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 $\log\left(2\pi\right)$ 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,*

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

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 $\mathbb{E}[x_t^{*2}] < \infty$, it follows directly that

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

Likewise

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

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

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[Y_t^2\mathbb{I}\left(|Y_t| > \delta\sqrt{T}\right)|\mathcal{F}_{t-1}\right] &\leq \frac{1}{T^2\delta^2}\sum_{t=1}^{T}\mathbb{E}\left[Y_t^4|\mathcal{F}_{t-1}\right] \\ &= \frac{1}{T\delta^2}\left(\frac{1}{T}\sum_{t=1}^{T}x_{t-1}^4\right)\mathbb{E}\left[\varepsilon_t^4\right] \xrightarrow{p} 0.\end{aligned} \tag{II.8}$$

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, $\text{LR}(\rho = \rho_0)$:

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,

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

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

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

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 $\text{LR}(\rho = \rho_0) = -2 \log Q$, where

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

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

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

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 $\mathrm{LR}(\rho = \rho_0) = -2\log Q = T\log(1 + W_T)$.

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

$$W_T \xrightarrow{P} 0 \text{ and } TW_T \xrightarrow{D} \chi_1^2.$$

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

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

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 $\mathbb{E}[x_t^*] = \frac{\mu}{1-\rho}$, 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*

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

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

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

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

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. $N(0, \Omega)$.

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,

$$\begin{aligned} \log L(\beta, \Omega) &= \log\left(\prod_{t=1}^{T} f(Y_t|Z_t)\right) \qquad \text{(II.22)} \\ &= -\frac{Tp}{2}\log(2\pi) - \frac{T}{2}\log\det(\Omega) - \frac{1}{2}\sum_{t=1}^{T}(Y_t - \beta' Z_t)' \Omega^{-1} (Y_t - \beta' Z_t) \end{aligned}$$

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, $df(B; dB) = \operatorname{tr}\{A dB\}$, $dg(B; dB) = \operatorname{tr}\{B^{-1} dB\}$, 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 $X_T \xrightarrow{D} X$, as $T \to \infty$, if and only if, for any $\lambda \in \mathbb{R}^p$, $\lambda \neq 0$,

$$\lambda' X_T \xrightarrow{D} \lambda' X.$$

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

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

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

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

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

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

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

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 $\operatorname{vec}(M)$ is (vector) $\mathrm{N}(0, \Omega)$ 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 $\mathrm{N}\left(0, \left(\Sigma_{qq}^{-1} \otimes \Sigma_{pp}\right)\right)$ distributed. Thus by applying (II.27)

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

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 $v_t \equiv (\varepsilon_t', Z_t')' \in \mathbb{R}^{p+q}$,

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

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

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\left(\operatorname{tr}(V'\varepsilon_t Z_t')\right)^2 | \mathcal{F}_{t-1}\right] \xrightarrow{P} \tag{II.32}$$
$$\operatorname{tr}\left(\Sigma_{\varepsilon\varepsilon} V \Sigma_{zz} V'\right) = \operatorname{vec}(V)'\left(\Sigma_{zz} \otimes \Sigma_{\varepsilon\varepsilon}\right)\operatorname{vec}(V) > 0.$$

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

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

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

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

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

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

Next, by (II.28),

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

and we have

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

And by the Cramer-Wold device, we get

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

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,*

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

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

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

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, $E\left(\varepsilon_t Z_t'\right) = \theta E\left(Z_{t-1} Z_t'\right) \neq 0$, 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*

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

$$\begin{aligned} \tilde{\Omega} &= \frac{1}{T} \sum_{t=1}^{T} \left(Y_t - \tilde{\beta}' Z_t\right) \left(Y_t - \tilde{\beta}' Z_t\right)' \\ &= S_{yy} - S_{yz} H \left(H' S_{zz} H\right)^{-1} H' S_{zy} \end{aligned} \tag{II.39}$$

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

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

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 $W_T \xrightarrow{P} 0$. Moreover, as will be argued below,

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

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

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

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

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

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

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

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

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

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

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

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

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

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

with $A_i \in \mathbb{R}^{p \times p}$, initial values $X_0, ..., X_{-k+1}$ fixed and ε_t i.i.d. $N(0, \Omega)$. 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 $df(x;dx) = \operatorname{tr}\{x^{-1}dx\}$, where dx a small $k \times k$ matrix, and hence, similar to the univariate case,

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

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

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

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

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

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

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

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

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

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

or alternatively, a Taylor expansion around x_0,

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

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

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

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

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

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

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

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

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

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

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

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,

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

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

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

Define $\tilde{f}(x) = [f(x) - f(c) - df(c; x - c)]/||x - c||$ 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*

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

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

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

□

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.*