STATIONARY PROCESSES – BASICS

Often we observe a time series whose fluctuations appear random, but with the same type of random behaviour from one time period to the next.

e.g. returns on stocks are random and the returns one year can be very different from the previous year, but the mean and standard deviation are often similar from year to the next.

Intuitively, a process $\{Y_t\}$ is said to be stationary if all aspects of its behaviour are unchanged by shifts in time. Various def's:

<u>Def</u> A sequence $\{Y_t\}_{t\in\mathbb{Z}}$ is <u>strongly stationary</u> if

$$(Y_{t_1}, Y_{t_2}, \dots, Y_{t_k}) \overset{D}{=} (Y_{t_1+h}, \dots, Y_{t_k+h})$$ for all sets of time points $t_1, t_2, \dots, t_k$ and any ("lag") integer h

<u>Def.</u> A sequence $\{Y_t\}_{t\in\mathbb{Z}}$ is <u>weakly stationary</u> if

a) $\mathbb{E}[Y_t] = \mu$

b) $\operatorname{cov}(Y_t, Y_{t+k}) = \gamma_k$ where μ, γ_k are constants independent of t.

<u>Note</u>: $\operatorname{cov}(Y_t, Y_{t+k}) = \mathbb{E}[(Y_t - \mathbb{E}[Y_t])(Y_{t+k} - \mathbb{E}[Y_{t+k}])] = \mathbb{E}[Y_t Y_{t+k}] - \mu^2$.

<u>Def</u>. The sequence $\{\gamma_k\}_{k\in\mathbb{Z}}$ is called the <u>autocovariance function</u>.

The function $\rho_k := \gamma_k/\gamma_0 = \operatorname{corr}(Y_t, Y_{t+k})$ is called the <u>autocorrelation function</u> (ACF).

Clearly, $\gamma_0 = \operatorname{var}(Y_t)$ and $\gamma_k = \gamma_{-k}$ for all k, by symmetry.

<u>Def</u>. A sequence $\{Y_t\}_{t\in\mathbb{Z}}$ is <u>Gaussian</u> if the joint density

$$f_{Y_{t_1}, Y_{t_2}, \dots, Y_{t_k}}(y_{t_1}, \dots, y_{t_k})$$ is multivariate normal for all $t_1, \dots, t_k$.

<u>Note</u>:

$$\text{Strong stationary} \; \underset{\not\longleftarrow}{\longrightarrow} \; \text{Weak stationary}$$

only if Gaussian as well

We will work mostly with weak stationary time series, which we'll call stationary from now on.

$\boxed{1.1}$ White noise

The basic building block for all stationary TS is a sequence $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ whose elements satisfy:

$$\left.\begin{aligned} &\mathbb{E}[\varepsilon_t]=0 \\ &\mathbb{E}[\varepsilon_t^2]=\sigma^2 \\ &\mathbb{E}[\varepsilon_t\varepsilon_\tau]=0 \quad \text{for all } t\neq\tau \end{aligned}\right\} \underline{\text{White noise process}} \quad WN(0,\sigma^2)$$

we mostly work with this

Clearly, $\gamma_0=\sigma^2$, $\gamma_k=0$ for $k\neq 0$.

$\rho_0=1$, $\rho_k=0$ for $k\neq 0$

If, in addition, we assume that $\varepsilon_t, \varepsilon_\tau$ are independent for $t\neq\tau$, then we have independent white noise.

If we also assume that $\varepsilon_t\sim N(0,\sigma^2)$, then we have Gaussian white noise.

$\boxed{1.2.}$ Moving Average process

$\boxed{1.2.1}$ MA(1) $\{Y_t\}_{t\in\mathbb{Z}}$

$$Y_t=\mu+\varepsilon_t+\theta\varepsilon_{t-1} \qquad \text{where } \mu, \theta \text{ are any constants.}$$

Intuition: Y_t is constructed from a weighted sum of the two most recent shocks ε.

Expectation: $\mathbb{E}[Y_t]=\mathbb{E}[\mu+\varepsilon_t+\theta\varepsilon_{t-1}]=\mu+\mathbb{E}[\varepsilon_t]+\theta\mathbb{E}[\varepsilon_{t-1}]=\mu$.

Variance: $\mathbb{E}[(Y_t-\mu)^2]=\mathbb{E}[(\varepsilon_t+\theta\varepsilon_{t-1})^2]=\mathbb{E}[\varepsilon_t^2+2\theta\varepsilon_{t-1}\varepsilon_t+\theta^2\varepsilon_{t-1}^2]=$
$=\sigma^2+0+\theta^2\sigma^2=(1+\theta^2)\sigma^2$.

Autocovariance: $\mathbb{E}[(Y_t-\mu)(Y_{t-1}-\mu)]=\mathbb{E}[(\varepsilon_t+\theta\varepsilon_{t-1})(\varepsilon_{t-1}+\theta\varepsilon_{t-2})]=$
$=\mathbb{E}[\varepsilon_t\varepsilon_{t-1}+\theta\varepsilon_{t-1}^2+\theta\varepsilon_t\varepsilon_{t-2}+\theta^2\varepsilon_{t-1}\varepsilon_{t-2}]=0+\theta\sigma^2+0+0=\theta\sigma^2$

For $j>1$: $\mathbb{E}[(Y_t-\mu)(Y_{t-j}-\mu)]=\mathbb{E}[(\varepsilon_t+\theta\varepsilon_{t-1})(\varepsilon_{t-j}+\theta\varepsilon_{t-j-1})]=0$.

So, $\gamma_0=(1+\theta^2)\sigma^2$, $\gamma_1=\theta\sigma^2$, $\gamma_k=0$ for $k>1$.

$\rho_0=1$, $\rho_1=\dfrac{\theta}{1+\theta^2}$, $\rho_k=0$ for $k>1$.

Note: In general (not just for MA(1)), $|\gamma_k|\leq|\gamma_0|$, i.e. $|\rho_k|\leq 1$ for all k. Why?

Schwartz Inequality: $\int(f(x))^2dx\int(g(x))^2dx\geq\left(\int f(x)g(x)dx\right)^2$

$$\mathbb{E}[Y_t^2]\,\mathbb{E}[Y_{t-k}^2]\geq(\mathbb{E}[Y_tY_{t-k}])^2$$

$$\gamma_0\gamma_0\geq\gamma_k^2 \Rightarrow |\gamma_0|\geq|\gamma_k|.$$

<u>Note</u>: The fact that ACF of MA(1) is 0 for $k>1$ is used as a good diagnostic that a given TS can be modelled as a MA(1) process.

<u>Note</u>: Identification problem: Value of $\frac{\theta}{1+\theta^2}=\rho_1$ is unchanged after $\theta \to 1/\theta$.

E.g. the processes $Y_t=\varepsilon_t+\frac{1}{2}\varepsilon_{t-1}$ and $Y_t=\varepsilon_t+2\varepsilon_{t-1}$ have the same ACF.

We can avoid this by considering only <u>invertible</u> MA(1)'s, i.e. those for which $|\theta|<1$. ⊕

1.2.2 <u>MA(q)</u> $\{y_t\}_{t\in\mathbb{Z}}$ $\qquad Y_t=\mu+\sum_{j=0}^{q}\theta_j\varepsilon_{t-j}$.

$$y_t=\mu+\varepsilon_t+\theta_1\varepsilon_{t-1}+\theta_2\varepsilon_{t-2}+\dots+\theta_q\varepsilon_{t-q}$$, $\theta_1,\dots,\theta_q$ any real numbers.

Expectation: $\mathbb{E}[Y_t]=\mu$

Variance: $\gamma_0=\mathbb{E}[(Y_t-\mu)^2]=\mathbb{E}[(\varepsilon_t+\theta_1\varepsilon_{t-1}+\dots+\theta_q\varepsilon_{t-q})^2]=$ ← ε's are uncorrelated

$$=\sigma^2+\theta_1^2\sigma^2+\theta_2^2\sigma^2+\dots+\theta_q^2\sigma^2=\sigma^2(1+\theta_1^2+\theta_2^2+\dots+\theta_q^2)$$

Autocovariance: $\gamma_j=\mathbb{E}[(\varepsilon_t+\theta_1\varepsilon_{t-1}+\dots+\theta_q\varepsilon_{t-q})\cdot(\varepsilon_{t-j}+\theta_1\varepsilon_{t-j-1}+\dots+\theta_q\varepsilon_{t-q-j})]=$

$$=\begin{cases}\mathbb{E}[\theta_j\varepsilon_{t-j}^2+\theta_{j+1}\theta_1\varepsilon_{t-j-1}^2+\dots+\theta_q\theta_{q-j}\varepsilon_{t-q}^2] & \text{for } j=1,2,\dots,q\\ 0 & \text{for } j>q\end{cases}$$

So, $$\gamma_j=\begin{cases}(\theta_j+\theta_{j+1}\theta_1+\dots+\theta_q\theta_{q-j})\sigma^2 & \text{for } j=1,2,\dots,q\\ 0 & \text{for } j>q\end{cases}$$

(Weak) stationarity of MA(q) for any q is now obvious.

Again, the ACF is zero after q lags, which is a good diagnostic.

1.2.3* <u>MA(∞)</u> $\{Y_t\}_{t\in\mathbb{Z}}$

$$Y_t=\mu+\sum_{j=0}^{\infty}\theta_j\varepsilon_{t-j}$$

It is stationary if $\sum_{j=0}^{\infty}\theta_j^2<\infty$.

MA(q) is <u>invertible</u> (i.e. it has an AR(∞) representation) when all the roots of

$$1+\theta_1 z+\theta_2 z^2+\dots+\theta_q z^q=0$$ lie outside the unit circle.

$$\mathbb{E}[Y_t]=\mu,\quad \gamma_0=\lim_{T\to\infty}(\theta_0^2+\theta_1^2+\dots+\theta_T^2)\sigma^2,\quad \gamma_j=\sigma^2(\theta_j\theta_0+\theta_{j+1}\theta_1+\theta_{j+2}\theta_2+\dots)$$

⊕ Invertible MA(1) models have AR(∞) representations (you'll see AR's later on)

$$(Y_t-\mu)=(1+\theta L)\varepsilon_t,\quad |\theta|<1$$
$$=(1-(-\theta)L)\varepsilon_t$$
$$\Rightarrow (1-(-\theta)L)^{-1}(Y_t-\mu)=\varepsilon_t \Rightarrow \varepsilon_t=\sum_{j=0}^{\infty}(-\theta)^j L^j(Y_t-\mu)$$

11.3 Autoregressive processes

11.3.1 AR(1)

$$Y_t = c + \phi Y_{t-1} + \varepsilon_t \quad , \; c, \phi \text{ constants}$$

Resembles linear regression (Ch. 6) $Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t$

$\beta_0 = c$, $\beta_1 = \phi$.

It's like regression of the process on its own past values, hence the name.

Intuition: think of ϕY_{t-1} as representing "memory" or "feedback" of the past into the present value of the process. It introduces correlation btw. Y_t and the past.

If $\phi = 0$, then $\{Y_t\}$ is $WN(c, \sigma^2)$. Think of ε_t as representing "new information" at time t, that cannot be anticipated so that the effects of today's new information is independent of the effects of yesterday's news.

If $|\phi| < 1$, then $\{Y_t\}_{t\in\mathbb{Z}}$ is a (weakly) stationary process. So, we assume $|\phi| < 1$.

Recursive substitution:

$$Y_t = c + \varepsilon_t + \phi(c + \varepsilon_{t-1} + \phi Y_{t-2}) = (c + \varepsilon_t) + \phi(c + \varepsilon_{t-1}) + \phi^2(c + \varepsilon_{t-2}) + \phi^3(c + \varepsilon_{t-3}) + \dots$$

i.e. $$Y_t = \frac{c}{1-\phi} + \varepsilon_t + \phi\varepsilon_{t-1} + \phi^2\varepsilon_{t-2} + \phi^3\varepsilon_{t-3} + \dots$$

since $|\phi| < 1$, then $\sum_{j=0}^{\infty} \phi^j = \frac{1}{1-\phi}$

this is an MA(∞) representation of an AR(1) process with $\theta_j = \phi^j$.

Y_t depends on all previous shocks with varying significance.

Expectation: $\mathbb{E}[Y_t] = \frac{c}{1-\phi} =: \mu$

Variance: $\gamma_0 = \mathbb{E}[(Y_t - \mu)^2] = \mathbb{E}[(\varepsilon_t + \phi\varepsilon_{t-1} + \phi^2\varepsilon_{t-2} + \phi^3\varepsilon_{t-3} + \dots)^2] \Rightarrow$

$$\gamma_0 = (1 + \phi^2 + \phi^4 + \phi^6 + \dots)\sigma^2 \quad , \text{ i.e. } \gamma_0 = \frac{\sigma^2}{1-\phi^2}.$$

Autocovariance: $\gamma_j = \mathbb{E}[(Y_t - \mu)(Y_{t-j} - \mu)] =$

$$= \mathbb{E}\left[(\varepsilon_t + \phi\varepsilon_{t-1} + \dots + \phi^j\varepsilon_{t-j} + \phi^{j+1}\varepsilon_{t-j-1} + \phi^{j+2}\varepsilon_{t-j-2} + \dots)(\varepsilon_{t-j} + \phi\varepsilon_{t-j-1} + \phi^2\varepsilon_{t-j-2} + \dots)\right]$$

So, $\gamma_j = (\phi^j + \phi^{j+2} + \phi^{j+4} + \dots)\sigma^2 = \phi^j \cdot (1 + \phi^2 + \phi^4 + \dots)\sigma^2$

$$\gamma_j = \frac{\phi^j}{1-\phi^2}\sigma^2.$$

ACF: $\rho_j = \frac{\gamma_j}{\gamma_0} \Rightarrow \rho_j = \phi^j.$

Note: This is not as good of a diagnostic as we had for MA(q).

Note: If ϕ is larger, then mean-reversion is slower, i.e. strong shocks need considerable time to die out. ACF depends on only one parameter, ϕ, which is remarkable parsimony. ACF decays geometrically to zero (actually, if $\phi < 0$, then the sign of ACF oscillates as its magnitude decays geometrically).

Note: We obtained above formulae by viewing AR(1) as MA(∞). However, if we assume stationarity, we can get those formulae even easier!

$$Y_t = c + \phi Y_{t-1} + \varepsilon_t \Rightarrow \mathbb{E}[Y_t] = c + \phi \mathbb{E}[Y_{t-1}] + \mathbb{E}[\varepsilon_t]$$

$$\underset{\text{stationarity}}{\Longrightarrow} \mu = c + \phi\mu + 0 \Rightarrow \mu = \frac{c}{1-\phi}.$$

Now, $Y_t = \mu(1-\phi) + \phi Y_{t-1} + \varepsilon_t$, i.e. $Y_t - \mu = \phi(Y_{t-1} - \mu) + \varepsilon_t$. (*) MEAN-ADJUSTED FORM OF AR(1)

$$\Rightarrow \mathbb{E}[(Y_t-\mu)^2] = \phi^2 \mathbb{E}[(Y_{t-1}-\mu)^2] + 2\phi\mathbb{E}[(Y_{t-1}-\mu)\varepsilon_t] + \mathbb{E}[\varepsilon_t^2]$$

new information ε_t is uncorrelated to $Y_{t-1} \Rightarrow \mathbb{E}[(Y_{t-1}-\mu)\varepsilon_t] = 0 \Rightarrow$

& stationarity $\Rightarrow \gamma_0 = \phi^2\gamma_0 + 0 + \sigma^2 \Rightarrow \gamma_0 = \dfrac{\sigma^2}{1-\phi^2}$

Also from (*), we have $\mathbb{E}[(Y_t-\mu)(Y_{t-j}-\mu)] = \phi\mathbb{E}[(Y_{t-1}-\mu)(Y_{t-j}-\mu)] + \mathbb{E}[\varepsilon_t(Y_{t-j}-\mu)]$

stationarity $\Rightarrow \quad \gamma_j = \phi\gamma_{j-1} + 0 \Rightarrow \gamma_j = \phi\gamma_{j-1} \Rightarrow \gamma_j = \phi^j\gamma_0$.

Note: If $\phi = 1$, then the mean-adjusted form gives $Y_t = Y_{t-1} + \varepsilon_t$

This is a random walk $Y_t = Y_0 + \sum_{j=1}^{t} \varepsilon_j$ with $var(Y_t) = \sigma^2 t$ depending on t.

1.3.2 Wold's decomposition theorem

So far, every process had a representation $Y_t = \mu + \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}$ with $\varepsilon_t \sim WN(0,\sigma^2)$.

This holds in general for ~~every~~ weakly stationary process!

Wold's thm Any weakly stationary timeseries $\{Y_t\}$ can be represented in the form

$$Y_t = \mu + \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}, \quad \varepsilon_t \sim WN(0,\sigma^2), \quad \psi_0 = 1 \text{ and } \sum_{j=0}^{\infty} \psi_j^2 < \infty.$$

Example: For MA(q), we have $\psi_j = \theta_j$, $j = 1, \dots, q$ and $\psi_j = 0$ for $j > q$.

For AR(1), we have $\psi_j = \phi^j$.

Note: This is one of the three fundamental representations of any weakly stationary TS.

Note: Once you find Wold's representation, then easily: $\mathbb{E}[Y_t] = \mu$, $\gamma_0 = \sigma^2 \sum_{j=0}^{\infty} \psi_j^2 < \infty$

and $\gamma_j = \sigma^2 \sum_{k=0}^{\infty} \psi_k \psi_{k+j}$

1.3.3 Lag operator L

Def. $LY_t = Y_{t-1}$, $L^j Y_t = Y_{t-j}$

<u>Example</u> AR(1) in Lag operator notation (assuming $\mu=0$)

$$\underbrace{(1-\phi L)}_{}Y_t = \varepsilon_t \iff Y_t = \phi Y_{t-1} + \varepsilon_t.$$

lag polynomial $\phi(L) = 1-\phi L$

If $|\phi|<1$, then the inverse of the lag polynomial exists $\psi(L) = \phi(L)^{-1}$

$$\psi(L) = (1-\phi L)^{-1} = \sum_{j=0}^{\infty} \phi^j L^j = 1 + \phi L + \phi^2 L^2 + \dots$$

Now, $Y_t = (1-\phi L)^{-1}\varepsilon_t = \sum_{j=0}^{\infty}\phi^j L^j \varepsilon_t = \sum_{j=0}^{\infty}\phi^j \varepsilon_{t-j}$

This is exactly Wold's representation of AR(1), where $\psi_j = \phi^j$.

<u>Def.</u> ψ_j is also known as the impulse response function (IRF).

<u>Note:</u> <u>Half-life of real exchange rates</u>

The real exchange rate is defined as $z_t = s_t - p_t + p_t^*$

s_t: log nominal exchange rate; p_t: log of domestic price level; p_t^*: log of foreign price level

Purchasing power parity (PPP) suggests that z_t should be stationary.

Half life: lag at which IRF decreases by one half. (a measure of the speed of mean-reversion)

For AR(1): $\psi_j = \phi^j = 1/2 \Rightarrow j = \dfrac{\ln(0.5)}{\ln\phi}$

$\boxed{1.3.4}$ <u>AR(p)</u>

Mean-adjusted form $\quad Y_t - \mu = \phi_1(Y_{t-1}-\mu) + \dots + \phi_p(Y_{t-p}-\mu) + \varepsilon_t$

OR Regression form $\quad Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \dots + \phi_p Y_{t-p} + \varepsilon_t.$

OR Lag operator form $\quad \phi(L)(Y_t-\mu) = \varepsilon_t$ where $\phi(L) = 1-\phi_1 L - \dots - \phi_p L^p$.

When $p=1$, we know that $\{Y_t\}$ is stationary when $|\phi|<1$. But, what if $p>1$?

<u>Trick:</u> Write the AR(p) in yet another form, so called <u>state space model form</u>.

Let $X_t = Y_t - \mu$. Then AR(p): $\phi(L)X_t = \varepsilon_t$.

Rewrite $\phi(L)X_t = \varepsilon_t$ as follows:

$$\begin{pmatrix} X_t \\ X_{t-1} \\ \vdots \\ X_{t-p+1} \end{pmatrix} = \begin{pmatrix} \phi_1 & \phi_2 & \cdots & \phi_p \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots 1 & 0 \end{pmatrix} \begin{pmatrix} X_{t-1} \\ X_{t-2} \\ \vdots \\ X_{t-p} \end{pmatrix} + \begin{pmatrix} \varepsilon_t \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

$$\underset{(p\times 1)}{\xi_t} = \underset{(p\times p)}{F} \cdot \underset{(p\times 1)}{\xi_{t-1}} + \underset{(p\times 1)}{V_t}$$

Recurse as with AR(1) $\Rightarrow$ $\xi_t = F^{j+1}\xi_{t-j-1} + F^j V_{t-j} + F^{j-1}V_{t-j+1} + \dots + FV_{t-1} + V_t$

Intuition: Stationarity should require $\lim_{j\to\infty} F^j = 0$ ← zero matrix.

Example AR(2) (mean 0)

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \varepsilon_t \iff \begin{pmatrix} X_t \\ X_{t-1} \end{pmatrix} = \begin{pmatrix} \phi_1 & \phi_2 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} X_{t-1} \\ X_{t-2} \end{pmatrix} + \begin{pmatrix} \varepsilon_t \\ 0 \end{pmatrix}$$

iterate j lags $\Rightarrow$ $\begin{pmatrix} X_t \\ X_{t-1} \end{pmatrix} = \begin{pmatrix} \phi_1 & \phi_2 \\ 1 & 0 \end{pmatrix}^{j+1}\begin{pmatrix} X_{t-j-1} \\ X_{t-j-2} \end{pmatrix} + \begin{pmatrix} \phi_1 & \phi_2 \\ 1 & 0 \end{pmatrix}^{j}\begin{pmatrix} \varepsilon_{t-j} \\ 0 \end{pmatrix} + \dots + \begin{pmatrix} \phi_1 & \phi_2 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} \varepsilon_{t-1} \\ 0 \end{pmatrix} + \begin{pmatrix} \varepsilon_t \\ 0 \end{pmatrix}$

first row $\Rightarrow$ $X_t = \left[(F^{j+1})_{1,1}X_{t-j-1} + (F^{j+1})_{1,2}X_{t-j-2}\right] + \underbrace{(F^j)_{1,1}}_{\psi_j}\varepsilon_{t-j} + \dots + \underbrace{(F)_{1,1}}_{\psi_1}\varepsilon_{t-1} + \varepsilon_t$

It's convincing now that the following result holds:

FACT The AR(p) model is stationary and has Wold representation

$$Y_t = \mu + \sum_{j=0}^{\infty}\psi_j\varepsilon_{t-j}, \quad \psi_0 = 1 \quad \text{with}$$

ψ_j being the (1,1)-entry of F^j PROVIDED that all the eigenvalues of F have modulus less than one.

How do we find eigenvalues? We solve $\det(F - \lambda I_p) = 0$ ← identity $p\times p$ matrix

In the above example: $\det(F-\lambda I_2) = \det\begin{pmatrix} \phi_1-\lambda & \phi_2 \\ 1 & -\lambda \end{pmatrix} = \lambda^2 - \phi_1\lambda - \phi_2 = 0$.

So, the eigenvalues of F solve the reverse characteristic equation $\lambda^p - \phi_1\lambda^{p-1} - \phi_2\lambda^{p-2} - \dots - \phi_p = 0$

Reminder: What is the modulus? Well, the roots could be complex. $\lambda_i = a + bi$

$$a = R\cos\alpha,\ b = R\sin\alpha,\ R = \sqrt{a^2+b^2} = \text{modulus}$$

To see why $|\lambda_i| < 1$ implies $\lim_{j\to\infty} F^j = 0$, let's consider the AR(2) with real eigenvalues.

By diagonalization:

$$F = T\Lambda T^{-1} \text{ for some } T^{-1} = T^t, \text{ where } \Lambda = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$$

Then $F^j = T\Lambda^j T^{-1}$ and $\lim_{j\to\infty} F^j = T \lim_{j\to\infty}\Lambda^j T^{-1} = 0$ since $|\lambda_1| < 1$ and $|\lambda_2| < 1$.

$$\Lambda^j = \begin{pmatrix} \lambda_1^j & 0 \\ 0 & \lambda_2^j \end{pmatrix}$$

Stationarity conditions on the lag polynomial $\phi(L) = 1-\phi_1 L - \ldots - \phi_p L^p$

Consider the AR(2) model: $(1-\phi_1 L - \phi_2 L^2) X_t = \varepsilon_t$

characteristic equation $1-\phi_1 z - \phi_2 z^2 = 0$. By fundamental theorem of algebra, it can be written as $(1-\lambda_1 z)(1-\lambda_2 z) = 0$ so that $z_1 = 1/\lambda_1$ and $z_2 = 1/\lambda_2$ are the roots of the characteristic equation. The values λ_1 and λ_2 are the eigenvalues of F.

FACT The inverses of the roots of the characteristic equation $\phi(z) = 1-\phi_1 z - \phi_2 z^2 \ldots - \phi_p z^p = 0$ are the eigenvalues of the companion matrix F. Hence, the AR(p) model is stationary provided the roots of $\phi(z)=0$ have modulus greater than unity. (roots lie outside the complex unit circle)

Note: Given that $\{X_t\}$ is a zero-mean AR(p) TS, it's easy to find its Wold representati

$$\phi(L) = 1-\phi_1 L - \ldots - \phi_p L^p = (1-\lambda_1 L)(1-\lambda_2 L) \cdot \ldots \cdot (1-\lambda_p L)$$

Then $\Psi(L) = \phi(L)^{-1} = (1-\lambda_1 L)^{-1}(1-\lambda_2 L)^{-1} \cdots (1-\lambda_p L)^{-1}$

Suppose λ_i real (of course, $|\lambda_i| < 1$)

$$\Psi(L) = \left(\sum_{j=0}^{\infty} \lambda_1^j L^j\right)\left(\sum_{j=0}^{\infty} \lambda_2^j L^j\right) \cdots \left(\sum_{j=0}^{\infty} \lambda_p^j L^j\right)$$

So, the Wold form can be found using

$$X_t = \Psi(L)\varepsilon_t = \left(\sum_{j=0}^{\infty} \lambda_1^j L^j\right) \cdots \left(\sum_{j=0}^{\infty} \lambda_p^j L^j\right)\varepsilon_t .$$

Note: Sometimes, we can use other tricks. To illustrate, consider the AR(2) model, whose Wold form we want to find.

$$\underset{(1-\phi_1 L - \phi_2 L^2)^{-1}}{\underset{\parallel}{\phi(L)^{-1}}} = \Psi(L) = \sum_{j=0}^{\infty} \psi_j L^j \implies 1 = (1-\phi_1 L - \phi_2 L^2)(1+\psi_1 L + \psi_2 L^2 + \ldots)$$

collect the coefficients of powers of L $\implies$

$$\implies 1 = 1 + (\psi_1 - \phi_1) L + (\psi_2 - \phi_1 \psi_1 - \phi_2) L^2 + \ldots$$

all coefficients on powers of L must be equal to zero, so we have:

$$\psi_1 = \phi_1$$
$$\psi_2 = \phi_1 \psi_1 + \phi_2$$
$$\psi_3 = \phi_1 \psi_2 + \phi_2 \psi_1$$
$$\vdots$$
$$\psi_j = \phi_1 \psi_{j-1} + \phi_2 \psi_{j-2}$$
$$\vdots$$

recursion to get Wold coefficients ψ_j.

What about expectation and ACF? Assume the series $\{Y_t\}$ is stationary $\boxed{AR(2)}$.

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \varepsilon_t$$ Taking expectations, we get:

$$\mathbb{E}[Y_t] = c + \phi_1 \mathbb{E}[Y_{t-1}] + \phi_2 \mathbb{E}[Y_{t-2}] + \mathbb{E}[\varepsilon_t] \Rightarrow \mu = c + \phi_1\mu + \phi_2\mu + 0 \Rightarrow \mu = \frac{c}{1-\phi_1-\phi_2}$$

To find second moments, use the mean-adjusted form

$$(Y_t - \mu) = \phi_1 (Y_{t-1} - \mu) + \phi_2 (Y_{t-2} - \mu) + \varepsilon_t$$

$$\mathbb{E}[(Y_t-\mu)(Y_{t-j}-\mu)] = \phi_1 \mathbb{E}[(Y_{t-1}-\mu)(Y_{t-j}-\mu)] + \phi_2 \mathbb{E}[(Y_{t-2}-\mu)(Y_{t-j}-\mu)] + \varepsilon_t$$

$$\gamma_j = \phi_1 \gamma_{j-1} + \phi_2 \gamma_{j-2} \quad \text{for } j = 1, 2, \ldots$$ ← all you need to do is to solve a second order difference equation

i.e. $\rho_j = \phi_1 \rho_{j-1} + \phi_2 \rho_{j-2}$ for $j = 1, 2, \ldots$

How do we find the variance γ_0?

$$Y_t - \mu = \phi_1 (Y_{t-1}-\mu) + \phi_2 (Y_{t-2}-\mu) + \varepsilon_t \Rightarrow \mathbb{E}[(Y_t-\mu)^2] = \phi_1 \mathbb{E}[(Y_{t-1}-\mu)(Y_t-\mu)] + \phi_2 \mathbb{E}[(Y_{t-2}-\mu)(Y_t-\mu)] + \mathbb{E}[\varepsilon_t(Y_t-\mu)]$$

$$\Rightarrow \gamma_0 = \phi_1\gamma_1 + \phi_2\gamma_2 + \sigma^2$$

why? well, $\mathbb{E}[\varepsilon_t(Y_t-\mu)] = \mathbb{E}[\varepsilon_t(\phi_1(Y_{t-1}-\mu) + \phi_2(Y_{t-2}-\mu) + \varepsilon_t)] = 0 + 0 + \sigma^2$

$$\Rightarrow \gamma_0 = \phi_1 \rho_1 \gamma_0 + \phi_2 \rho_2 \gamma_0 + \sigma^2$$

But, $\rho_1 = \phi_1 + \phi_2\rho_1$ $\left(\text{or } \rho_1 = \frac{\phi_1}{1-\phi_2}\right)$ and $\rho_2 = \phi_1\rho_1 + \phi_2$, so

$$\gamma_0 = \left[\frac{\phi_1^2}{1-\phi_2} + \frac{\phi_2\phi_1^2}{1-\phi_2} + \phi_2^2\right]\gamma_0 + \sigma^2 \quad \text{or} \quad \gamma_0 = \frac{(1-\phi_2)\sigma^2}{(1+\phi_2)\left((1-\phi_2)^2 - \phi_1^2\right)}$$

What about stationary $\boxed{AR(p)}$?

Again, we have $\mu = c + \phi_1\mu + \ldots + \phi_p\mu \Rightarrow \mu = \frac{c}{1-\phi_1-\ldots-\phi_p}$. Now, one can jump to mean-adjusted form

Using $$Y_t - \mu = \phi_1(Y_{t-1}-\mu) + \phi_2(Y_{t-2}-\mu) + \ldots + \phi_p(Y_{t-p}-\mu) + \varepsilon_t \quad (**)$$

Multiply both sides of $(*)$ by $Y_{t-j}-\mu$ and take expectations!

$$\gamma_j = \phi_1\gamma_{j-1} + \phi_2\gamma_{j-2} + \ldots + \phi_p\gamma_{j-p} \quad \text{for } j = 1, 2, \ldots$$

and $$\gamma_0 = \phi_1\gamma_1 + \phi_2\gamma_2 + \ldots + \phi_p\gamma_p + \sigma^2 \quad \text{for } j = 0$$

It can also be shown that $(\gamma_0, \gamma_1, \ldots, \gamma_{p-1})$ is the first p elements of the 1^{st} column of the $p^2 \times p^2$ matrix

$$\sigma^2\left[I_{p^2} - (F \otimes F)\right]^{-1}$$

Kronecker product

For ACF, we get <u>YULE-WALKER EQUATIONS</u>: $\rho_i = \phi_1\rho_{i-1} + \phi_2\rho_{i-2} + \ldots + \phi_p\rho_{i-p}$, $i = 1, 2, \ldots$

1.4 ARMA(p, q) (Mixed Autoregressive Moving Average Process)

$ARMA(p,0) \equiv AR(p)$
$ARMA(0,q) \equiv MA(q)$

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \dots + \phi_p Y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \dots + \theta_q \varepsilon_{t-q}$$

Lag operator form: $(1-\phi_1 L - \phi_2 L^2 - \dots - \phi_p L^p) Y_t = c + (1+\theta_1 L + \theta_2 L^2 + \dots + \theta_q L^q)\varepsilon_t$.

Provided that the roots of $1-\phi_1 z - \phi_2 z^2 \dots - \phi_p z^p = 0$ lie outside the complex unit circle, we can write this further as $Y_t = \mu + \psi(L)\varepsilon_t$, where

$$\psi(L) = \frac{1+\theta_1 L + \theta_2 L^2 + \dots + \theta_q L^q}{1-\phi_1 L - \phi_2 L^2 - \dots - \phi_p L^p}, \quad \sum_{j=0}^{\infty} |\psi_j| < \infty, \text{ and } \quad \mu = \frac{c}{1-\phi_1 - \phi_2 - \dots - \phi_p}$$

Stationarity of an ARMA process depends entirely on the autoregressive parameters $(\phi_1, \dots, \phi_p)$ and not on the moving average parameters $(\theta_1, \dots, \theta_q)$

Mean-adjusted form: $Y_t - \mu = \phi_1(Y_{t-1}-\mu) + \phi_2(Y_{t-2}-\mu) + \dots + \phi_p(Y_{t-p}-\mu) + \varepsilon_t + \theta_1\varepsilon_{t-1} + \dots + \theta_q \varepsilon_{t-q}$ (***)

How do we find autocovariances? As before, multiply (***) by $(Y_{t-j}-\mu)$ and take expectation.

For $j = q+1, q+2, \dots$ we get $\gamma_j = \phi_1 \gamma_{j-1} + \phi_2 \gamma_{j-2} + \dots + \phi_p \gamma_{j-p}$

Thus, after q lags the autocovariances follow the AR(p) model.

For $j \le q$, we have correlation between $\theta_j \varepsilon_{t-j}$ and Y_{t-j}, which will result in very complex autocovariance behaviour for lags 1 through q (much more complex than for the AR(p) process)

1.5 Model identification

1.5.1 Estimation of the parameters of a stationary process

Suppose we have data $(X_1, \dots, X_T)$ from a stationary TS. We can estimate

- the mean by $\hat{\mu} = \frac{1}{T}\sum_{t=1}^{T} X_t$
- the autocovariance by $\hat{\gamma}_k = \frac{1}{T}\sum_{t=k+1}^{T} (X_t - \hat{\mu})(X_{t-k}-\hat{\mu})$ ← don't forget: covariance btw. X_t & X_{t-k} is independent of t.
- the autocorrelation by $\hat{\rho}_k = \hat{\gamma}_k / \hat{\gamma}_0$.

The plot of $\hat{\rho}_k$ versus k is known as the correlogram.

If it is known that $\mu = 0$, then γ_k is estimated by $\hat{\gamma}_k = \frac{1}{T}\sum_{t=k+1}^{T} X_t X_{t-k}$.

In defining $\hat{\gamma}_k$ we divide by T rather than by T−k. It does not really matter since T is usually large relative to k.

Note: Suppose that a stationary process $\{X_t\}$ has autocovariance function $\{\gamma_k\}$.

Then $var\left(\sum_{t=1}^{T} a_t X_t\right) = \sum_{t=1}^{T}\sum_{s=1}^{T} a_t a_s Cov(X_t, X_s) = \sum_{t=1}^{T}\sum_{s=1}^{T} a_t a_s \gamma_{|t-s|} \ge 0$

A sequence $\{\gamma_k\}$ for which this holds for every $T \geq 1$ and a set of constants $(a_1, \dots, a_T)$ is called a <u>nonnegative definite sequence</u>. Blochner's theorem states that $\{\gamma_k\}$ is a valid autocovariance function <u>iff</u> it is nonnegative definite.

Dividing by T rather than by $T-k$ in the definition of $\hat{\gamma}_k$ ensures that $\{\hat{\gamma}_k\}$ is nonnegative definite (and thus that it could be the autocovariance function of a stationary process).

1.5.2 Identifying a MA(q) process

The MA(q) process Y_t has $\rho_k = 0$ for all k, $|k| > q$. So, a diagnostic for MA(q) is that $|\hat{\rho}_k|$ drops to near zero beyond some threshold.

1.5.3 Identifying an AR(p) process

The AR(1) process has $\rho_k = \phi^k$, decaying exponentially. This can be difficult to recognize in the correlogram. Moreover, for the AR(p) process, the autocorrelation was even more complex: Yule-Walker equation $\rho_j = \phi_1 \rho_{j-1} + \phi_2 \rho_{j-2} + \dots + \phi_p \rho_{j-p}$, $j = 1, 2, \dots$

This leads to solving a p^{th} order difference equation and even more complex exponential decay (depending on the roots of the reverse characteristic equation $z^p - \phi_1 z^{p-1} - \phi_2 z^{p-2} - \dots - \phi_p = 0$).

There is a better measure for identifying the AR(p) process than the correlogram: <u>partial autocorrelation function (PACF)</u>.

It is based on the <u>one-step linear predictor</u> $\hat{Y}_{n+1|n}$ for Y_{n+1} based on a linear combination on n previous values $\hat{Y}_{n+1|n} = a_0 + a_1 Y_n + a_2 Y_{n-1} + \dots + a_n Y_1$.

Prediction/Forecasting will be covered in much more detail later.

Now, we take just a little detour.

1.5.3.1 Linear predictor $\hat{Y}_{n+h|n}$

Suppose $\{Y_t\}$ is a stationary process with mean μ and autocovariance function $\{\gamma_k\}$.

Goal: Predict Y_{n+h} given $Y_1, Y_2, \dots, Y_n$.

We use linear predictor $\hat{Y}_{n+h|n} = a_0 + a_1 Y_n + a_2 Y_{n-1} + \dots + a_n Y_1$.

We need to find $a_0, a_1, \dots, a_n$ so that the <u>mean squared error</u> (MSE)

$$S(a_0, a_1, \dots, a_n) = \mathbb{E}\left[(Y_{n+h} - \hat{Y}_{n+h|n})^2\right] \text{ is minimized.}$$

$$S(a_0, a_1, \dots, a_n) = \mathbb{E}\left[\left(Y_{n+h} - a_0 - a_1 Y_n - a_2 Y_{n-1} - \dots a_n Y_1\right)^2\right]$$

Take partial derivatives of $S \Rightarrow$

$$\frac{\partial S}{\partial a_i} = \mathbb{E}\left[\frac{\partial}{\partial a_i}\left(Y_{n+h} - a_0 - a_1 Y_n - a_2 Y_{n-1} - \dots - a_n Y_1\right)^2\right]$$

it's ok to swap (dominated convergence theorem)

For $i=0$, we get $\dfrac{\partial S}{\partial a_0} = \mathbb{E}\left[-2\left(Y_{n+h} - a_0 - a_1 Y_n - a_2 Y_{n-1} - \dots - a_n Y_1\right)\right] = 0$

$$\Rightarrow \frac{\partial S}{\partial a_0} = \mu - a_0 - a_1\mu - a_2\mu - \dots - a_n\mu = 0 \Rightarrow a_0 = \mu\left(1 - \sum_{j=1}^{n} a_j\right)$$

So, once we figure out $a_1, a_2, \dots, a_n$, we'll have a_0 as well.

For $i = 1, 2, \dots, n$

$$\frac{\partial S}{\partial a_i} = \mathbb{E}\left[-2 Y_{n+1-i}\left(Y_{n+h} - a_0 - a_1 Y_n - a_2 Y_{n-1} - \dots - a_n Y_1\right)\right] = 0.$$

$$\Rightarrow \mathbb{E}\left[Y_{n+1-i} Y_{n+h} - a_0 Y_{n+1-i} - a_1 Y_n Y_{n+1-i} - \dots - a_n Y_1 Y_{n+1-i}\right] = 0$$

$$\Rightarrow \gamma_{h+i-1} = \underbrace{a_0\mu - \mu^2 - \sum_{j=1}^{n} a_j \mu^2}_{\text{this is 0, since } a_0 = \mu - \mu\sum_{j=1}^{n} a_j} + \sum_{j=1}^{n} \gamma_{j-i} a_j$$

$$\Rightarrow \gamma_{h+i-1} = \sum_{j=1}^{n} \gamma_{j-i} a_j \quad \text{for } i = 1, 2, \dots, n$$

Using $\gamma_k = \gamma_{-k}$, we can rewrite this in a matrix form:

$$\underbrace{\begin{pmatrix} \gamma_0 & \gamma_1 & \cdots & \gamma_{n-1} \\ \gamma_1 & \gamma_0 & \cdots & \gamma_{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{n-1} & \gamma_{n-2} & \cdots & \gamma_0 \end{pmatrix}}_{\Gamma_n} \underbrace{\begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix}}_{\vec{a}_n} = \underbrace{\begin{pmatrix} \gamma_h \\ \gamma_{h+1} \\ \vdots \\ \gamma_{h+n-1} \end{pmatrix}}_{\vec{\gamma}_{h,n}}$$

prediction equation: $\boxed{\Gamma_n \vec{a}_n = \vec{\gamma}_{h,n}}$

So, the best h-step linear predictor is given by $\vec{a}_n = \Gamma_n^{-1} \vec{\gamma}_{h,n}$ and $a_0 = \mu - \mu \sum_{j=1}^{n} a_n$.

We'll worry about MSE later when we ~~talk~~ about forecasting in more detail

Example: AR(1) with zero mean ($\mu = 0$)

$Y_t = \phi Y_{t-1} + \varepsilon_t$, $|\phi| < 1$, $\varepsilon_t \sim WN(0, \sigma^2)$

We know that $\gamma_h = \dfrac{\phi^h \sigma^2}{1-\phi^2}$. Suppose we need one step ahead forecast with observations $\{Y_1, Y_2\}$

Then $\hat{Y}_{3|2} = a_1 Y_2 + a_2 Y_1$, $\gamma_0 = \frac{\sigma^2}{1-\phi^2}$, $\gamma_1 = \frac{\phi\sigma^2}{1-\phi^2}$, $\gamma_2 = \frac{\phi^2\sigma^2}{1-\phi^2}$.

Prediction equation is

$$\frac{\sigma^2}{1-\phi^2}\begin{pmatrix} 1 & \phi \\ \phi & 1 \end{pmatrix}\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} \phi \\ \phi^2 \end{pmatrix}\frac{\sigma^2}{1-\phi^2} \Rightarrow \left.\begin{matrix} a_1 + \phi a_2 = \phi \\ \phi a_1 + a_2 = \phi^2 \end{matrix}\right\} \Rightarrow \begin{matrix} a_1 = \phi \\ a_2 = 0. \end{matrix}$$

and $\hat{Y}_{3|2} = \phi Y_2$.

<u>Example</u> Forecasting $WN(0,\sigma^2)$, based on $\{Y_1, \dots, Y_n\}$

prediction equation is

$$\begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{pmatrix}\begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}\sigma^2 \Rightarrow \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \text{ so}$$

$\hat{Y}_{n+1|n} = 0$. In a WN process, information from past does not help determine the forecast. Makes sense.

1.5.3.2 The PACF

<u>Def.</u> The partial autocorrelation function is defined as:

$$\alpha_n = \begin{cases} 1 & , \text{if } n = 0 \\ \text{the last entry of } \vec{a_n} & , \text{if } n > 0. \end{cases}$$

where $\vec{a_n}$ is the solution to the prediction equation $\vec{a_n} = \Gamma_n^{-1}\vec{\gamma}_{1,n}$.

<u>Intuition:</u> For an AR(p) process with zero mean:

$$\underbrace{Y_t - \phi_1 Y_{t-1} - \phi_2 Y_{t-2} - \dots - \phi_{p-1} Y_{t-p+1}}_{\text{part of } Y_t \text{ which can't be described by } (Y_{t-1}, Y_{t-2}, \dots, Y_{t-p+1})} = \phi_p \underbrace{Y_{t-p}} + \varepsilon_t$$

PACF α_p describes the relationship between these two

<u>Example</u> <u>PACF for AR(p)</u> Let's find it!

Assume $\mu = 0$ for simplicity.

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \dots + \phi_{p-1} Y_{t-p+1} + \phi_p Y_{t-p} + \varepsilon_t$$

Choose some $r \geq p$, and rewrite:

$$Y_t - \phi_1 Y_{t-1} - \phi_2 Y_{t-2} - \dots - \phi_p Y_{t-p} - \phi_{p+1} Y_{t-p-1} - \dots - \phi_r Y_{t-r} = \varepsilon_t \text{ where}$$

$$\phi_{p+1} = \phi_{p+2} = \dots = \phi_r = 0.$$

Multiply both sides by Y_{t-j} $(j = 1, \dots, r)$ and take expectation:

$$\mathbb{E}\left[Y_tY_{t-j}-\phi_1Y_{t-1}Y_{t-j}-\phi_2Y_{t-2}Y_{t-j}-\dots-\phi_rY_{t-r}Y_{t-j}\right]=\mathbb{E}\left[\varepsilon_tY_{t-j}\right],\quad j=1,2,\dots,r$$

By common sense (or by Wold: $\mathbb{E}[\varepsilon_tY_{t-j}]=\mathbb{E}[\varepsilon_t\sum_{i=0}^{\infty}\psi_i\varepsilon_{t-j-i}]=\sum_{i=0}^{\infty}\psi_i\mathbb{E}[\varepsilon_t\varepsilon_{t-j-i}]=0$)

we have

$$\mathbb{E}\left[Y_tY_{t-j}-\phi_1Y_{t-1}Y_{t-j}-\phi_2Y_{t-2}Y_{t-j}-\dots-\phi_rY_{t-r}Y_{t-j}\right]=0,\quad j=1,2,\dots,r$$

$\Rightarrow \gamma_j-\phi_1\gamma_{j-1}-\phi_2\gamma_{j-2}-\dots-\phi_r\gamma_{j-r}=0$, $j=1,2,\dots,r$; or in the matrix form

$$\begin{pmatrix}\gamma_0 & \gamma_1 & \dots & \gamma_{r-1}\\ \gamma_1 & \gamma_0 & \dots & \gamma_{r-2}\\ \vdots & \vdots & \ddots & \vdots\\ \gamma_{r-1} & \gamma_{r-2} & \dots & \gamma_0\end{pmatrix}\begin{pmatrix}\phi_1\\ \phi_2\\ \vdots\\ \phi_r\end{pmatrix}=\begin{pmatrix}\gamma_1\\ \gamma_2\\ \vdots\\ \gamma_r\end{pmatrix}\quad\text{or}\quad\Gamma_r\begin{pmatrix}\phi_1\\ \phi_2\\ \vdots\\ \phi_r\end{pmatrix}=\vec{\gamma}_{1,r}$$

Now, by definition α_r is the last entry in $\begin{pmatrix}\phi_1\\ \phi_2\\ \vdots\\ \phi_r\end{pmatrix}$, so $\boxed{\alpha_r=\begin{cases}1, & r=0\\ \phi_r, & r\le p\\ 0, & r>p\end{cases}}$ PACF for the AR(P).

PUNCH:

→ PACF is a very nice diagnostic, since the cutoff point at <u>sample PACF</u> determines p.

What is sample PACF?

$$\widehat{\vec{a}_n}=\hat{\Gamma}_n^{-1}\widehat{\vec{\gamma}}_{1,n},\ \text{i.e.}\quad\begin{pmatrix}\hat{a}_1\\ \hat{a}_2\\ \vdots\\ \hat{a}_n\end{pmatrix}=\begin{pmatrix}\hat{\gamma}_0 & \hat{\gamma}_1 & \dots & \hat{\gamma}_{n-1}\\ \hat{\gamma}_1 & \hat{\gamma}_0 & \dots & \hat{\gamma}_{n-2}\\ \vdots & \vdots & \ddots & \vdots\\ \hat{\gamma}_{n-1} & \hat{\gamma}_{n-2} & \dots & \hat{\gamma}_0\end{pmatrix}^{-1}\begin{pmatrix}\hat{\gamma}_1\\ \hat{\gamma}_2\\ \vdots\\ \hat{\gamma}_n\end{pmatrix}$$

the last entry is the sample PACF $\hat{\alpha}_n$!

It can be shown that the PACF of a MA(q) process, α_k, is not zero for all k.

<u>Summary</u> Given the data $Y_1,\dots,Y_n$, we plot the sample ACF and the sample PACF.
A rule of thumb is that if $\hat{\rho}_k$ is negligible beyond some cutoff point q, then we decide to fit a MA(q) model to $\{Y_t\}$. If $\hat{\alpha}_k$ is negligible beyond some cutoff point p, then we decide to fit an AR(p) model to $\{Y_t\}$. What does "negligible" mean?

Both the sample ACF and PACF are approximately normally distributed about their population means, and have standard deviation of about $1/\sqrt{n}$, where n is the length of the sample from TS. A rule of thumb is that $\hat{\rho}_k$ (and similarly $\hat{\alpha}_k$) is negligible if it lies between $\pm 2/\sqrt{n}$. Here 2 is an approximation to 1.96. Recall that if $Z_1,\dots,Z_n\sim N(\mu,1)$, a test of size 0.05 of the hypothesis $H_0:\mu=0$ against $H_1:\mu\neq 0$ rejects H_0 if and only if $\bar{Z}$ lies outside $\pm 1.96/\sqrt{n}$.

Some care is needed in applying this rule of thumb. It is important to realise that the sample autocorrelations $\hat{\rho}_1, \hat{\rho}_2, \ldots$ (and the sample PACF $\hat{\alpha}_1, \hat{\alpha}_2, \ldots$) are not independent. The probability that any one $\hat{\rho}_k$ should lie outside $\pm 2/\sqrt{n}$ depends on the values of the other $\hat{\rho}_k$

Unfortunately, there is no easy diagnostics (such as ACF or PACF) for general ARMA(p,q) processes. We'll have to use different methods to identify these, which will be done later on. NOTE: An ARMA(p,q) process has $\hat{\rho}_k$ and $\hat{\alpha}_k$ decaying geometrically for $k > \max(p,q)$.

$\boxed{1.5.4}$ Identifying the white noise / Box-Jenkins modelling strategy

Box-Jenkins modeling strategy for fitting ARMA (p,q) models is as follows:

Step 1 Transform the data, if necessary, so that the assumption of weak stationarity is a reasonable one.
This involves detrending, removing seasonality, etc. (see Chapter 5)

Step 2. Make an initial guess for the values of p and/or q.
We saw two useful diagnostics for fitting MA(q) or AR(p), but no useful diagnostic yet for general ARMA(p,q) models. See 3.7. for ARMA(p,q).

Step 3. Estimate the parameters of the proposed ARMA (p,q) model.
This is done by maximum likelihood estimation (see Chapter 3)

Step 4. Perform diagnostic analysis to confirm that the proposed model adequately describes the data. We need to examine residuals from fitted model $\varepsilon_t = Y_t - \hat{Y}_t$, and test whether ε_t are white noise indeed. NOTE: Also, see overfitting in 3.6.

Step 5: If the residuals pass the white noise test, our fitted model $\hat{Y}_t$ is good.
Use this ARMA (p,q) model for forecasting the future. See Ch.2. for forecasting.
Invert all the transformations from Step 1 and obtain the forecast for the original TS.

Now let's talk about Step 4. "model checking". How do we test whether residuals $\varepsilon_t = Y_t - \hat{Y}_t$ are white noise?

Autocorrelation test for residuals

In the sample ACF or PACF for residuals, with 95% confidence, non-zero lags should only appear significantly different from zero one out [illegible]

Box-Pierce / Portmanteau test for residuals

Similar to the sample ACF $\hat{\rho}_k^Y$ for $\{Y_t\}$, if the residuals are i.i.d., then

$$\sqrt{n}\,\hat{\rho}_k^{\varepsilon} \sim N(0,1),$$ i.e. the sample ACF for ε_t is normally distributed with mean zero and variance $1/n$. Define statistic Q:

$$Q = n \sum_{k=1}^{h} \left(\hat{\rho}_k^{\varepsilon}\right)^2,$$ which has χ^2 distribution with h degrees of freedom

So, we reject $H_0 : \{\varepsilon_t\}$ are i.i.d. if

$$Q > \underbrace{\chi^2_{1-\alpha}}(h)$$

$(1-\alpha)$-quantile of χ^2 with h DOF.
α is the size of the test.

Ljung-Box test

It's based on the statistic $Q_{LB} = n(n+2) \sum_{k=1}^{h} \frac{\left(\hat{\rho}_k^{\varepsilon}\right)^2}{n-k}$

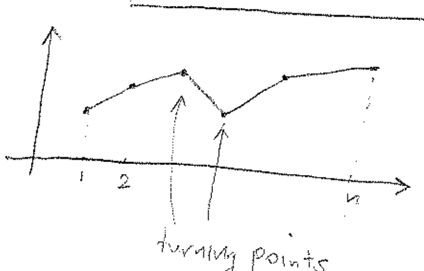
The distribution of Q_{LB} is better approximated by $\chi^2(h)$ than the Q-statistic above.

Question: How large should h be? The sensitivity of the test to departure from white noise depends on the choice of h. If the true model is ARMA(p,q) then the greatest power is obtained (rejection of the white noise hypothesis is most probable) when h is about $p+q$.

Turning point test for residuals

This one is probably the simplest.
If $\{\varepsilon_t\}$ is a sequence of residuals, we say that there is a turning point at time i, if one of the two conditions happens:

$$\boxed{\varepsilon_i > \varepsilon_{i+1} \text{ and } \varepsilon_i > \varepsilon_{i-1}} \quad \text{OR} \quad \boxed{\varepsilon_i < \varepsilon_{i+1} \text{ and } \varepsilon_i < \varepsilon_{i-1}}$$



Our statistic will be T: number of turning points

i.e. $T = \sum_{j=2}^{n-1} T_j$ where $T_j = \begin{cases} 1, & \text{if turning point @} j \\ 0, & \text{otherwise} \end{cases}$

$$\mathbb{E}[T] = \sum_{j=2}^{n-1} \mathbb{E}[T_j] = \sum_{j=2}^{n-1} \mathbb{P}(T_j = 1)$$

Claim $\mathbb{P}(T_j = 1) = 4/6$

proof: Consider $\varepsilon_{j-1}, \varepsilon_j, \varepsilon_{j+1}$ and possible orders in their realization:

$\square$

So, $\mathbb{E}[T] = \frac{4}{6}(n-2) = \frac{2}{3}(n-2)$

Also, it's easy to derive $\operatorname{var}(T) = \frac{16n-22}{90}$.

For large n, $T \sim N\left(\frac{2}{3}(n-2), \frac{16n-22}{90}\right)$, i.e. $T^* = \frac{\left|T - \frac{2}{3}(n-2)\right|}{\sqrt{\frac{16n-22}{90}}} \sim N(0,1)$

So, we reject $H_0 : \{\varepsilon_t\}$ is i.i.d. if $T^* > \Phi_{1-\frac{\alpha}{2}}$ ← $(1-\frac{\alpha}{2})$-quantile of standard normal distribution

2 FORECASTING

2.1 Basic principles (without proofs)

Setup: Want to forecast the value of a variable Y_{t+1} based on a set of variables X_t observed at date t. For example, we might want to forecast Y_{t+1} based on its m most recent values, in which case $X_t = \{Y_t, Y_{t-1}, \dots, Y_{t-m+1}, \text{constant}\}$

Notation: $Y^*_{t+1|t}$ denotes a forecast Y_{t+1} based on X_t.

We want to minimize a certain "loss function", which is usually the mean squared error (MSE):

$$MSE(Y^*_{t+1|t}) = \mathbb{E}\left[(Y_{t+1} - Y^*_{t+1|t})^2\right]$$

FACT The forecast with the smallest MSE turns out to be the expectation of Y_{t+1} conditional on X_t:

$$Y^*_{t+1|t} = \mathbb{E}[Y_{t+1} | X_t]$$

CAVEAT The computation of $\mathbb{E}[Y_{t+1}|X_t]$ depends on the distribution of $\{\varepsilon_t\}$ and may be a very complicated nonlinear function of the history of $\{\varepsilon_t\}$

So, what is usually done with forecasting of TS is linear prediction (we saw some of this in 1.5.3.1.)

We'll require the forecast $Y^*_{t+1|t}$ to be a linear function of X_t: $Y^*_{t+1|t} = \alpha^T X_t$. We need to find a value of α such that the forecast error $(Y_{t+1} - \alpha^T X_t)$ is uncorrelated with X_t, i.e. $\mathbb{E}[(Y_{t+1} - \alpha^T X_t) X_t^T] = 0^T$. $(*)$

If $(*)$ holds, then $\alpha^T X_t$ is called the linear projection of Y_{t+1} on X_t.

FACT Among all possible linear prediction, the linear projection of Y_{t+1} on X_t (which satisfies $(*)$) turns out to produce the smallest MSE.

Notation: The linear projection of Y_{t+1} on X_t is usually denoted by $\hat{Y}_{t+1|t}$

Properties of $\hat{Y}_{t+1|t}$: $(*)$ $\mathbb{E}[(Y_{t+1} - \alpha^T X_t) X_t^T] = 0^T \Rightarrow \mathbb{E}[Y_{t+1} X_t^T] = \alpha^T \mathbb{E}[X_t X_t^T]$

OR $\boxed{\alpha^T = \mathbb{E}[Y_{t+1} X_t^T] \left(\mathbb{E}[X_t X_t^T]\right)^{-1}}$ (assuming $\mathbb{E}[X_t X_t^T]$ is a nonsingular matrix)

so, the projection coefficients α^T are easy to find in terms of the moments of Y_{t+1} and X_t.

The MSE of $\hat{Y}_{t+1|t}$ is given by:

$$\mathbb{E}[(Y_{t+1} - \hat{Y}_{t+1|t})^2] = \mathbb{E}[(Y_{t+1} - \alpha^T X_t)^2] = \mathbb{E}[Y^2_{t+1}] - 2\mathbb{E}[\alpha^T X_t Y_{t+1}] + \mathbb{E}[\alpha^T X_t X_t^T \alpha] =$$

$$\overset{*}{=} \mathbb{E}[Y^2_{t+1}] - 2\mathbb{E}[Y_{t+1} X_t^T] \left(\mathbb{E}[X_t X_t^T]\right)^{-1} \mathbb{E}[X_t Y_{t+1}] + \mathbb{E}[Y_{t+1} X_t^T] \cancel{\left(\mathbb{E}[X_t X_t^T]\right)^{-1}} \cancel{\mathbb{E}[X_t X_t^T]} \left(\mathbb{E}[X_t X_t^T]\right)^{-1} \mathbb{E}[X_t Y_{t+1}]$$

$$= \mathbb{E}[Y^2_{t+1}] - \mathbb{E}[Y_{t+1} X_t^T] \left(\mathbb{E}[X_t X_t^T]\right)^{-1} \mathbb{E}[X_t Y_{t+1}]$$

So, the MSE is $\boxed{\mathbb{E}[(Y_{t+1}-\hat{Y}_{t+1|t})^2] = \mathbb{E}[Y_{t+1}^2] - \mathbb{E}[Y_{t+1}X_t^T](\mathbb{E}[X_tX_t^T])^{-1}\mathbb{E}[X_tY_{t+1}]}$

2.2. FORECASTS BASED ON AN INFINITE NUMBER OF OBSERVATIONS

2.2.1 Forecasts based on lagged ε's

Let $\{Y_t\}$ have a Wold representation $Y_t - \mu = \Psi(L)\varepsilon_t$, where $\Psi(L) = \sum_{j=0}^{\infty}\psi_j L^j$, $\psi_0 = 1$ and $\sum_{j=0}^{\infty}\psi_j^2 < \infty$

Suppose X_t is the infinite set $\{\varepsilon_t, \varepsilon_{t-1}, \varepsilon_{t-2}, \dots\}$

(Also, suppose we know values of μ and ψ_j, for all j. This is the topic of Step 3. Maximum likelihood estimation in Box-Jenkins strategy covered in Ch3).

We want a s-step forecast $\hat{Y}_{t+s|t}$.

First of all, from Wold representation: $Y_{t+s} = \mu + \varepsilon_{t+s} + \psi_1\varepsilon_{t+s-1} + \dots + \psi_{s-1}\varepsilon_{t+1} + \psi_s\varepsilon_t + \psi_{s+1}\varepsilon_{t-1} + \dots$

Then $\boxed{\hat{Y}_{t+s|t} = \mu + \psi_s\varepsilon_t + \psi_{s+1}\varepsilon_{t-1} + \psi_{s+2}\varepsilon_{t-2} + \dots}$ Why? Intuitively, the unknown future ε's are set to their expected values of zero. Formally, look at the error

$$Y_{t+s} - \hat{Y}_{t+s|t} = \varepsilon_{t+s} + \psi_1\varepsilon_{t+s-1} + \dots + \psi_{s-1}\varepsilon_{t+1}$$ which is clearly uncorrelated with $\varepsilon_t, \varepsilon_{t-1}, \dots$, i.e. with X_t; hence $(*)$ holds, and $\hat{Y}_{t+s|t}$ given by above formula is indeed a linear projection.

MSE: $\mathbb{E}[(Y_{t+s} - \hat{Y}_{t+s|t})^2] = (1 + \psi_1^2 + \psi_2^2 + \dots + \psi_{s-1}^2)\sigma^2$

Example: Forecasting the MA(q) process based on lagged ε's.

For MA(q): $\Psi(L) = 1 + \theta_1 L + \theta_2 L^2 + \dots + \theta_q L^q$

$$\hat{Y}_{t+s|t} = \begin{cases} \mu + \theta_s\varepsilon_t + \theta_{s+1}\varepsilon_{t-1} + \theta_{s+2}\varepsilon_{t-2} + \dots + \theta_q\varepsilon_{t-q+s}, & \text{for } s = 1, 2, \dots, q \\ \mu, & \text{for } s > q \end{cases}$$

the MSE is $$\begin{cases} \sigma^2 & s = 1 \\ (1 + \theta_1^2 + \dots + \theta_{s-1}^2)\sigma^2, & s = 2, 3, \dots, q \\ (1 + \theta_1^2 + \dots + \theta_q^2)\sigma^2, & s > q \end{cases}$$

Common sense check: If we try to forecast MA(q) further than q periods in the future, the forecast is simply the unconditional mean $\mathbb{E}[Y_t] = \mu$ and the MSE is the unconditional variance $var(Y_t) = (1 + \theta_1^2 + \dots + \theta_q^2)\sigma^2$.

Some new notation $[\ ]_+$

Annihilation operator: $\left[\frac{\Psi(L)}{L^s}\right]_+ = \psi_s + \psi_{s+1}L + \psi_{s+2}L^2 + \dots$, i.e. remove all negative power terms after dividing by L^s

s-step forecast becomes

$$\hat{Y}_{t+s|t} = \mu + \left[\frac{\Psi(L)}{L^s}\right]_+ \varepsilon_t$$

2.2.2 Forecasts based on lagged Y's

What didn't make much practical sense in the previous sense is that ε_t is not observed directly in practice. In the usual forecasting situation, we actually have observations on lagged Y's, not lagged ε's. So, now let $X_t = \{Y_t, Y_{t-1}, Y_{t-2}, \dots\}$

Idea is very simple. Invert MA(∞) (i.e. Wold) representation $Y_t - \mu = \psi(L)\varepsilon_t$ into an AR(∞) representation $\psi(L)^{-1}(Y_t - \mu) = \varepsilon_t$.

The s-step forecast formula becomes $\hat{Y}_{t+s|t} = \mu + \left[\frac{\psi(L)}{L^s}\right]_+ \psi(L)^{-1}(Y_t - \mu)$.

This is known as the Wiener-Kolmogorov prediction formula (WKP)

Example 1 AR(1)

$$(1-\phi L)(Y_t - \mu) = \varepsilon_t \quad \Rightarrow \psi(L) = \frac{1}{1-\phi L} = 1 + \phi L + \phi^2 L^2 + \phi^3 L^3 + \dots$$

$$\left[\frac{\psi(L)}{L^s}\right]_+ = \phi^s + \phi^{s+1} L + \phi^{s+2} L^2 + \phi^{s+3} L^3 + \dots = \frac{\phi^s}{1-\phi L}$$

So, from (WKP) $\hat{Y}_{t+s|t} = \mu + \frac{\phi^s}{1-\phi L} \cdot (1-\phi L)(Y_t - \mu)$, i.e.

$$\hat{Y}_{t+s|t} = \mu + \phi^s (Y_t - \mu).$$

Here $\psi_j = \phi^j$, so the MSE is $(1 + \phi^2 + \phi^4 + \dots + \phi^{2s-2})\sigma^2$. Note that

$\lim_{s\to\infty} \hat{Y}_{t+s|t} = \mu = E[Y_t]$ and $MSE \to \frac{\sigma^2}{1-\phi^2} = Var(Y_t)$, since $|\phi| < 1$

Example 2 AR(p)

There are two ways to forecast the AR(p) process based on the infinitely many observations $Y_t, Y_{t-1}, Y_{t-2}, \dots$

Look back to 1.3.4. We had from the state space model form

$$\xi_{t+s} = F^s \xi_t + F^{s-1} V_{t+1} + F^{s-2} V_{t+2} + \dots + F V_{t+s-1} + V_{t+s} \quad \text{where}$$

$$\xi_t = \begin{pmatrix} Y_t - \mu \\ Y_{t-1} - \mu \\ \vdots \\ Y_{t-p+1} - \mu \end{pmatrix}, \quad F = \begin{pmatrix} \phi_1 & \phi_2 & \cdots & \phi_p \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots 1 & 0 \end{pmatrix} \quad \text{and} \quad V_t = \begin{pmatrix} \varepsilon_t \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

So, from the first row, we get:

$$Y_{t+s} - \mu = (F^s)_{1,1}(Y_t - \mu) + (F^s)_{1,2}(Y_{t-1} - \mu) + \dots + (F^s)_{1,p}(Y_{t-p+1} - \mu) + \varepsilon_{t+s} + \psi_1 \varepsilon_{t+s-1} + \dots + \psi_{s-1}\varepsilon_{t+1}$$

where $\psi_1 = (F)_{1,1}$, $\psi_2 = (F^2)_{1,1}$, $\dots$, $\psi_{s-1} = (F^{s-1})_{1,1}$

The optimal s-step forecast is thus:

$$\hat{Y}_{t+s|t} = \mu + (F^s)_{1,1}(Y_t - \mu) + (F^s)_{1,2}(Y_{t-1} - \mu) + \ldots + (F^s)_{1,p}(Y_{t-p+1} - \mu) \quad (**)$$

The forecast error is $Y_{t+s} - \hat{Y}_{t+s|t} = \varepsilon_{t+s} + \psi_1 \varepsilon_{t+s-1} + \ldots + \psi_{s-1}\varepsilon_{t+1}$

whose expectation is clearly 0.

The alternative way to calculate (**) is through a <u>principle of iterated projections</u>.

Suppose that at date t we wish to make a one-period ahead forecast of Y_{t+1}. The optimal forecast is clearly $\left(\hat{Y}_{t+1|t} - \mu\right) = \phi_1(Y_t - \mu) + \phi_2(Y_{t-1} - \mu) + \ldots + \phi_p(Y_{t-p+1} - \mu) \quad (1)$

Consider next a two-step forecast. Suppose that at date t+1 we were to make a one-step forecast of Y_{t+2}. Replacing t by t+1, we get that the optimal forecast is

$$\left(\hat{Y}_{t+2|t+1} - \mu\right) = \phi_1(Y_{t+1} - \mu) + \phi_2(Y_t - \mu) + \ldots + \phi_p(Y_{t-p+2} - \mu) \quad (2)$$

The law of iterated projections asserts that <u>if this date t+1 forecast of Y_{t+2} is projected on date t information, the result is the date t forecast of Y_{t+2}</u>. At date t values of $Y_t, Y_{t-1}, \ldots, Y_{t-p+2}$ are known, so we get:

$$\left(\hat{Y}_{t+2|t} - \mu\right) = \phi_1\left(\hat{Y}_{t+1|t} - \mu\right) + \phi_2(Y_t - \mu) + \ldots + \phi_p(Y_{t-p+2} - \mu) \quad (3)$$

Now substitute (1) into (3) to get a two-step forecast formula for an AR(p) process.

$$\left(\hat{Y}_{t+2|t} - \mu\right) = \phi_1\left[\phi_1(Y_t - \mu) + \phi_2(Y_{t-1} - \mu) + \ldots + \phi_p(Y_{t-p+1} - \mu)\right] + \phi_2(Y_t - \mu) + \ldots + \phi_p(Y_{t-p+2} - \mu)$$

i.e. $\left(\hat{Y}_{t+2|t} - \mu\right) = (\phi_1^2 + \phi_2)(Y_t - \mu) + (\phi_1\phi_2 + \phi_3)(Y_{t-1} - \mu) + \ldots + (\phi_1\phi_{p-1} + \phi_p)(Y_{t-p+2} - \mu) + \phi_1\phi_p(Y_{t-p+1} - \mu)$

In general, the s-step forecasts for an AR(p) process can be obtained from iterating on:

$$\left(\hat{Y}_{t+j|t} - \mu\right) = \phi_1\left(\hat{Y}_{t+j-1|t} - \mu\right) + \phi_2\left(\hat{Y}_{t+j-2|t} - \mu\right) + \ldots + \phi_p\left(\hat{Y}_{t+j-p|t} - \mu\right)$$

for $j = 1, 2, \ldots, s$

where $\hat{Y}_{\tau|t} = Y_\tau$ for $\tau \le t$.

<u>Example 3.</u> <u>MA(1) process</u>

invertible MA(1) process $Y_t - \mu = (1 + \theta L)\varepsilon_t$, $|\theta| < 1$

so $\psi(L) = 1 + \theta L$

(WKP) $\Rightarrow \hat{Y}_{t+s|t} = \mu + \left[\frac{1+\theta L}{L^s}\right]_+ \frac{1}{1+\theta L}(Y_t - \mu)$

For one-step forecast (s=1) we get $\left[\frac{1+\theta L}{L}\right]_+ = \theta$, so

(***) $\hat{Y}_{t+1|t} = \mu + \frac{\theta}{1+\theta L}(Y_t - \mu) = \mu + \theta(Y_t - \mu) - \theta^2(Y_{t-1} - \mu) + \theta^3(Y_{t-2} - \mu) - \dots$

Next, we introduce a useful notation which will enable us to write (***)
(and s-step forecast for MA(q) and ARMA(p,q) series) in a more concise form)

Rewrite $Y_t - \mu = (1+\theta L)\varepsilon_t$ as $\varepsilon_t = \frac{1}{1+\theta L}(Y_t - \mu)$ and view ε_t as the outcome of the infinite recursion $\hat{\varepsilon}_t = (Y_t - \mu) - \theta\hat{\varepsilon}_{t-1}$. Formula (***) can then be rewritten as $\hat{Y}_{t+1|t} = \mu + \theta\hat{\varepsilon}_t$

Note: The idea is to use the recursion $\hat{\varepsilon}_t = (Y_t - \mu) - \theta\hat{\varepsilon}_{t-1}$ to generate approximations $\hat{\varepsilon}_t$ of the white noise; then plug these into $\hat{Y}_{t+1|t} = \mu + \theta\hat{\varepsilon}_t$ to obtain forecasts.

Example 4. MA(q) (invertible)

$$Y_t - \mu = (1 + \theta_1 L + \theta_2 L^2 + \dots + \theta_q L^q)\varepsilon_t$$

(WKP) $\Rightarrow$ $\hat{Y}_{t+s|t} = \mu + \left[\frac{1+\theta_1 L + \theta_2 L^2 + \dots + \theta_q L^q}{L^s}\right]_+ \cdot \frac{1}{1+\theta_1 L + \dots + \theta_q L^q}(Y_t - \mu)$

Now, $$\left[\frac{1+\theta_1 L + \theta_2 L^2 + \dots + \theta_q L^q}{L^s}\right]_+ = \begin{cases} \theta_s + \theta_{s+1} L + \theta_{s+2} L^2 + \dots + \theta_q L^{q-s} & \text{for } s = 1, 2, \dots, q \\ 0 & \text{for } s = q+1, q+2, \dots \end{cases}$$

So, the formula for the s-step forecast is:

$$\hat{Y}_{t+s|t} = \mu + (\theta_s + \theta_{s+1} L + \dots + \theta_q L^{q-s})\hat{\varepsilon}_t \quad , s = 1, 2, \dots, q$$

where $\hat{\varepsilon}_t = (Y_t - \mu) - \theta_1\hat{\varepsilon}_{t-1} - \theta_2\hat{\varepsilon}_{t-2} - \dots - \theta_q\hat{\varepsilon}_{t-q}$.

A forecast farther than q steps into the future is simply the unconditional mean μ.

Example 5. ARMA(1,1) (stationary $|\phi| < 1$ and invertible $|\theta| < 1$)

$$(1 - \phi L)(Y_t - \mu) = (1 + \theta L)\varepsilon_t$$

(WKP) $\Rightarrow$ $\hat{Y}_{t+s|t} = \mu + \left[\frac{1+\theta L}{(1-\phi L)L^s}\right]_+ \cdot \frac{1-\phi L}{1+\theta L}(Y_t - \mu)$

Now, $$\left[\frac{1+\theta L}{(1-\phi L)L^s}\right]_+ = \left[\frac{(1+\phi L + \phi^2 L^2 + \dots)}{L^s} + \frac{\theta L(1 + \phi L + \phi^2 L^2 + \dots)}{L^s}\right]_+ =$$

$$= (\phi^s + \phi^{s+1} L + \phi^{s+2} L^2 + \dots) + \theta(\phi^{s-1} + \phi^s L + \phi^{s+1} L^2 + \dots) =$$

$$= (\phi^s + \theta\phi^{s-1})(1 + \phi L + \phi^2 L^2 + \dots) = \frac{\phi^s + \theta\phi^{s-1}}{1-\phi L}$$

So, $\hat{Y}_{t+s|t} = \mu + \dfrac{\phi^s + \theta\phi^{s-1}}{1+\theta L}(Y_t - \mu)$

Notice that for $s=2,3,\dots$ this formula obeys the recursion $\boxed{(\hat{Y}_{t+s|t} - \mu) = \phi(\hat{Y}_{t+s-1|t} - \mu)}$, so beyond one step in the future, the forecast decays geometrically at the rate ϕ toward the unconditional mean μ. Look at $s=1$.

$$\hat{Y}_{t+1|t} = \mu + \frac{\phi+\theta}{1+\theta L}(Y_t - \mu) \quad \text{or equivalently}$$

$$\hat{Y}_{t+1|t} - \mu = \frac{\phi(1+\theta L) + \theta(1-\phi L)}{1+\theta L}(Y_t - \mu), \text{ i.e. } \boxed{\hat{Y}_{t+1|t} - \mu = \phi(Y_t - \mu) + \theta\hat{\varepsilon}_t} \quad (4)$$

where $\hat{\varepsilon}_t = \dfrac{1-\phi L}{1+\theta L}(Y_t - \mu)$ or

$$\hat{\varepsilon}_t = (1-\phi L)(1-\theta L + \theta^2 L^2 - \theta^3 L^3 + \dots)(Y_t - \mu) \quad \text{or}$$

$$\hat{\varepsilon}_t = (Y_t - \mu) - \phi L(Y_t - \mu) - \theta L(1 - \theta L + \theta^2 L^2 - \dots)(1-\phi L)(Y_t - \mu) \quad \text{or}$$

$$\hat{\varepsilon}_t = (Y_t - \mu) - \phi(Y_{t-1} - \mu) - \theta L \cdot \frac{1-\phi L}{1+\theta L}(Y_t - \mu) \quad \text{or}$$

$$\hat{\varepsilon}_t = (Y_t - \mu) - \phi(Y_{t-1} - \mu) - \theta \cdot \frac{1-\phi L}{1+\theta L}(Y_{t-1} - \mu) \quad \text{or}$$

$$\hat{\varepsilon}_t = (Y_t - \mu) - \phi(Y_{t-1} - \mu) - \theta\hat{\varepsilon}_{t-1} \quad \text{or}$$

$$\hat{\varepsilon}_t = (Y_t - \mu) - (\hat{Y}_{t|t-1} - \mu) \quad \text{or finally} \quad \boxed{\hat{\varepsilon}_t = Y_t - \hat{Y}_{t|t-1}} \quad (5)$$

recursion for $\hat{\varepsilon}_t$.

<u>Example 6.</u> <u>ARMA (p,q)</u> (stationary and invertible)

$$(1 - \phi_1 L - \phi_2 L^2 - \dots - \phi_p L^p)(Y_t - \mu) = (1 + \theta_1 L + \theta_2 L^2 + \dots + \theta_q L^q)\varepsilon_t$$

Similarly to example 5. one can derive generalizations of (4) and (5).

<u>1-step forecast</u>: $\boxed{(\hat{Y}_{t+1|t} - \mu) = \phi_1(Y_t - \mu) + \phi_2(Y_{t-1} - \mu) + \dots + \phi_p(Y_{t-p+1} - \mu) + \theta_1\hat{\varepsilon}_t + \theta_2\hat{\varepsilon}_{t-1} + \dots + \theta_q\hat{\varepsilon}_{t-q+1}}$,

where $\boxed{\hat{\varepsilon}_t = Y_t - \hat{Y}_{t|t-1}}$

<u>s-step forecasts</u>

$$(\hat{Y}_{t+s|t} - \mu) = \phi_1(\hat{Y}_{t+s-1|t} - \mu) + \phi_2(\hat{Y}_{t+s-2|t} - \mu) + \dots + \phi_p(\hat{Y}_{t+s-p|t} - \mu) + \theta_s\hat{\varepsilon}_t + \theta_{s+1}\hat{\varepsilon}_{t-1} + \dots + \theta_q\hat{\varepsilon}_{t+s-q} \quad \text{for } s = 1,2,\dots,q$$

and

$$(\hat{Y}_{t+s|t} - \mu) = \phi_1(\hat{Y}_{t+s-1|t} - \mu) + \phi_2(\hat{Y}_{t+s-2|t} - \mu) + \dots + \phi_p(\hat{Y}_{t+s-p|t} - \mu) \quad \text{for } s > q.$$

where $\hat{Y}_{\tau|t} = Y_\tau$ for $\tau \le t$.

2.3. FORECASTS BASED ON A FINITE NUMBER OF OBSERVATIONS

In 2.2 we assumed that we had an infinite number of past observations $\{Y_t, Y_{t-1}, \ldots\}$ and knew with certainty population parameters such as μ, ϕ and θ.
We'll still assume we know the population parameters (how to find those is the topic of Section 3 and MLE's). However, we'll develop here methods for forecasting based on a finite number of observations $\{Y_t, Y_{t-1}, \ldots, Y_{t-m+1}\}$, which is what happens in practice.
Notice that for AR(p) models, an optimal s-step forecast formulae based on an infinite number of observations $\{Y_t, Y_{t-1}, \ldots\}$ in fact makes use of only the p most recent values $\{Y_t, Y_{t-1}, \ldots, Y_{t-p+1}\}$. Hence, the formulae from 2.2. are still used for AR(p) processes.
However, for an MA or ARMA series, we need new formulae.

2.3.1 Approximations to optimal forecasts

Idea: Assume presample ε's are all equal to 0, i.e. $\varepsilon_{t-m}=0$, $\varepsilon_{t-m-1}=0, \ldots$

example MA(q)

From example 4. in 2.2.2. we have:

$$\hat{Y}_{t+s|t} = \mu + \theta_s \hat{\varepsilon}_t + \theta_{s+1} \hat{\varepsilon}_{t-1} + \ldots + \theta_q \hat{\varepsilon}_{t-q+s}, \quad \text{for } s=1,2,\ldots,q$$

and $\hat{Y}_{t+s|t} = \mu$ for $s = q+1, q+2, \ldots$

where $\hat{\varepsilon}_t = (Y_t - \mu) - \theta_1 \hat{\varepsilon}_{t-1} - \theta_2 \hat{\varepsilon}_{t-2} - \ldots - \theta_q \hat{\varepsilon}_{t-q}$. ← infinite recursion which "needs" q values to get started

This last recursion for $\hat{\varepsilon}_t$'s is then started by setting

$$\hat{\varepsilon}_{t-m} = \hat{\varepsilon}_{t-m-1} = \ldots = \hat{\varepsilon}_{t-m-q+1} = 0.$$

Then we generate $\hat{\varepsilon}_{t-m+1}, \hat{\varepsilon}_{t-m+2}, \ldots, \hat{\varepsilon}_t$ by iterating the recursion:

$$\hat{\varepsilon}_{t-m+1} = Y_{t-m+1} - \mu$$

$$\hat{\varepsilon}_{t-m+2} = (Y_{t-m+2} - \mu) - \theta_1 \hat{\varepsilon}_{t-m+1}$$

$$\hat{\varepsilon}_{t-m+3} = (Y_{t-m+3} - \mu) - \theta_1 \hat{\varepsilon}_{t-m+2} - \theta_2 \hat{\varepsilon}_{t-m+1} \quad \text{and so on.}$$

The resulting values for $(\hat{\varepsilon}_t, \hat{\varepsilon}_{t-1}, \ldots, \hat{\varepsilon}_{t-q+s})$ are then substituted directly into the above formula for $\hat{Y}_{t+s|t}$.

In practice, for m large and $|\theta|$ small, these approximations are very good.
If $|\theta|$ is close to one, [illegible]

2.3.2 Exact formula for finite sample forecast

We have already done this in 1.5.3.1 where we needed a PACF for an AR(p) process.
It also follows from the formula for α^T in 2.1. for $X_t = (Y_t-\mu, Y_{t-1}-\mu, \dots, Y_{t-m+1}-\mu)^T$
Observations: $Y_t, Y_{t-1}, \dots Y_{t-m+1}$ for a stationary process $\{Y_t\}$ with mean μ and auto-cov. function $\{\gamma_k\}$

s-step ahead forecast:

$$\boxed{\hat{Y}_{t+s|t} = \mu + \alpha_1^{(m,s)}(Y_t-\mu) + \alpha_2^{(m,s)}(Y_{t-1}-\mu) + \dots + \alpha_m^{(m,s)}(Y_{t-m+1}-\mu)}$$

where

$$\boxed{\begin{pmatrix} \alpha_1^{(m,s)} \\ \alpha_2^{(m,s)} \\ \vdots \\ \alpha_m^{(m,s)} \end{pmatrix} = \begin{pmatrix} \gamma_0 & \gamma_1 & \dots & \gamma_{m-1} \\ \gamma_1 & \gamma_0 & \dots & \gamma_{m-2} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{m-1} & \gamma_{m-2} & \dots & \gamma_0 \end{pmatrix}^{-1} \begin{pmatrix} \gamma_s \\ \gamma_{s+1} \\ \vdots \\ \gamma_{s+m-1} \end{pmatrix}}$$

In practice, this is not so easy to use, since we need to invert an $m \times m$ matrix.
One usually uses some kind of factorization for this positive definite symmetric matrix, such as the Cholesky factorization.

3 PARAMETER ESTIMATION

In the previous chapters, we assumed that the population parameters such as $c, \phi_1, \phi_2, \dots, \phi_p, \theta_1, \theta_2, \dots, \theta_q, \sigma^2$ were known and then we showed how covariances and forecasts could be calculated as functions of these parameters. In this chapter we explore how to estimate the values of $c, \phi_1, \phi_2, \dots, \phi_p, \theta_1, \theta_2, \dots, \theta_q, \sigma^2$ on the basis of observations on Y.

3.1. Yule-Walker Equations for the AR(p) process

Consider an AR(p) process. Recall from 1.3.4 the Yule-Walker equations

$$\gamma_0 - \phi_1\gamma_1 - \phi_2\gamma_2 - \dots - \phi_p\gamma_p = \sigma^2$$

$$\begin{pmatrix} \gamma_0 & \gamma_1 & \dots & \gamma_{p-1} \\ \gamma_1 & \gamma_0 & \dots & \gamma_{p-2} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{p-1} & \gamma_{p-2} & \dots & \gamma_0 \end{pmatrix} \begin{pmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_p \end{pmatrix} = \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_p \end{pmatrix}$$

So, the parameters could be obtained by:

$$\begin{pmatrix} \hat\phi_1 \\ \hat\phi_2 \\ \vdots \\ \hat\phi_p \end{pmatrix} = \begin{pmatrix} \hat\gamma_0 & \hat\gamma_1 & \dots & \hat\gamma_{p-1} \\ \hat\gamma_1 & \hat\gamma_0 & \dots & \hat\gamma_{p-2} \\ \vdots & \vdots & \ddots & \vdots \\ \hat\gamma_{p-1} & \hat\gamma_{p-2} & \dots & \hat\gamma_0 \end{pmatrix}^{-1} \begin{pmatrix} \hat\gamma_1 \\ \hat\gamma_2 \\ \vdots \\ \hat\gamma_p \end{pmatrix} \quad \text{and} \quad \hat\gamma_0 - \hat\phi_1\hat\gamma_1 - \hat\phi_2\hat\gamma_2 - \dots - \hat\phi_p\hat\gamma_p = \hat\sigma^2$$

3.1.1. Limitations of Yule-Walker approach

Consider an MA(1) process (with mean $\mu = 0$) : $Y_t = \varepsilon_t + \theta_1\varepsilon_{t-1}$

Multiply both sides by Y_{t-j} ($j = 0, 1$) and take expectations:

$$\mathbb{E}[Y_tY_{t-j}] = \mathbb{E}[\varepsilon_tY_{t-j} + \theta_1\varepsilon_{t-1}Y_{t-j}] \Rightarrow$$

$$\Rightarrow \gamma_j = \mathbb{E}[\varepsilon_tY_{t-j}] + \theta_1\mathbb{E}[\varepsilon_{t-1}Y_{t-j}] \qquad j = 0, 1 \Rightarrow$$

$$\Rightarrow \gamma_0 = \mathbb{E}[\varepsilon_tY_t] + \theta_1\mathbb{E}[\varepsilon_{t-1}Y_t] \Rightarrow \gamma_0 = \sigma^2 + \theta_1^2\sigma^2 \quad \text{(here we use } Y_t = \varepsilon_t + \theta_1\varepsilon_{t-1}\text{)}$$

$$\gamma_1 = \mathbb{E}[\varepsilon_tY_{t-1}] + \theta_1\mathbb{E}[\varepsilon_{t-1}Y_{t-1}] \Rightarrow \gamma_1 = 0 + \theta_1\sigma^2 \quad \text{(here we use } Y_{t-1} = \varepsilon_{t-1} + \theta_1\varepsilon_{t-2}\text{)}$$

So, $\left.\begin{array}{l}\gamma_0=\sigma^2+\theta_1^2\sigma^2\\ \gamma_1=\theta_1\sigma^2\end{array}\right\}$ this is a non-linear system of equation, which indicates that Yule-Walker approach is <u>impractical</u> for $MA(q)$ and $ARMA(p,q)$ with $q>0$.

3.2. Maximum Likelihood estimation (MLE)

- less estimation bias
- less estimation standard error
- can be used for $MA(q)$ and $ARMA(p,q)$

However, it's more computationally intensive.

This approach also requires specifying particular distribution for the white noise ε_t

So, assume that ε_t is Gaussian WN, i.e. $\varepsilon_t \sim$ i.i.d. $N(0,\sigma^2)$

2 steps: (1) calculate the <u>likelihood function</u>

$$f_{Y_T,Y_{T-1},\ldots Y_1}(y_T,y_{T-1},\ldots,y_1;\vec{\theta})$$ where

$(y_1,y_2,\ldots,y_T)$ is an observed sample of size T and

$\vec{\theta}=(c,\phi_1,\phi_2,\ldots,\phi_p,\theta_1,\theta_2,\ldots,\theta_q,\sigma^2)^T$ is the <u>parameter vector</u> of an $ARMA(p,q)$ process: $Y_t=c+\phi_1Y_{t-1}+\ldots+\phi_pY_{t-p}+\varepsilon_t+\theta_1\varepsilon_{t-1}+\ldots+\theta_q\varepsilon_{t-q}$

$\varepsilon_t \sim$ i.i.d. $N(0,\sigma^2)$

(2) find the values of $\vec{\theta}$ that maximize this function.

3.2.1. AR(1) process

$$Y_t=c+\phi Y_{t-1}+\varepsilon_t \qquad \vec{\theta}=(c,\phi,\sigma^2)^T$$

Consider the prob. distr. of the first observation Y_1 in our sample $(Y_1,\ldots,Y_T)$.

We know $\mathbb{E}[Y_1]=\mu=\frac{c}{1-\phi}$ and $Var(Y_1)=\mathbb{E}[(Y_1-\mu)^2]=\frac{\sigma^2}{1-\phi^2}$

Since $\{\varepsilon_t\}_{t=-\infty}^{+\infty}$ is Gaussian, then Y_1 is also Gaussian. Hence, the density of Y_1 is given by

$$f_{Y_1}(y_1;\vec{\theta})=\frac{1}{\sqrt{2\pi}\sqrt{\sigma^2/(1-\phi^2)}}\,e^{\frac{-(y_1-[c/(1-\phi)])^2}{2\sigma^2/(1-\phi^2)}}$$

Next, consider the density of Y_2, conditional on observing $Y_1=y_1$.

Now, $Y_2=c+\phi Y_1+\varepsilon_2 \Rightarrow (Y_2|Y_1=y_1)\sim N(c+\phi y_1,\sigma^2)$

So, $f_{Y_2|Y_1}(y_2|y_1;\vec{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_2-c-\phi_1 y_1)^2}{2\sigma^2}}$

The joint density of the first two observations is just the product of the two above:

$$f_{Y_2,Y_1}(y_2,y_1;\vec{\theta}) = f_{Y_2|Y_1}(y_2|y_1;\vec{\theta}) \cdot f_{Y_1}(y_1;\vec{\theta}).$$

Similarly, we get

$$f_{Y_t|Y_{t-1},\dots,Y_1}(y_t|y_{t-1},\dots,y_1;\vec{\theta}) = f_{Y_t|Y_{t-1}}(y_t|y_{t-1};\vec{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_t-c-\phi_1 y_{t-1})^2}{2\sigma^2}}$$

Since the values of $Y_1,\dots,Y_{t-1}$ matter for Y_t only through the value of Y_{t-1}.

The joint density of the first t observations is then

$$f_{Y_t,Y_{t-1},\dots,Y_1}(y_t,y_{t-1},\dots,y_1;\vec{\theta}) = f_{Y_t|Y_{t-1}}(y_t|y_{t-1};\vec{\theta}) \cdot f_{Y_{t-1},Y_{t-2},\dots,Y_1}(y_{t-1},\dots,y_1;\vec{\theta})$$

The likelihood function of the complete sample is:

$$f_{Y_T,Y_{T-1},\dots,Y_1}(y_T,y_{T-1},\dots,y_1;\vec{\theta}) = f_{Y_1}(y_1;\vec{\theta}) \cdot \prod_{t=2}^{T} f_{Y_t|Y_{t-1}}(y_t|y_{t-1};\vec{\theta})$$

Often, when maximizing, we work with the <u>LOG LIKELIHOOD FUNCTION</u> $\mathcal{L}(\vec{\theta})$.

So, for AR(1) process, $\mathcal{L}(\vec{\theta}) = \log f_{Y_1}(y_1;\vec{\theta}) + \sum_{t=2}^{T} \log f_{Y_t|Y_{t-1}}(y_t|y_{t-1};\vec{\theta})$, i.e.

$$\boxed{\begin{aligned}\mathcal{L}(\vec{\theta}) = &-\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\left(\frac{\sigma^2}{1-\phi^2}\right) - \frac{(y_1-[c/(1-\phi)])^2}{2\sigma^2/(1-\phi^2)} - \left(\frac{T-1}{2}\right)\cdot\log 2\pi - \left(\frac{T-1}{2}\right)\log\sigma^2 - \\ (*) \quad &- \sum_{t=2}^{T}\left[\frac{(y_t-c-\phi y_{t-1})^2}{2\sigma^2}\right].\end{aligned}}$$

<u>ALTERNATIVE WAY OF DERIVING (*)</u>

Let $\vec{y} = (y_1, y_2, \dots, y_T)^T$ ← vector of observations

$\vec{Y} = (Y_1, Y_2, \dots, Y_T)^T$ ← T-dimensional Gaussian distribution

$\mathbb{E}[\vec{Y}] = \vec{\mu}$, where $\vec{\mu} = \overbrace{(\mu, \mu, \dots, \mu)}^{T}$ and $\mu = \frac{c}{1-\phi}$.

The variance-covariance matrix of $\vec{Y}$ is $\Omega = \mathbb{E}\left[(\vec{Y}-\vec{\mu})(\vec{Y}-\vec{\mu})^T\right]$, which is (as we know already) given by $\Omega = \sigma^2 V$, where

$$V = \frac{1}{1-\phi^2}\begin{pmatrix} 1 & \phi & \phi^2 & \dots & \phi^{T-1} \\ \phi & 1 & \phi & \dots & \phi^{T-2} \\ \phi^2 & \phi & 1 & \dots & \phi^{T-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi^{T-1} & \phi^{T-2} & \phi^{T-3} & \dots & 1 \end{pmatrix}, \text{ since } \mathbb{E}\left[(Y_t-\mu)(Y_{t-j}-\mu)\right] = \frac{\sigma^2\phi^j}{1-\phi^2} = \gamma_j$$

We can consider our observed sample $\vec{y}$ as a single draw from a multivariate Normal $N(\vec{\mu}, \Omega)$ distribution, so from the formula for the multivariate Gaussian density (see Lecture Notes 1C), we have the likelihood function

$$f_{\vec{Y}}(\vec{y}; \vec{\theta}) = (2\pi)^{-T/2}\left(\det(\Omega^{-1})\right)^{1/2} \cdot e^{-\frac{1}{2}(\vec{y}-\vec{\mu})^T\Omega^{-1}(\vec{y}-\vec{\mu})}$$

The log likelihood is then

$$(**)\quad \boxed{\mathcal{L}(\vec{\theta}) = -\frac{T}{2}\log(2\pi) + \frac{1}{2}\log\left(\det(\Omega^{-1})\right) - \frac{1}{2}(\vec{y}-\vec{\mu})^T\Omega^{-1}(\vec{y}-\vec{\mu})}$$

This is the same formula as $(*)$. Why? Well, $V^{-1} = L^T L$, where

$$L = \begin{pmatrix} \sqrt{1-\phi^2} & 0 & 0 & \dots & 0 & 0 \\ -\phi & 1 & 0 & \dots & 0 & 0 \\ 0 & -\phi & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \dots & -\phi & 1 \end{pmatrix}, \text{ so } \Omega^{-1} = \sigma^{-2}L^T L. \text{ Hence,}$$

$$\mathcal{L}(\vec{\theta}) = -\frac{T}{2}\log(2\pi) + \frac{1}{2}\log\left(\det(\sigma^{-2}L^T L)\right) - \frac{1}{2}(\vec{y}-\vec{\mu})^T\sigma^{-2}L^T L(\vec{y}-\vec{\mu})$$

Finally let

$$\tilde{y} = L(\vec{y}-\vec{\mu}) = \begin{pmatrix} \sqrt{1-\phi^2}(y_1-\mu) \\ (y_2-\mu)-\phi(y_1-\mu) \\ (y_3-\mu)-\phi(y_2-\mu) \\ \vdots \\ (y_T-\mu)-\phi(y_{T-1}-\mu) \end{pmatrix} \underset{\mu=\frac{c}{1-\phi}}{=} \begin{pmatrix} \sqrt{1-\phi^2}(y_1-c/(1-\phi)) \\ y_2-c-\phi y_1 \\ y_3-c-\phi y_2 \\ \vdots \\ y_T-c-\phi y_{T-1} \end{pmatrix}$$

So,

$$\frac{1}{2}(\vec{y}-\vec{\mu})^T\sigma^{-2}L^T L(\vec{y}-\vec{\mu}) = \frac{1}{2\sigma^2}\tilde{y}^T\tilde{y} = \frac{1}{2\sigma^2}(1-\phi^2)\left(y_1-\frac{c}{1-\phi}\right)^2 + \frac{1}{2\sigma^2}\sum^{T}(y_t-c-\phi y_{t-1})^2$$

while the middle term

$$\frac{1}{2}\log\left(\det(\sigma^{-2}L^TL)\right)=\frac{1}{2}\log\left(\sigma^{-2T}\det(L^TL)\right)=-\frac{1}{2}\log\sigma^{2T}+\frac{1}{2}\log\left(\det(L^TL)\right)=$$

$$=-\frac{T}{2}\log\sigma^2+\log\det(L)=-\frac{T}{2}\log\sigma^2+\frac{1}{2}\log(1-\phi^2).$$

det(L)=det(L^T) L is lower triangular

Now, it's evident that (*) and (**) are the same formula.

(*) is preferred for computation purposes, since it does not involve matrix inversion

(*) is known as the prediction-error decomposition of the log-likelihood function.

Once we have found $\mathcal{L}(\vec{\theta})$, we would differentiate it w.r.t. to $\vec{\theta}$ and set derivatives equal to 0. This usually results in a system of nonlinear equations in $\vec{\theta}$ and $(y_1, y_2, \dots, y_T)$ for which there is no simple solution for $\vec{\theta}$ in terms of $(y_1, \dots, y_T)=\vec{y}$.

So, numerical procedures are required.

3.2.2 Conditional maximum likelihood function

Instead of doing the numerical maximization, once $\mathcal{L}(\vec{\theta})$ is found, it makes sense to regard the value of y_1 as deterministic and maximize the likelihood conditional on this first observation, i.e. maximize

$$\log f_{Y_T,Y_{T-1},\dots,Y_2|Y_1}(y_T,y_{T-1},\dots,y_2|y_1;\vec{\theta})=\log\prod_{t=2}^{T}f_{Y_t|Y_{t-1}}(y_t|y_{t-1};\vec{\theta})=$$

$$=-\frac{T-1}{2}\log(2\pi)-\frac{T-1}{2}\log\sigma^2-\sum_{t=2}^{T}\frac{(y_t-c-\phi y_{t-1})^2}{2\sigma^2}$$

Now, the maximization w.r.t. c and ϕ is equivalent to minimization of $\sum_{t=2}^{T}(y_t-c-\phi y_{t-1})^2$. This is just ordinary least squares (OLS) regression of y_t on a constant and its own lagged value. We'll see later that this gives conditional mle's for c, ϕ:

$$\begin{pmatrix}\hat{c}\\ \hat{\phi}\end{pmatrix}=\begin{pmatrix}T-1 & \sum_{t=2}^{T}y_{t-1}\\ \sum_{t=2}^{T}y_{t-1} & \sum_{t=2}^{T}y_{t-1}^2\end{pmatrix}^{-1}\begin{pmatrix}\sum_{t=2}^{T}y_t\\ \sum_{t=2}^{T}y_t y_{t-1}\end{pmatrix}$$

What about conditional m.l.e. for σ^2? Well, differentiate w.r.t. σ^2 to get

$$-\frac{T-1}{2\sigma^2} + \sum_{t=2}^{T} \frac{(y_t - c - \phi y_{t-1})^2}{2\sigma^4} = 0 \Rightarrow \hat{\sigma}^2 = \sum_{t=2}^{T} \frac{(y_t - \hat{c} - \hat{\phi} y_{t-1})^2}{T-1}$$

which is just the average squared residual from the regression.

So, in contrast to real m.l.e.'s for c, ϕ, σ^2; the conditional m.l.e's are trivial to compute. Moreover, if the sample size T is sufficiently large, the 1st observation makes a negligible contribution to the total likelihood.

So, in most applications, conditional m.l.e's are computed instead.

(It also turns out that for $|\phi|<1$, the exact m.l.e. and the cond m.l.e. have the same large-sample distribution).

3.2.3 AR(p) $\quad Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \dots + \phi_p Y_{t-p} + \varepsilon_t \qquad \varepsilon_t \sim \text{i.i.d. } N(0, \sigma^2)$

$$\vec{\theta} = (c, \phi_1, \phi_2, \dots, \phi_p, \sigma^2)^T$$

We use a combination of the two methods we used for AR(1). First we collect the first p observations in the sample into a $p\times 1$ vector $\vec{y}_p = (y_1, y_2, \dots, y_p)$ which is viewed as a single realization of a p-dim. multivariate normal variable.

Let $\vec{\mu}_p = \mathbb{E}[\vec{y}_p] = \begin{pmatrix} \mu \\ \vdots \\ \mu \end{pmatrix}_{p \times 1}$, where $\mu = \dfrac{c}{1-\phi_1 - \dots - \phi_p}$.

Let $\sigma^2 V_p = \begin{pmatrix} \gamma_0 & \gamma_1 & \dots & \gamma_{p-1} \\ \gamma_1 & \gamma_0 & \dots & \gamma_{p-2} \\ \vdots & & \ddots & \vdots \\ \gamma_{p-1} & \gamma_{p-2} & \dots & \gamma_0 \end{pmatrix}$ be the variance-covariance matrix of $(Y_1, Y_2, \dots, Y_p)$

where $\gamma_0, \gamma_1, \dots, \gamma_{p-1}$ can be found from Yule-Walker equations (see 1.3.4)

The density of the first p observations is then that of a $N(\vec{\mu}_p, \sigma^2 V_p)$ multivariate normal variable

$$f_{Y_p, Y_{p-1}, \dots, Y_1}(y_p, y_{p-1}, \dots, y_1; \vec{\theta}) = \frac{1}{(2\pi)^{p/2}} \cdot \frac{1}{(\sigma^2)^{p/2}} \cdot \left(\det(V_p^{-1})\right)^{1/2} \cdot e^{-\frac{1}{2\sigma^2}(\vec{y_p}-\vec{\mu_p})^T V_p^{-1}(\vec{y_p}-\vec{\mu_p})}$$

Now, let's consider the remaining observations in our sample $(y_{p+1}, y_{p+2}, \dots, y_T)$.
Conditional on the first $t-1$ observations, the t^{th} observation is normal with mean $c+\phi_1 y_{t-1}+\phi_2 y_{t-2}+\dots+\phi_p y_{t-p}$ and variance σ^2.
Hence, for $t>p$, we have:

$$f_{Y_t|Y_{t-1}, Y_{t-2}, \dots, Y_1}(y_t|y_{t-1}, y_{t-2}, \dots, y_1; \vec{\theta}) = f_{Y_t|Y_{t-1}, \dots, Y_{t-p}}(y_t|y_{t-1}, \dots, y_{t-p}; \vec{\theta})$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_t-c-\phi_1 y_{t-1}-\dots-\phi_p y_{t-p})^2}{2\sigma^2}}$$

Now, $\mathcal{L}(\vec{\theta}) = \log f_{Y_T, Y_{T-1}, \dots, Y_1}(y_T, y_{T-1}, \dots, y_1; \vec{\theta}) =$

$$= \log f_{Y_p, Y_{p-1}, \dots, Y_1}(y_p, \dots, y_1; \vec{\theta}) + \sum_{t=p+1}^{T} \log f_{Y_t|Y_{t-1}, \dots, Y_{t-p}}(y_t|y_{t-1}, \dots, y_{t-p}; \vec{\theta})$$

$$= -\frac{p}{2}\log(2\pi) - \frac{p}{2}\log(\sigma^2) + \frac{1}{2}\log\left(\det(V_p^{-1})\right) - \frac{1}{2\sigma^2}(\vec{y_p}-\vec{\mu_p})^T V_p^{-1}(\vec{y_p}-\vec{\mu_p})$$

$$- \frac{T-p}{2}\log(2\pi) - \frac{T-p}{2}\log(\sigma^2) - \sum_{t=p+1}^{T}\frac{(y_t-c-\phi_1 y_{t-1}-\dots-\phi_p y_{t-p})^2}{2\sigma^2}, \text{ i.e.}$$

$$\boxed{\mathcal{L}(\vec{\theta}) = -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\sigma^2) + \frac{1}{2}\log\left(\det(V_p^{-1})\right) - \frac{1}{2\sigma^2}(\vec{y_p}-\vec{\mu_p})^T V_p^{-1}(\vec{y_p}-\vec{\mu_p}) - \sum_{t=p+1}^{T}\frac{(y_t-c-\phi_1 y_{t-1}-\dots-\phi_p y_{t-p})^2}{2\sigma^2}}.$$

Using this formula requires inverting V_p. One can use Galbraith's equations: the (i,j)-entry $v^{ij}(p)$ of V_p^{-1} is:

$$v^{ij}(p) = \sum_{k=0}^{i-1}\phi_k \phi_{k+j-i} - \sum_{k=p+1-j}^{p+i-j}\phi_k \phi_{k+j-i} \quad \text{for } 1 \le i \le j \le p, \text{ and } v^{ij}(p) = v^{ji}(p).$$

<u>Example</u>: AR(2) process ($p=2$). Galbraith's equations give:

$$V_2^{-1} = \begin{pmatrix} 1-\phi_2^2 & -(\phi_1+\phi_1\phi_2) \\ -(\phi_1+\phi_1\phi_2) & 1-\phi_2^2 \end{pmatrix}, \text{ so } \det(V_2^{-1}) = (1+\phi_2)^2\left[(1-\phi_2)^2-\phi_1^2\right]$$

and $$(\vec{y}_2-\vec{\mu}_2)^T V_2^{-1}(\vec{y}_2-\vec{\mu}_2) = (1+\phi_2)\left((1-\phi_2)(y_1-\mu)^2 - 2\phi_1(y_1-\mu)(y_2-\mu) + (1-\phi_2)(y_2-\mu)^2\right)$$

So, for AR(2) process, the exact log likelihood is

$$\mathcal{L}(\vec{\theta}) = -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\sigma^2) + \frac{1}{2}\log\left((1+\phi_2)^2\left[(1-\phi_2)^2-\phi_1^2\right]\right) - \left(\frac{1+\phi_2}{2\sigma^2}\right)\Big((1-\phi_2)(y_1-\mu)^2 -$$
$$-2\phi_1(y_1-\mu)(y_2-\mu) + (1-\phi_2)(y_2-\mu)^2\Big) - \sum_{t=3}^{T}\frac{(y_t-c-\phi_1 y_{t-1}-\phi_2 y_{t-2})^2}{2\sigma^2}$$

where $$\mu = \frac{c}{1-\phi_1-\phi_2}.$$

3.2.4. <u>Conditional MLE's for AR(p)</u>

The log of the likelihood function conditional on the first p observations is

$$\log f_{Y_T, Y_{T-1},\dots,Y_{p+1}|Y_p,\dots,Y_1}(y_T, y_{T-1},\dots,y_{p+1}|y_p,\dots,y_1;\vec{\theta}) =$$
$$= -\frac{T-p}{2}\log(2\pi) - \frac{T-p}{2}\log(\sigma^2) - \sum_{t=p+1}^{T}\frac{(y_t-c-\phi_1 y_{t-1}-\dots-\phi_p y_{t-p})^2}{2\sigma^2}$$

So, in order to maximize this, we need to find $c, \phi_1, \dots, \phi_p$ that minimize

$$\sum_{t=p+1}^{T}\frac{(y_t-c-\phi_1 y_{t-1}-\dots-\phi_p y_{t-p})^2}{2\sigma^2}$$

This is just OLS regression of y_t on a constant and p of its own lagged values.

Again, as before,

$$\hat{\sigma}^2 = \frac{1}{T-p}\sum_{t=p+1}^{T}(y_t-\hat{c}-\hat{\phi}_1 y_{t-1}-\dots-\hat{\phi}_p y_{t-p})^2$$

Again, the conditional and exact m.l.e's are pretty much the same if T is large.

3.2.5. What if the time series is not Gaussian?

We assumed so far in this Chapter that $\varepsilon_t \sim$ i.i.d. $N(0,\sigma^2)$. But what if this is not true;
For a positive random variable Y_t, Box and Cox proposed the general set of transformations:

$$Y_t^{(\lambda)} = \begin{cases} \dfrac{Y_t^{\lambda}-1}{\lambda}, & \text{for } \lambda \neq 0 \\ \log Y_t, & \text{for } \lambda = 0. \end{cases}$$

These transformations often produce a Gaussian time series.
So, the approach would be to pick a particular value of λ and maximize the likelihood function for $Y_t^{(\lambda)}$ under the assumption that $Y_t^{(\lambda)}$ is a Gaussian ARMA process.
The value of λ that is associated with the highest value of the maximized likelihood is taken as the best transformation.

3.3. Fitting the MA processes using MLE approach

3.3.1. Conditional MLE

Calculation of the likelihood function for AR(p) processes turned out to be much simpler if we condition on initial values for the Y's.
Similarly, calculation of the likelihood function for an MA process is simpler if we condition on initial values for the ε's.
Let's look at MA(1): $Y_t = \mu + \varepsilon_t + \theta\varepsilon_{t-1}$, with $\varepsilon_t \sim$ i.i.d. $N(0,\sigma^2)$.
$\vec{\theta} = (\mu, \theta, \sigma^2)^T \leftarrow$ population parameters to be estimated.
If ε_{t-1} were known with certainty, then

$$Y_t \mid \varepsilon_{t-1} \sim N\left((\mu+\theta\varepsilon_{t-1}), \sigma^2\right) \text{ or } f_{Y_t|\varepsilon_{t-1}}(y_t \mid \varepsilon_{t-1}; \vec{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_t-\mu-\theta\varepsilon_{t-1})^2}{2\sigma^2}} \quad \oplus$$

So, suppose that we knew for certain that $\varepsilon_0 = 0$. Then

$$(Y_1 \mid \varepsilon_0 = 0) \sim N(\mu, \sigma^2)$$

Moreover, given observation of y_1, the value of ε_1 is then known with certainty as well;

$$\varepsilon_1 = y_1 - \mu, \text{ allowing application of } \oplus \text{ again!}$$

$$f_{Y_2|Y_1,\varepsilon_0=0}(y_2 \mid y_1, \varepsilon_0 = 0; \vec{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_2-\mu-\theta\varepsilon_1)^2}{2\sigma^2}}.$$

Since ε_1 is known with certainty, ε_2 can be calculated from $\varepsilon_2 = y_2 - \mu - \theta\varepsilon_1$. We proceed in this fashion, so it's clear that given knowledge $\varepsilon_0 = 0$, the sequence $\{\varepsilon_1, \varepsilon_2, \dots, \varepsilon_T\}$ can be calculated from $\{y_1, y_2, \dots, y_T\}$ by iterating on

$$\varepsilon_t = y_t - \mu - \theta\varepsilon_{t-1} \text{ for } t = 1, 2, \dots, T \text{ starting from } \varepsilon_0 = 0.$$

The conditional density of the t^{th} observation is:

$$f_{Y_t|Y_{t-1}, Y_{t-2}, \dots, Y_1, \varepsilon_0 = 0}(y_t | y_{t-1}, y_{t-2}, \dots, y_1, \varepsilon_0 = 0; \vec{\theta}) = f_{Y_t|\varepsilon_{t-1}}(y_t|\varepsilon_{t-1}; \vec{\theta}) =$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\varepsilon_t^2/2\sigma^2}$$

The sample likelihood is then:

$$f_{Y_T, Y_{T-1}, \dots, Y_1|\varepsilon_0 = 0}(y_T, y_{T-1}, \dots, y_1 | \varepsilon_0 = 0; \vec{\theta}) =$$

$$= f_{Y_1|\varepsilon_0}(y_1|\varepsilon_0 = 0; \vec{\theta}) \cdot \prod_{t=2}^{T} f_{Y_t|\varepsilon_{t-1}}(y_t|\varepsilon_{t-1}; \vec{\theta})$$, so the conditional log likelihood

$$\mathcal{L}(\vec{\theta}) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2) - \frac{\varepsilon_1^2}{2\sigma^2} - \frac{T-1}{2}\log(2\pi) - \frac{T-1}{2}\log(\sigma^2) - \sum_{t=2}^{T}\frac{\varepsilon_t^2}{2\sigma^2}$$

i.e. $$\boxed{\mathcal{L}(\vec{\theta}) = -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\sigma^2) - \sum_{t=1}^{T}\frac{\varepsilon_t^2}{2\sigma^2}}$$

So, how would the procedure work? For a particular numerical value of $\vec{\theta} = (\mu, \theta, \sigma^2)$ we calculate the sequence of ε's implied by the data from $\varepsilon_t = Y_t - \mu - \theta\varepsilon_{t-1}$. The conditional log likelihood is then a function of the sum of squares of these ε's. Although it's simple to program this iteration, the log likelihood is fairly complicated nonlinear function of μ and θ. So, unlike for fitting AR models, even the conditional MLE for an MA(1) process must be found by numerical optimization.

Q: How good is the assumption that $\varepsilon_0 = 0$?

If MA(1) is invertible, i.e. $|\theta| < 1$, this assumption will result in a very good approximation to the exact MLE's for a reasonably large sample size.

3.3.2. Exact MLE's for MA(1)

Let $\vec{y} = (y_1, y_2, \ldots, y_T)^T \leftarrow$ observed data into $T \times 1$ vector

$\vec{\mu} = \mathbb{E}[\vec{y}] = \begin{pmatrix} \mu \\ \vdots \\ \mu \end{pmatrix}$ and $\Omega = \mathbb{E}[(\vec{Y}-\vec{\mu})(\vec{Y}-\vec{\mu})^T]$. We know from 1.2.1 that the autocovariance matrix Ω is:

$$\Omega = \sigma^2 \begin{pmatrix} 1+\theta^2 & \theta & 0 & \cdots & 0 \\ \theta & 1+\theta^2 & \theta & \cdots & 0 \\ 0 & \theta & 1+\theta^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & & 1+\theta^2 \end{pmatrix}$$

The likelihood function is: $f_{\vec{Y}}(\vec{y};\vec{\theta}) = (2\pi)^{-T/2} (\det(\Omega^{-1}))^{1/2} e^{-\frac{1}{2}(\vec{y}-\vec{\mu})^T \Omega^{-1} (\vec{y}-\vec{\mu})}$

Let's use the triangular decomposition of Ω: $\Omega = ADA^T$, where

$$A = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ \frac{\theta}{1+\theta^2} & 1 & 0 & \cdots & 0 & 0 \\ 0 & \frac{\theta(1+\theta^2)}{1+\theta^2+\theta^4} & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \frac{\theta(1+\theta^2+\ldots+\theta^{2(T-2)})}{1+\theta^2+\theta^4+\ldots+\theta^{2(T-1)}} & 1 \end{pmatrix}$$ and

$$D = \sigma^2 \begin{pmatrix} 1+\theta^2 & 0 & 0 & \cdots & 0 \\ 0 & \frac{1+\theta^2+\theta^4}{1+\theta^2} & 0 & \cdots & 0 \\ 0 & 0 & \frac{1+\theta^2+\theta^4+\theta^6}{1+\theta^2+\theta^4} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \frac{1+\theta^2+\ldots+\theta^{2T}}{1+\theta^2+\ldots+\theta^{2(T-1)}} \end{pmatrix}$$

Now, A lower triangular $\Rightarrow \det(A) = 1$ (ones on the main diagonal)

$\Rightarrow \det(\Omega) = \det(A)\det(D)\det(A^T) = \det(D)$

If we define $\tilde{y} = A^{-1}(\vec{y}-\vec{\mu})$, then the likelihood becomes:

$$f_{\vec{Y}}(\vec{y};\vec{\theta}) = (2\pi)^{-T/2} \cdot (\det D)^{-1/2} \cdot e^{-\frac{1}{2}\tilde{y}^T D^{-1} \tilde{y}}$$

Since $A\tilde{y} = \vec{y} - \vec{\mu}$, we have $\tilde{y}_1 = y_1 - \mu$ (from the 1^{st} row) and

$$\tilde{y}_t = y_t - \mu - \frac{\theta(1+\theta^2+\theta^4+\dots+\theta^{2(t-2)})}{1+\theta^2+\theta^4+\dots+\theta^{2(t-1)}}\,\tilde{y}_{t-1}, \text{ for } t=2,3,\dots,T$$

from the t^{th} row.

Now, $\det D = \prod_{t=1}^{T} d_{tt}$ (product of entries on the main diagonal), where

$$d_{tt} = \sigma^2 \cdot \frac{1+\theta^2+\dots+\theta^{2t}}{1+\theta^2+\dots+\theta^{2(t-1)}}.$$ Hence, $\tilde{y}^T D^{-1}\tilde{y} = \sum_{t=1}^{T} \frac{\tilde{y}_t^2}{d_{tt}}$.

So, $$f_{\vec{Y}}(\vec{y};\vec{\theta}) = (2\pi)^{-T/2}\cdot\left(\prod_{t=1}^{T} d_{tt}\right)^{-1/2}\cdot e^{-\frac{1}{2}\sum_{t=1}^{T}\tilde{y}_t^2/d_{tt}}$$

$\Rightarrow$ exact log likelihood for an MA(1) process is:

$$\mathcal{L}(\vec{\theta}) = \log f_{\vec{Y}}(\vec{y};\vec{\theta}) = -\frac{T}{2}\log(2\pi) - \frac{1}{2}\sum_{t=1}^{T}\log(d_{tt}) - \frac{1}{2}\sum_{t=1}^{T}\frac{\tilde{y}_t^2}{d_{tt}}.$$

So, given a numerical value for μ, θ, σ^2, first calculate the sequence $\tilde{y}_t$ by iterating this, starting at $\tilde{y}_1 = y_1 - \mu$, while d_{tt} is given by this.

3.3.2. <u>MA(q) – conditional MLE's</u>

$$Y_t = \mu + \varepsilon_t + \theta_1\varepsilon_{t-1} + \dots + \theta_q\varepsilon_{t-q} \qquad \vec{\theta} = (\mu, \theta_1, \theta_2, \dots, \theta_q, \sigma^2)^T.$$

simple approach: condition on $\varepsilon_0 = \varepsilon_{-1} = \dots = \varepsilon_{-q+1} = 0$.

then iterate on $\varepsilon_t = y_t - \mu - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} - \dots - \theta_q\varepsilon_{t-q}$ for $t = 1, 2, \dots, T$

The conditional log likelihood is then

$$\mathcal{L}(\vec{\theta}) = \log f_{Y_T,\dots,Y_1|\varepsilon_0=\varepsilon_{-1}=\dots=\varepsilon_{-q+1}=0}(y_T,\dots,y_1|\varepsilon_0=\varepsilon_{-1}=\dots=\varepsilon_{-q+1}=0;\vec{\theta})$$

$$\mathcal{L}(\vec{\theta}) = -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\sigma^2) - \sum_{t=1}^{T}\frac{\varepsilon_t^2}{2\sigma^2}.$$

(this works only if MA(q) is invertible, i.e. if all roots of $1+\theta_1 z+\theta_2 z^2+\dots+\theta_q z^q = 0$ are outside the unit circle)

5.4. Fitting an ARMA(p,q) process

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \dots + \phi_p Y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \dots + \theta_q \varepsilon_{t-q}$$

$\varepsilon_t \sim$ i.i.d. $N(0,\sigma^2)$, $\vec{\theta} = (c, \phi_1, \phi_2, \dots, \phi_p, \theta_1, \theta_2, \dots \theta_q, \sigma^2)^T$

A common approximation to the likelihood function for an ARMA(p,q) process conditions on both y's and ε's. Taking initial values for $\vec{y_0} \equiv (y_0, y_{-1}, \dots, y_{-p+1})^T$ and $\vec{\varepsilon_0} = (\varepsilon_0, \varepsilon_{-1}, \dots, \varepsilon_{-q+1})^T$ as given, the sequence $\{\varepsilon_1, \varepsilon_2, \dots, \varepsilon_T\}$ can be calculated from $\{y_1, y_2, \dots, y_T\}$ by iterating on:

$$\varepsilon_t = y_t - c - \phi_1 y_{t-1} - \phi_2 y_{t-2} - \dots - \phi_p y_{t-p} - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \dots - \theta_q \varepsilon_{t-q}, \quad t = 1, 2, \dots, T$$

The conditional log likelihood is then

$$\mathcal{L}(\vec{\theta}) = \log f_{Y_T, Y_{T-1}, \dots, Y_1 | \vec{Y_0}, \vec{\varepsilon_0}}(y_T, y_{T-1}, \dots, y_1 | \vec{y_0}, \vec{\varepsilon_0}; \vec{\theta}) =$$

$$= -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\sigma^2) - \sum_{t=1}^{T} \frac{\varepsilon_t^2}{2\sigma^2}$$

1st Approach: Set initial y's and ε's equal to their expected values. In other words,

set $y_s = \dfrac{c}{1-\phi_1-\phi_2-\dots-\phi_p}$ for $s = 0, -1, \dots, -p+1$ and set

$\varepsilon_s = 0$ for $s = 0, -1, \dots, -q+1$, and then proceed with iteration for $t = 1, \dots, T$

2nd Approach: (Box-Jenkins)

Start the iteration at date $t = p+1$ with $y_1, \dots, y_p$ set to their observed values and $\varepsilon_p = \varepsilon_{p-1} = \dots = \varepsilon_{p-q+1} = 0$.

The conditional log likelihood is then

$$-\frac{T-p}{2}\log(2\pi) - \frac{T-p}{2}\log(\sigma^2) - \sum_{t=p+1}^{T} \frac{\varepsilon_t^2}{2\sigma^2}$$

3.5. Numerical optimization

Once we have found the log likelihood function $\mathcal{L}(\vec{\theta})$, we need to find the value of $\vec{\theta}$ that maximizes it. This is usually done by computers (numerical optimization).

We assume we have a black box that enables a computer to calculate the numerical value of $\mathcal{L}(\vec{\theta})$ given any particular values for $\vec{\theta}$ and the observed data $y_T, \dots, y_1$.

$$\underset{\text{Input}}{\boxed{\text{values of } y_T, \dots, y_1 \text{ and } \vec{\theta}}} \longrightarrow \underset{\text{procedure}}{\boxed{\text{Calculates } \mathcal{L}(\vec{\theta})}} \longrightarrow \underset{\text{output}}{\boxed{\text{value of } \mathcal{L}(\vec{\theta})}}$$

The idea of numerical optimization is to make a series of different guesses for $\vec{\theta}$, compare the value of $\mathcal{L}(\vec{\theta})$ for each guess, and try to infer from these values the value $\hat{\theta}$ for which $\mathcal{L}(\vec{\theta})$ is largest. There are many methods that computers use; here we describe only one!

Newton-Raphson method

Assumptions: ① Second derivatives of the log likelihood $\mathcal{L}(\vec{\theta})$ exist

② $\mathcal{L}(\vec{\theta})$ is concave, i.e. -1 times the matrix of second derivatives (known as Hessian) is everywhere positive definite.

Suppose $\vec{\theta}$ is an $a \times 1$ vector of parameters to be estimated. Let $\vec{g}(\vec{\theta}^{(0)})$ be the gradient vector of the log likelihood function at $\vec{\theta}^{(0)}$, i.e.

$$\underset{(a\times 1)}{\vec{g}(\vec{\theta}^{(0)})} = \left.\frac{\partial \mathcal{L}(\vec{\theta})}{\partial \vec{\theta}}\right|_{\vec{\theta}=\vec{\theta}^{(0)}}.$$

Let $H(\vec{\theta}^{(0)})$ be -1 times the matrix of 2^{nd} derivatives, i.e. $\underset{(a\times a)}{H(\vec{\theta}^{(0)})} = -\left.\frac{\partial^2 \mathcal{L}(\vec{\theta})}{\partial \vec{\theta}\, \partial \vec{\theta}^T}\right|_{\vec{\theta}=\vec{\theta}^{(0)}}$

Taylor series approximation of $\mathcal{L}(\vec{\theta})$ around $\vec{\theta}^{(0)}$:

$$(\circledast) \quad \mathcal{L}(\vec{\theta}) \approx \mathcal{L}(\vec{\theta}^{(0)}) + [\vec{g}(\vec{\theta}^{(0)})]^T [\vec{\theta} - \vec{\theta}^{(0)}] - \frac{1}{2} [\vec{\theta} - \vec{\theta}^{(0)}]^T H(\vec{\theta}^{(0)}) [\vec{\theta} - \vec{\theta}^{(0)}]$$

Idea is to choose $\vec{\theta}$ so as to maximize $\circledast$. Take a derivative of $\circledast$ w.r.t. $\vec{\theta}$, set it to $\vec{0} \Rightarrow$

$$(\oplus) \quad \vec{g}(\vec{\theta}^{(0)}) - H(\vec{\theta}^{(0)}) [\vec{\theta} - \vec{\theta}^{(0)}] = \vec{0}.$$

Let $\vec{\theta}^{(0)}$ denote the initial guess for the value of $\vec{\theta}$.

How do we calculate the derivatives of $\mathcal{L}(\vec{\theta})$, i.e. $\vec{g}(\vec{\theta}^{(0)})$ and $H(\vec{\theta}^{(0)})$??

Well, for example, the i^{th} element of $\vec{g}(\vec{\theta}^{(0)})$ might be approximated by:

$$g_i(\vec{\theta}^{(0)}) \cong \frac{1}{\Delta}\left[\mathcal{L}(\theta_1^{(0)}, \ldots, \theta_{i-1}^{(0)}, \theta_i^{(0)}+\Delta, \theta_{i+1}^{(0)}, \ldots, \theta_a^{(0)}) - \mathcal{L}(\vec{\theta}^{(0)})\right]$$

where Δ is some very small scalar, e.g. $\Delta = 10^{-6}$

Now, ⊕ suggests that an improved estimate of $\vec{\theta}$, denoted by $\vec{\theta}^{(1)}$ would satisfy

$$\vec{g}(\vec{\theta}^{(0)}) = H(\vec{\theta}^{(0)})\left[\vec{\theta}^{(1)} - \vec{\theta}^{(0)}\right], \text{ i.e. } \vec{\theta}^{(1)} - \vec{\theta}^{(0)} = \left[H(\vec{\theta}^{(0)})\right]^{-1} \cdot \vec{g}(\vec{\theta}^{(0)})$$

In other words, $H(\vec{\theta}^{(0)})$ specifies the "search" direction for maximum. What is usually done here is the combination w/ grid-search method. Instead of the formula above, we use $\vec{\theta}^{(1)} = \vec{\theta}^{(0)} + s\left[H(\vec{\theta}^{(0)})\right]^{-1} \cdot \vec{g}(\vec{\theta}^{(0)})$, where s is a scalar controlling the step length. So, we calculate the value of $\mathcal{L}(\vec{\theta}^{(1)})$ for $s = \frac{1}{16}, \frac{1}{8}, \frac{1}{4}, \frac{1}{2}, 1, 2, 4, 8, 16$ and choose a new estimate $\vec{\theta}^{(1)}$ to be the value of $\vec{\theta}^{(0)} + s\left[H(\vec{\theta}^{(0)})\right]^{-1} \cdot \vec{g}(\vec{\theta}^{(0)})$ for which $\mathcal{L}(\vec{\theta})$ is the largest. Next, one could calculate $\vec{g}(\vec{\theta}^{(1)})$ and $H(\vec{\theta}^{(1)})$ and use these to find a new estimate $\vec{\theta}^{(2)}$, and continue iterating. The m^{th} step of iteration would be

$$\vec{\theta}^{(m+1)} = \vec{\theta}^{(m)} + s \cdot \left[H(\vec{\theta}^{(m)})\right]^{-1} \vec{g}(\vec{\theta}^{(m)}).$$

Drawback: $H(\vec{\theta})$ has $\frac{a(a+1)}{2}$ significant entries (it's symmetric).

So, calculating the inverse could be extremely time consuming, if a is large.

There are other procedures such as Fletcher–Powell, etc.

3.6. Likelihood ratio tests

In the previous section we discussed one method how to find $\vec{\theta}$ that maximizes $\mathcal{L}(\vec{\theta})$ once we have calculated $\mathcal{L}(\vec{\theta})$. Now, we want to discuss one method that can be used to test a hypothesis about $\vec{\theta}$. A popular approach to test hypothesis about parameters that are estimated by MLE's is the likelihood ratio test.

Suppose a null hypothesis implies a set of m different restrictions on the value of the $(a \times 1)$ vector $\vec{\theta}$. First, we maximize the likelihood function ignoring these restrictions to obtain the unrestricted m.l.e. $\hat{\theta}$. Next, we find an estimate $\tilde{\theta}$ that makes the likelihood as large as possible while still satisfying all the restrictions. This is achieved by defining a new $(q-m)\times 1$ vector $\vec{\lambda}$ in terms of which all the elements of $\vec{\theta}$ can be expressed when the restrictions are satisfied. For example, if the restriction is that the last m entries of $\vec{\theta}$ are zero, then $\vec{\lambda}$ consists of the first $q-m$ entries of $\vec{\theta}$. Clearly, $\mathcal{L}(\hat{\theta}) > \mathcal{L}(\tilde{\theta})$. What is important, however, is that it often holds:

$$2[\mathcal{L}(\hat{\theta}) - \mathcal{L}(\tilde{\theta})] \approx \chi^2_m \leftarrow \text{chi-squared distr. w/ } m \text{ DOF.}$$

Simple example: Suppose that the log-likelihood is $\mathcal{L}(\vec{\theta}) = -1.5\theta_1^2 - 2\theta_2^2$, $q=2$, $\vec{\theta}=(\theta_1,\theta_2)^T$

Suppose we're interested in testing $H_0: \theta_2 = \theta_1 + 1$. Under H_0, $\vec{\theta}$ can be written as $(\lambda, \lambda+1)$ where $\lambda = \theta_1$.

Let's find the restricted m.l.e $\tilde{\theta}$.

$$\tilde{\mathcal{L}}(\theta_1) = -1.5\theta_1^2 - 2(\theta_1+1)^2 \Rightarrow -3\theta_1 - 4(\theta_1+1) = 0 \Rightarrow \theta_1 = -4/7.$$

restricted m.l.e. is $\tilde{\theta} = (-4/7, 3/7)^T$, and $\mathcal{L}(\tilde{\theta}) = -6/7$.

Unrestricted m.l.e. $\hat{\theta}$ is clearly $\hat{\theta} = (0,0)^T$ at which $\mathcal{L}(\hat{\theta}) = 0$.

So, $2[\mathcal{L}(\hat{\theta}) - \mathcal{L}(\tilde{\theta})] = 12/7 = 1.71$

$m=1$, so the probability that a χ_1^2-variable exceeds 3.84 is 0.05

Since $1.71 < 3.84$ we accept $H_0: \theta_2 = \theta_1 + 1$ at the 5% significance level.

Likelihood ratio tests are often used for overfitting. We add extra parameters to the model and use likelihood ratio tests to check whether they are significant.

3.7. Model selection for ARMA(p,q) processes

The sample ACF and the sample PACF (see 1.5.3.2) were excellent model selection criteria for MA(q) and AR(p) processes, respectively. However, we did not have a good diagnostic for ARMA(p,q) processes.

AIC and SBC are model selection criteria based on the log-likelihood and can be used to select p and q.

AIC (Akaike's information criterion) is defined as $\boxed{-2\mathcal{L}(\hat{\theta})+2(p+q)}$ where $\mathcal{L}(\hat{\theta})$ is the log likelihood evaluated at the MLE $\hat{\theta}$.

SBC (Schwarz's Bayesian criterion) is defined as $\boxed{-2\mathcal{L}(\hat{\theta})+\log(T)(p+q)}$, where T is the length of the time series.

The best model according to either criterion is the model that minimizes that criterion.

Both criteria tend to select models with large values of the likelihood.

The terms $2(p+q)$ in AIC and $\log(T)(p+q)$ in SBC are penalties on having too many parameters (i.e. lack of parsimony). So, both AIC and SBC both try to trade off a good fit to the data measured by $\mathcal{L}$ with the desire to use as few parameters as possible.

SBC penalizes p+q more than AIC does. Hence, AIC tends to choose models with more parameters than SBC.

In practice, the best AIC and the best SBC models are the same model often.