-x_{t+1,h}$ denote the h-period loss of the share of some bank (say, Goldman Sachs), and let $-y_{t+1,h}$ denote the loss of the entire financial sector (e.g., as mention as the loss of an equity index consisting of financial companies). Then the LRMES is given by

$$\mathrm{LRMES}_{t,h}^{\kappa} = \mathbb{E}[-x_{t+1,h}| - y_{t+1,h} > 40\%],$$

where the event $-y_{t+1,h} > 40\%$ is interpreted as a major financial crisis. Note that the computation of CoVaR and LRMES requires a specification of the joint conditional distribution of (x_{t+1}, y_{t+1}), which may be done in terms of a multivariate GARCH model.

The notion of VaR may be used outside of financial risk management. For instance the International Monetary Fund (IMF), as well as several central banks, make use of the so-called Growth-at-Risk (GaR) measure that quantifies the smallest possible (that is, the worst-case scenario) GDP growth rate at a given confidence level $(1-\kappa)$; see e.g. Brownlees and Souza (2021). For a given country, let Y_t denote the GDP growth rate in, say, quarter t, then the 1-step ahead GaR at risk level κ, $\mathrm{GaR}_{t,1}^{\kappa}$, is defined by

$$\mathbb{P}(Y_{t+1} < \mathrm{GaR}_{t,1}^{\kappa} | \mathcal{F}_t) = \kappa, \quad \mathrm{GaR}_{t,1}^{\kappa} \in \mathcal{F}_t,$$

for some information set $\mathcal{F}_t$, potentially containing various macroeconomic variables and indicators.

References

Abadir, K.M., Luati, A., & Paruolo, P., 2023, "GARCH density and functional forecasts“, *Journal of Econometrics*, Vol. 235, pp. 470–483.

Adrian, T., & Brunnermeier, M.K., 2016, "CoVaR“, *American Economic Review*, Vol. 106, pp. 1705–1741.

Avarucci, M., Beutner, E., & Zaffaroni, P., 2012, "On moment conditions for quasi-maximum likelihood estimation of multivariate ARCH models“, *Econometric Theory*, Vol.29, pp. 545–566.

Banulescu-Radu, D., Hurlin, C., Leymarie, J., & Scaillet, O., 2021, "Backtesting Marginal Expected Shortfall and Related Systemic Risk Measures“, *Management Science*, Vol. 67, pp. 5730-5754.

Barone-Adesi, G., Giannopoulos, K., & Vosper, L., 1999, "VaR without correlations for portfolios of derivative securities“, *The Journal of Futures Markets*, Vol. 19, pp. 583–602.

Beutner, E., Heinemann, A., & Smeekes, S., 2024, "A residual bootstrap for conditional Value-at-Risk“, *Journal of Econometrics*, Vol. 238, 105554.

Blasques, F., Francq, C., & Laurent, S., 2024, "Autoregressive conditional betas“, *Journal of Econometrics*, Vol. 238, 105630.

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.

Bollerslev, T., 1986, "Generalized autoregressive conditional heteroskedasticity“, *Journal of Econometrics*, Vol. 31, pp. 307–327.

Bollerslev, T., 1990, "Modelling the coherence in short-run nominal exchange rates: A multivariate Generalized ARCH model Arch Model“, *Review of Economics and Statistics*, Vol. 72, pp. 498–505.

Bollerslev, T., 2010, "Glossary to ARCH (GARCH)“ in *Volatility and Time Series Econometrics: Essays in Honor of Robert Engle*, Oxford University Press.

Bougerol, P., & Picard, N., 1992, "Strict stationarity of generalized autoregressive processes“, *The Annals of Statistics*, Vol. 20, pp. 1714–1730.

Boussama, F., Fuchs, F., & Stelzer, R., 2011, "Stationarity and geometric ergodicity of BEKK multivariate GARCH models", *Stochastic Processes and their Applications*, Vol. 121, pp. 2331–2360.

Brownlees, C., & Engle, R., 2016, "SRISK: A Conditional Capital Shortfall Measure of Systemic Risk", *Review of Financial Studies, Vol. 30, pp. 48–79.*

Brownlees, C., & Souza, A.B.M., 2021, "Backtesting global Growth-at-Risk", *Journal of Monetary Economics*, Vol. 118, pp. 312–330.

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

Chorro, C., Guégan, D., & Ielpo, F., 2015, *A Time Series Approach to Option Pricing*, Springer.

Christoffersen, P., 1998, "Evaluating interval forecasts", *International Economic Review*, Vol. 39, pp. 841–862.

Conrad, C., & Karanasos, M., 2010, "Negative volatility spillovers in the unrestricted ECCC-GARCH model", *Econometric Theory*, Vol. 26, pp. 838–862.

Damek, E., Matsui, M., & Swiatkowski, W., (2019), "Componentwise different tail solutions for bivariate stochastic recurrence equations with application to GARCH (1,1) processes", *Colloquium Mathematicum*, Vol. 155, pp. 227–254.

D'Innocenzo, E., & Lucas, A., 2024, "Dynamic partial correlation models", *Journal of Econometrics*, Vol. 241, 105747.

Engle, R.F., 1982, "Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation", *Econometrica*, Vol. 50, pp. 987–1008.

Engle, R.F., 2002, "Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models", *Journal of Business & Economic Statistics*, Vol. 20, pp. 339–350.

Engle, R.F., 2016, "Dynamic conditional beta", *Journal of Financial Econometrics*, Vol. 14., pp. 643–667.

Engle, R.F., & Kroner, K.F., 1995, "Multivariate simultaneous generalized ARCH“, *Econometric Theory*, Vol. 11, pp. 122–150.

Escanciano, J.C., & Olmo, J., 2010, "Backtesting parametric Value-at-Risk with estimation risk“, *Journal of Business & Economic Statistics*, Vol. 28, pp. 36–51.

Francq, C. & Thieu, L.Q., 2019, "QML inference for volatility models with covariates“, *Econometric Theory*, Vol. 35, pp. 37–72.

Francq, C. & Zakoïan, J.-M., 2012, "QML estimation of a class of multivariate asymmetric GARCH models“, *Econometric Theory*, Vol. 28, pp. 179–206.

Francq, C. & Zakoïan, J.-M., 2016, "Estimating multivariate volatility models equation by equation“, *Journal of the Royal Statistical Society: Series B*, Vol. 78, pp. 613–635.

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

Glosten, L.R., Jagannathan, R., & Runkle, D.E., 1993, "On the relation between expected value and the volatility of the nominal excess return on stocks“, *The Journal of Finance*, Vol. XLVIII, pp. 1779–1801.

Hafner, C.M., & Preminger, A., 2009, "On asymptotic theory for multivariate GARCH models“, *Journal of Multivariate Analysis*, Vol. 100, pp. 2044–2054.

Han, H, & Kristensen, D., 2014, "Asymptotic theory for the QMLE in GARCH-X models with stationary and nonstationary covariates“, *Journal of Business & Economic Statistics*, Vol. 32, pp. 416–429.

Hetland, S., Pedersen, R.S., & Rahbek, A., 2023, "Dynamic conditional eigenvalue GARCH“, *Journal of Econometrics*, Vol. 237, 105175.

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

Jeantheau, T., 1998, "Strong consistency of estimators for multivariate ARCH models“, *Econometric Theory*, Vol. 14, pp. 70–86.

Matsui, M., & Pedersen, R.S., 2022, "Characterization of the tail behavior of a class of BEKK processes: A stochastic recurrence equation approach“, *Econometric Theory*, Vol. 38, pp. 1–34.

Mikosch, T., & Starica, C., 2000, "Limit theory for the sample autocorrelations and extremes of a GARCH(1,1) process", *The Annals of Statistics*, Vol. 28, pp. 1427–1451.

Mikosch, T., & Straumann, D., 2006, "'Stable limits of martingale transforms with application to the estimation of GARCH parameters", *The Annals of Statistics*, Vol. 34, pp. 493–522.

Neely, C.J., Rapach, D.E., Tu, J., & Zhou, G., 2014, "Forecasting the equity risk premium: The role of technical indicators", *Management Science*, Vol. 60, pp. 1772–1791.

Nielsen, H.B., & Rahbek, A., 2024, "Penalized quasi-likelihood estimation and model selection with parameters on the boundary of the parameter space", *Econometrics Journal*, Vol. 27, pp. 107–125.

Noureldin, D., Shephard, N., & Sheppard, K., 2014, "Multivariate rotated ARCH models", *Journal of Econometrics*, Vol. 179, pp. 16–30.

Pedersen, R.S., (2016), "Targeting estimation of CCC-GARCH models with infinite fourth moments", *Econometric Theory*, Vol. 32, pp. 498–531.

Pedersen, R.S., (2017), "Inference and testing on the boundary in extended constant conditional correlation GARCH models", *Journal of Econometrics*, Vol. 196, pp. 23–36.

Pedersen, R.S., & Rahbek, A., 2014, "Multivariate variance targeting in the BEKK-GARCH model", *Econometrics Journal*, Vol. 17, pp. 24–55.

Pedersen, R.S., & Rahbek, A., 2019, "Testing GARCH-X type models", *Econometric Theory*, Vol. 35, pp. 1012–1047.

Pelletier, D., 2006, "Regime switching for dynamic correlations", *Journal of Econometrics*, Vol. 131, pp. 445-473.

Taylor, S.J., 1986, *Modelling Financial Time Series*, Wiley.

Tsay, R.S., 2010, *Analysis of Financial Time Series*, 3rd edition, Wiley.