Financial Econometrics

Slides-06: Generalizing to ARMA and Forecasting

Dr. Rachida Ouysse
School of Economics[1]

- General $AR(p)$
  - Wold Decomposition
  - AF and PACF patterns
  - Impulse response function

- Yule-Walker equations
- AR & MA mix- ARMA models
  - AF and PACF patterns
  - Impulse response function

- Estimation of ARMA
- Forecasting in ARMA

The conditions for stationarity/invertibility of an AR(1) process can be extended to higher order AR processes.

- First consider an AR(2) process

$$\left(1-\alpha_1 L-\alpha_2 L^2\right) y_t=\alpha\left(L\right) y_t=\alpha_0+\varepsilon_t .$$

In general, the polynomial $\alpha\left(L\right)$ can be rewritten as

$$\left(1-\alpha_1 L-\alpha_2 L^2\right)=\left(1-\phi_1 L\right)\left(1-\phi_2 L\right).$$

where ϕ_1 and ϕ_2 can be solved from $\phi_1+\phi_2=\alpha_1$ and $-\phi_1\phi_2=\alpha_2$.

The conditions for invertibility of the second order polynomial are just the conditions that both the first order polynomials $\left(1-\phi_1 L\right)$ and $\left(1-\phi_2 L\right)$ are invertible, i.e. $|\phi_1|<1$ and $|\phi_2|<1$.

A more common way of presenting these conditions is in terms of the so-called **characteristic equation**

$$\left(1-\alpha_1 z-\alpha_2 z^2\right)=0,$$

$$\text{or} \quad \left(1-\phi_1 z\right)\left(1-\phi_2 z\right)=0.$$

This equation has two solutions, denoted z_1 and z_2

$$z_1, z_2=\frac{\alpha_1 \pm \sqrt{\alpha_1^2+4\alpha_2}}{-2\alpha_2},$$

referred to as the **characteristic roots** of the $\alpha(L)$ polynomial.

The requirement $|\phi_i|<1$ corresponds to $|z_i|>1$. If any solution satisfies $|z_i| \leq 1$, the polynomial $\alpha(L)$ is non-invertible. A solution that equals unity is referred to as a unit root.

General Conditions for Stationarity for an $AR(p)$

Calculating the roots of a higher order AR process is computationally not a trivial job. In most circumstances there is little need to directly calculate the characteristic roots, though, as there are some useful **simple rules** for checking stationarity/invertibility of higher order processes

- **Necessary condition**: $\sum_{i=1}^{p} \alpha_i < 1$

- **Sufficient condition**: $\sum_{i=1}^{p} |\alpha_i| < 1$

As, under appropriate conditions, an AR(p) process has an MA(∞) representation and an MA(q) has an AR(∞) representation, there is no fundamental difference between AR and MA models.

- The MA representation is convenient to derive the properties (mean, variance, ...) of a series
- The AR representation is convenient for making predictions conditional upon the past

When estimating time series models (cf. below), the choice is simply a matter of parsimony.

What is the AR process is stationary?

For a stationary AR(p) process, it is more convenient to derive the properties from imposing that the mean, variance and autocovariances do not depend on time.

For computational convenience consider an AR(2) process.

- The unconditional mean of y_t can be solved from

$$E(y_t) = \alpha_0 + \alpha_1 E(y_{t-1}) + \alpha_2 E(y_{t-2})$$

which, assuming that $E(y_t)$ does not depend on time allows us to write

$$E(y_t) = \alpha_0/(1 - \alpha_1 - \alpha_2)$$

- The variance of y_t can be solved by defining $x_t = y_t - E(y_t)$ which yields

$$x_t = \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + \varepsilon_t \tag{18}$$

The variance of y_t can be obtained by multiplying both sides by x_t and taking expectations

$$
\begin{aligned}
V(y_t) = \gamma_0 &= E\left(\alpha_1 x_t x_{t-1} + \alpha_2 x_t x_{t-2} + x_t \varepsilon_t\right) \\
&= \alpha_1 \gamma_1 + \alpha_2 \gamma_2 + E\left(x_t \varepsilon_t\right) \\
&= \alpha_1 \gamma_1 + \alpha_2 \gamma_2 + \sigma^2 \qquad (19)
\end{aligned}
$$

where $E\left(x_t \varepsilon_t\right) = \sigma^2$ is obtained from multiplying both sides of (18) by ε_t and taking expectations. Multiplying both sides by x_{t-1} and x_{t-2} and taking expectations we obtain

$$
\gamma_1 = \alpha_1 \gamma_0 + \alpha_2 \gamma_1 \qquad (20)
$$

$$
\gamma_2 = \alpha_1 \gamma_1 + \alpha_2 \gamma_0 \qquad (21)
$$

These equations can be solved for γ_0 to obtain

$$
\gamma_0 = \frac{(1-\alpha_2)}{(1+\alpha_2)(1-\alpha_1-\alpha_2)(1+\alpha_1-\alpha_2)}\sigma^2
$$

What is the AR process is stationary?

- The autocorrelation coefficients ρ_1 and ρ_2 can be obtained by dividing (20) and (21) by γ_0

$$\rho_1 = \alpha_1 + \alpha_2 \rho_1$$

$$\rho_2 = \alpha_1 \rho_1 + \alpha_2$$

and solving to obtain

$$\rho_1 = \alpha_1 / (1 - \alpha_2)$$

$$\rho_2 = \alpha_1^2 / (1 - \alpha_2) + \alpha_2$$

It is easily verified that the higher-order autocorrelation coefficients are given by

$$\rho_k = \alpha_1 \rho_{k-1} + \alpha_2 \rho_{k-2}$$

Yule Walker Equations

- **The beauty of the yule Walker Equations!**

$X_t = \alpha_1 X_{t-1} + \cdots + \alpha_p X_{t-1} + \epsilon_t$

$$
\begin{aligned}
x_t x_{t-1} &= \alpha_1 x_{t-1} x_{t-1} + \cdots + \alpha_p x_{t-p} x_{t-1} + \epsilon_t x_{t-1} \\
E(x_t x_{t-1}) &= \alpha_1 E(x_{t-1} x_{t-1}) + \cdots + \alpha_p E(x_{t-p} x_{t-1}) + E(\epsilon_t x_{t-1}) \\
\gamma_1 &= \alpha_1 \gamma_0 + \alpha_2 \gamma_1 + \cdots + \alpha_p \gamma_{p-1} \\
&\cdots \\
x_t x_{t-j} &= \alpha_1 x_{t-1} x_{t-j} + \cdots + \alpha_p x_{t-p} x_{t-j} + \epsilon_t x_{t-j} \\
E(x_t x_{t-j}) &= \alpha_1 E(x_{t-1} x_{t-j}) + \cdots + \alpha_p E(x_{t-p} x_{t-j}) + E(\epsilon_t x_{t-j}) \\
\gamma_{|j|} &= \alpha_1 \gamma_{|j-1|} + \alpha_2 \gamma_{|j-2|} + \cdots + \alpha_p \gamma_{|p-j|}, \\
&\cdots \\
\gamma &= \mathbf{\Gamma} \alpha
\end{aligned}
$$

Defining an ARMA Process

Let ε_t be a white noise process. Then:

$$\alpha(L)\, y_t = \alpha_0 + \beta(L)\, \varepsilon_t \tag{22}$$

with $\alpha(L)$ an AR polynomial of order p and $\beta(L)$ an MA polynomial of order q, is an **autoregressive moving average process** with orders p and q, denoted ARMA(p,q).
$\rightarrow$ y_t depends on its own lagged values and on current and past values of a white noise disturbance term ε_t.

Dynamic Behaviour and Impulse Response

If the AR polynomial $\alpha(L)$ is invertible, the ARMA(p, q) process can be written as a stable MA(∞) process of the form

$$y_t = \alpha(L)^{-1}\alpha_0 + \alpha(L)^{-1}\beta(L)\varepsilon_t$$
$$= \alpha_0' + \theta(L)\varepsilon_t$$

where $\alpha_0' = \alpha_0 / \left(1 - \sum_{i=1}^{p} \alpha_i\right)$ and $\theta(L) = \alpha(L)^{-1}\beta(L)$ $= 1 + \sum_{i=1}^{\infty} \theta_i L^i$, with $\theta_i =$ undetermined coefficients.

Even if the AR polynomial is non-invertible, we can still solve for the ε sequence but this solution will not be a stable MA process, i.e.

$$y_t = f(t) + \theta(L)\varepsilon_t$$

where $f(t)$ indicates that the mean is a function of time.

Dynamic Behaviour and Impulse Response

The impulse response function can be obtained from the MA representation.

Note that as a finite order MA process is stationary by construction, an ARMA process is stationary if the AR component is stationary (i.e. if the AR polynomial is invertible).

- In the stationary case the impact of shocks gradually dies out (i.e. $\sum_{i=1}^{\infty} \theta_i$ is finite)
- In the non-stationary case the impact of a shock never vanishes (i.e. $\sum_{i=1}^{\infty} \theta_i$ is infinite)

Unconditional Moments of an $ARMA(p,q)$

If the AR polynomial $\alpha(L)$ is non-invertible the mean, variance and covariances are time-varying

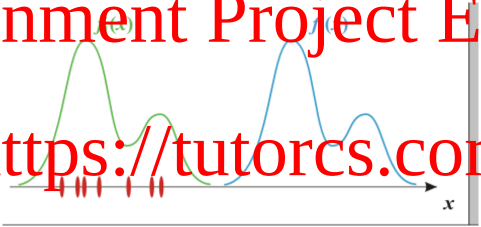
If the AR polynomial $\alpha(L)$ is invertible, the AR process can be rewritten as the stable infinite MA process. The properties of a stationary AR process can easily be derived from this MA representation.

- Unconditional mean: $E(y_t) = \alpha_0 / \left(1 - \sum_{i=1}^{p} \alpha_i\right)$
- Unconditional variance: $V(y_t) = \sigma^2 \sum_{i=0}^{\infty} \theta_i^2$
- Covariances: $\gamma_k = (\theta_k + \theta_1\theta_{k+1} + \theta_2\theta_{k+2} + \dots)\sigma^2$

As an ARMA(p,q) process includes both an AR and an MA component, both the ACF and the PACF do not cut off at some point. As such, it is difficult to determine the order of an ARMA model from the ACF and PACF.

Maximum Likelihood Estimation: Intuitive Illustration

This illustration shows a sample of n independent observations, and two continuous distributions $f_1(x)$ and $f_2(x)$,



Of these two distributions, which one is the most likely to have generated the sample?

Although it is not impossible, we don't believe that $f_2(x)$ generated the sample. Why?

On the other hand, the values taken by $f_1(x)$ are substantial for all the observations, which are then where one would expect them to be, would the sample be actually generated by $f_1(x)$.

- Maximum Likelihood Estimation is a general method of estimation that can be used for many different types of data and economic models. It has very wide applicability.

- The Maximum Likelihood Estimator (MLE) answers the following question: What are the parameter estimates that are most likely to have generated the observed data given the assumed model.

- Begin by assuming a model for the outcome variable including a distribution function for the underlying population error term (and hence a distribution for the outcome variable in the population.)

- Consider $AR(1)$ model: $y_t = \alpha_0 + \alpha_1 y_{t-1} + \epsilon_t$ where $\epsilon_t \sim$ i.i.d $N(0, \sigma^2)$.
  - it follows: $y_t|\Omega_{t-1} \sim N\left(\alpha_0 + \alpha_1 y_{t-1}, \sigma^2\right), \; t = 2, 3, \cdots$.

$$y_1 \sim N\left([1-\alpha_1]^{-1}\alpha_0, [1-\alpha_1^2]^{-1}\sigma^2\right)$$

  - Conditional pdf:

$$f(y_t|\Omega_{t-1}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y_t - \alpha_0 - \alpha_1 y_{t-1})^2}{2\sigma^2}\right\}$$

  - Information sets: $\Omega_1 = \{y_1\}, \Omega_2 = \{y_2, \Omega_1\}, \cdots, \Omega_t = \{y_t, \Omega_{t-1}\}$.
  - Joint pdf for a time series $\{y_1, \cdots, y_T\}$ can be factorised:

$$\begin{aligned} f(y_T, y_{T-1} \cdots, y_1) &= \\ &= f(y_T, y_{T-1} \cdots, y_2|\Omega_1) f(y_1) \\ &= f(y_T, y_{T-1} \cdots, y_3|\Omega_2) f(y_2|\Omega_1) f(y_1) \\ &= f(y_T|\Omega_{T-1}) f(y_{T-1}|\Omega_{T-2}) \cdots f(y_3|\Omega_2) f(y_2|\Omega_1) f(y_1) \end{aligned}$$

– Properties of ML estimators

- When the pdf (likelihood) is correctly specified, the ML estimators have nice large-T sampling properties:
  - consistent,
  - asymptotically normally distributed, and
  - asymptotically efficient.

Allow us to draw inference based on reported SEs.

- When the pdf (likelihood) is incorrect, the "ML" procedure is called quasi (or pseudo) ML.

  - When the normal pdf is used, which may be incorrect, the quasi ML estimators are still consistent and asymptotically normal, as long as the model is defined by the conditional mean and variance that are correctly specified.

Must use "robust" SEs

ARMA Process: Identification

The Box-Jenkins Approach

The so-called Box-Jenkins approach toward fitting ARMA models comprises three stages:

- **Identification**: determine tentative model(s)
  - Plot the time series to have a first idea on the DGP (stationary/non-stationary, structural break, ...)
  - Plot the ACF and the PACF to have a first idea on the order of the ARMA model
- **Estimation**: estimate the various tentative models
  - Compare the estimated models using information criteria
  - Select parsimonious model
- **Diagnostic checking**: check the selected model's diagnostics

AR Process: Estimation

- Consider the **AR(p) model**

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \dots + \alpha_p y_{t-p} + \varepsilon_t$$
$$\alpha(L)\, y_t = \varepsilon_t$$

with ε_t a zero-mean white noise process.

As $y_{t-1}, \dots, y_{t-p}$ are observed in the data, the model can be estimated using **OLS**.

The OLS estimator is

- **biased**, because $E(y_{t-j}\varepsilon_{t-j}) \neq 0$
- **consistent**, because $E(y_{t-j}\varepsilon_t) = 0 \quad \forall j > 0$
- **asymptotically normal**

Intuition: the error terms and the explanatory variables are not completely independent but contemporaneously uncorrelated.

- Consider the **MA(q) model**

$$y_t = \alpha_0 + \beta_1 \varepsilon_{t-1} + \ldots + \beta_q \varepsilon_{t-q} + \varepsilon_t$$
$$y_t = \alpha_0 + \beta(L)\,\varepsilon_t$$

with ε_t a zero-mean white noise process.

As $\varepsilon_{t-1}, \ldots, \varepsilon_{t-q}$ are NOT observed in the data, the model cannot be directly estimated using OLS.

A possible solution is to estimate the coefficients in $\beta(L)$ from the AR representation of the MA model. For an invertible MA(1) model, this is given by (cf. above):

$$y_t = \alpha_0 / (1 + \beta_1) - \sum_{i=1}^{\infty} (-\beta_1)^i \, y_{t-i} + \varepsilon_t$$

Information criteria

A fundamental idea in the Box - Jenkins approach is the principle of **parsimony** (meaning sparseness)

- A parsimonious model fits the data well without incorporating any needless coefficients
- In general parsimonious models produce better forecasts than over-parametrized models

Increasing the lag orders p and q will:

- Increase the goodness-of-fit of the model, i.e. reduce the RSS
- Reduce the degrees of freedom

$\rightarrow$ **information criteria** provide a **trade-off** between the goodness-of-fit of the model and the number of parameters used to obtain that fit.

Model Selection: order of the ARMA(p,q)

The two most commonly used information criteria are:

- **Akaike Information Criterion** (AIC)

$$AIC = T \ln (RSS) + 2k$$

- **Schwarz Bayesian Criterion** (SBC)

$$SIC = T \ln (RSS) + k \ln (T)$$

with $k = p + q + 1$ the number of estimated parameters.

The most appropriate model is the one that minimises AIC and/or SBC.

Model Selection: order of the ARMA(p,q)

Note that:

- When you estimate models with lagged variables some initial observations are lost. In order to compare models using information criteria, you should **keep** T **fixed**! Otherwise you will be comparing the performance of the models over different sample periods. Moreover, decreasing T has the direct effect of reducing the AIC and SBC.
- The SBC embodies a much stiffer penalty for the loss of degrees of freedom than the AIC. The main difference between the two in terms of performance is that SBC is consistent (i.e. asymptotically it delivers the correct model) while the AIC is biased toward selecting an over-parametrised model. However, in small samples, the AIC can work better than the SBC.

- If data are generated by an ARMA, the probability that SIC selects the correct model converges to one as $T \to \infty$.
  - In finite samples, SIC may select a too-small model.

eg. unanticipated US monthly inflation:

- SIC selects AR(1)
- AIC selects ARMA(1,1)

Sample (adjusted): 1978M02 1987M12
Included observations: 119 after adjustments
Convergence achieved after 2 iterations

Variable	Coefficient	Std. Error	t-Statistic	Prob.
AR(1)	0.574656	0.076201	7.541340	0.0000

R-squared	0.325217	Mean dependent var	0.000121
Adjusted R-squared	0.325217	S.D. dependent var	0.061977
S.E. of regression	0.050911	Akaike info criterion	-3.109100
Sum squared resid	0.305850	Schwarz criterion	-3.085746
Log likelihood	185.9914	Durbin-Watson stat	1.821211

Inverted AR Roots	.57

Sample (adjusted): 1978M02 1987M12
Included observations: 119 after adjustments
Convergence achieved after 6 iterations
Backcast: OFF

Variable	Coefficient	Std. Error	t-Statistic	Prob.
AR(1)	0.315395	0.141540	2.228308	0.0278
MA(1)	0.386570	0.136971	2.822271	0.0056

R-squared	0.350263	Mean dependent var	0.000121
Adjusted R-squared	0.344710	S.D. dependent var	0.061977
S.E. of regression	0.050170	Akaike info criterion	-3.130116
Sum squared resid	0.294498	Schwarz criterion	-3.083408
Log likelihood	188.2419	Durbin-Watson stat	2.004409

Inverted AR Roots	.32
Inverted MA Roots	-.39

Forecasting using ARMA models

- In-sample versus out-of-sample forecasts
  - **In-sample forecasts** are those generated for the same set of data that was used to estimate the model's parameters. Good performance may be due to fitting a spurious model to the noise in the sample, though!
  - **Out-of-sample forecasts** are those generated for a set of data that was not used to estimate the model, i.e. do not use all observations in estimating the model and evaluate the model from the forecasting accuracy in the holdout sample.
- Static versus dynamic forecasts
  - **Static forecasts** are a sequence of one-step-ahead forecasts, using actual, rather than forecasted values for lagged dependent variables.
  - **Dynamic forecasts** are a sequence of multi-step-ahead forecasts starting from the first period in the forecast sample, using forecasted values for lagged dependent variables.

Forecasting Accuracy

In addition to the prediction itself, it is important to know how accurate this prediction is. To judge forecasting accuracy, define the **prediction error** as

$$fe_{t,s} = y_{t+s} - f_{t,s}$$

and the **variance of the forecasting error** by

$$var\left(fe_{t,s}\right) = E\left(y_{t+s} - f_{t,s}\right)^2$$

For the **MA(q) model** we have

$$
\begin{aligned}
fe_{t,1} &= \varepsilon_{t+1} \\
fe_{t,2} &= \varepsilon_{t+2} + \beta_1 \varepsilon_{t+1} \\
fe_{t,3} &= \varepsilon_{t+3} + \beta_1 \varepsilon_{t+2} + \beta_2 \varepsilon_{t+1} \\
&\vdots \\
fe_{t,q} &= \varepsilon_{t+q} + \beta_1 \varepsilon_{t+q-1} + \ldots + \beta_{q-1} \varepsilon_{t+1} \\
fe_{t,q+1} &= \varepsilon_{t+q+1} + \beta_1 \varepsilon_{t+q} + \ldots + \beta_q \varepsilon_{t+1} \\
fe_{t,q+2} &= \varepsilon_{t+q+2} + \beta_1 \varepsilon_{t+q+1} + \ldots + \beta_q \varepsilon_{t+2} \\
&\vdots
\end{aligned}
$$

such that

$$var\left(fe_{t,1}\right) = E\left(\varepsilon_{t+1}\right)^2 = \sigma^2$$

$$var\left(fe_{t,2}\right) = E\left(\varepsilon_{t+2} + \beta_1\varepsilon_{t+1}\right)^2 = \left(1+\beta_1^2\right)\sigma^2$$

$$var\left(fe_{t,3}\right) = E\left(\varepsilon_{t+3} + \beta_1\varepsilon_{t+2} + \beta_2\varepsilon_{t+1}\right)^2 = \left(1+\beta_1^2+\beta_2^2\right)\sigma^2$$

$$\vdots$$

$$\begin{aligned} var\left(fe_{t,q}\right) &= E\left(\varepsilon_{t+q} + \beta_1\varepsilon_{t+q-1} + \ldots + \beta_{q-1}\varepsilon_{t+1}\right)^2 \\ &= \left(1+\beta_1^2+\ldots+\beta_{q-1}^2\right)\sigma^2 \end{aligned}$$

$$\begin{aligned} var\left(fe_{t,q+1}\right) &= E\left(\varepsilon_{t+q+1} + \beta_1\varepsilon_{t+q} + \ldots + \beta_q\varepsilon_{t+1}\right)^2 \\ &= \left(1+\beta_1^2+\ldots+\beta_q^2\right)\sigma^2 \end{aligned}$$

The accuracy of the prediction

- decreases as we predict further into the future
- does not decrease any further from $s = q + 1$ onward as the variance of the prediction error stabilises at the unconditional variance. This is the upper bound on the inaccuracy of the predictor.

Forecasting AR(p)

For a **stationary AR(p) model**, the prediction errors are most easily obtained from the MA(∞) representation

$$y_t = \alpha_0 / (1 - \alpha_1 - \ldots \alpha_p) + \sum_{i=0}^{\infty} \beta_i \varepsilon_{t-i}$$

with β_i undetermined coefficients.

Consequently, the s-period-ahead prediction error is given by

$$fe_{t,s} = \sum_{i=0}^{s-1} \beta_i \varepsilon_{t+s-i}$$

with variance

$$var\left(fe_{t,s}\right) = \sigma^2 \sum_{i=0}^{s-1} \beta_i^2$$

The accuracy of the prediction

- decreases as we predict further into the future as $\beta_i^2 > 0$
- converges to the stable unconditional variance $\sigma^2 \sum_{i=0}^{\infty} \beta_i^2$ as $t \to \infty$. This is the upper bound on the inaccuracy of the predictor.

As an illustration, consider an AR(1) model where $\beta_i = \alpha_1^i$. The forecasting errors are given by

$$fe_{t,1} = \varepsilon_{t+1}$$

$$fe_{t,2} = \varepsilon_{t+2} + \alpha_1 \varepsilon_{t+1}$$

$$\vdots$$

$$fe_{t,s} = \sum_{i=0}^{s-1} \alpha_1^i \varepsilon_{t+s-i}$$