R301a Time Series

James Legrand[1]

Lent Term, 2023-2024

Contents

[1] A changelog and archive can be found at github.com/james-legrand/Metrics-Notes.

1 Time Series Models

1.1 Basic Concepts

A model describes a stochastic process by which observations are generated. Each observation is a random variable, thus a stochastic process may be defined as a collection of random variables which are ordered in time.
To model in a meaningful way, we must assume some notion of stationarity, that properties of the process remain fixed in time.

Definition 1.1.1: Weak stationarity

A process y_t is weakly stationary when the following are satisfied for all t:

1. $\mathbb{E}(y_t) = \mu$
2. $\mathbb{E}[(y_t - \mu)^2] = \gamma(0)$
3. $\mathbb{E}[(y_t - \mu)(y_{t-\tau} - \mu)] = \gamma(\tau), \quad \tau = 1, 2, \ldots$

Thus the mean stays at a constant level (independent of t), as are the variances and autocovariances.

Definition 1.1.2: Strict stationarity

A process y_t is strictly stationary when the joint distribution of $y_{t_1}, y_{t_2}, \ldots, y_{t_k}$ is the same as the joint distribution of $y_{t_1+\tau}, y_{t_2+\tau}, \ldots, y_{t_k+\tau}$ for all τ and all k.

When a process is strictly stationary, the joint distribution of any finite number of observations is invariant to shifts in time. When a process is strictly stationary, and its first two moments exist, it must be weakly stationary. The converse is not true.
When a series is weakly stationary and the observations have multivariate normal distribution, then it must be strictly stationary also. This is because the entire distribution of a normal is described by its first two moments.

Example (White noise). The simplest example of a covariance (weakly) stationary process is a sequence of uncorrelated random variables with constant mean and variance. This is known as white noise.
Strict white noise are IID. Given the first two moments exist, this is also white noise. Gaussian white noise is strict white noise, for the same reason as above.

Definition 1.1.3: Autocorrelation function

The autocovariances may be standardised by dividing by the variance of the process - yielding the autocorrelation function (ACF)

$$\rho(\tau) = \frac{\gamma(\tau)}{\gamma(0)}$$

The information in the ACF is typically displayed in a plot of $\rho(\tau)$ against τ. For real series, $\rho(\tau) = \rho(-\tau)$, so it's not necessary to extend the plot to negative lags.

Note:-

Ergodicity

An ergodic sequence can be thought of as one for which, over an infinite time period, every event occurs with probability 0 or 1. When a process is ergodic with finite second moment, observations sufficiently far apart in time should be almost uncorrelated.

When the process is ergodic, the sample mean, variance and autocovariances give consistent estimators of the mean, variance and autocovariances of the process.

For all models considered here, stationarity implies ergodicity (although this is not generally the case).

1.2 Time series processes

Definition 1.2.1: MA(1) Process

$$y_t = \mu + \varepsilon_t + \theta\varepsilon_{t-1}, \quad \varepsilon_t \sim WN(0, \sigma^2)$$

We now derive some properties.

$$\begin{aligned}
\mathbb{E}(y_t) &= \mu + \mathbb{E}(\varepsilon_t) + \theta\mathbb{E}(\varepsilon_{t-1}) = \mu \\
\gamma(0) &= \mathbb{E}[(y_t - \mu)^2] = \mathbb{E}[(\varepsilon_t + \theta\varepsilon_{t-1})^2] = \mathbb{E}[\varepsilon_t^2] + 2\theta\mathbb{E}[\varepsilon_t\varepsilon_{t-1}] + \theta^2\mathbb{E}[\varepsilon_{t-1}^2] = \sigma^2(1+\theta^2) \\
\gamma(1) &= \mathbb{E}[(y_t - \mu)(y_{t-1} - \mu)] = \mathbb{E}[(\varepsilon_t + \theta\varepsilon_{t-1})(\varepsilon_{t-1} + \theta\varepsilon_{t-2})] = \mathbb{E}[\theta\varepsilon_{t-1}, \varepsilon_{t-1}] = \theta\sigma^2
\end{aligned}$$

Higher autocovariances are zero. The mean , variance and covariances are therefore independent of t; confirming that the process is stationary. Note that it is not necessary to specify the full distribution of the disturbances ε_t, or even to insist that they be serially independent rather than merely uncorrelated.

The autocorrelation function is

$$\rho(\tau) = \frac{\gamma(\tau)}{\gamma(0)} = \begin{cases} 1 & \tau = 0 \\ \frac{\theta}{1+\theta^2} & |\tau| = 1 \\ 0 & |\tau| > 1 \end{cases}$$

No restrictions need to be placed on θ for it to be stationary, although some ambiguity arises in the ACF because $\rho(1)$ can take the same value for $|\theta|$ and $1/|\theta|$.

When θ is positive, successive values ofy_t are positively correlated and so the process is smoother than the random series ε_t. When θ is negative the series is more irregular than the random series.

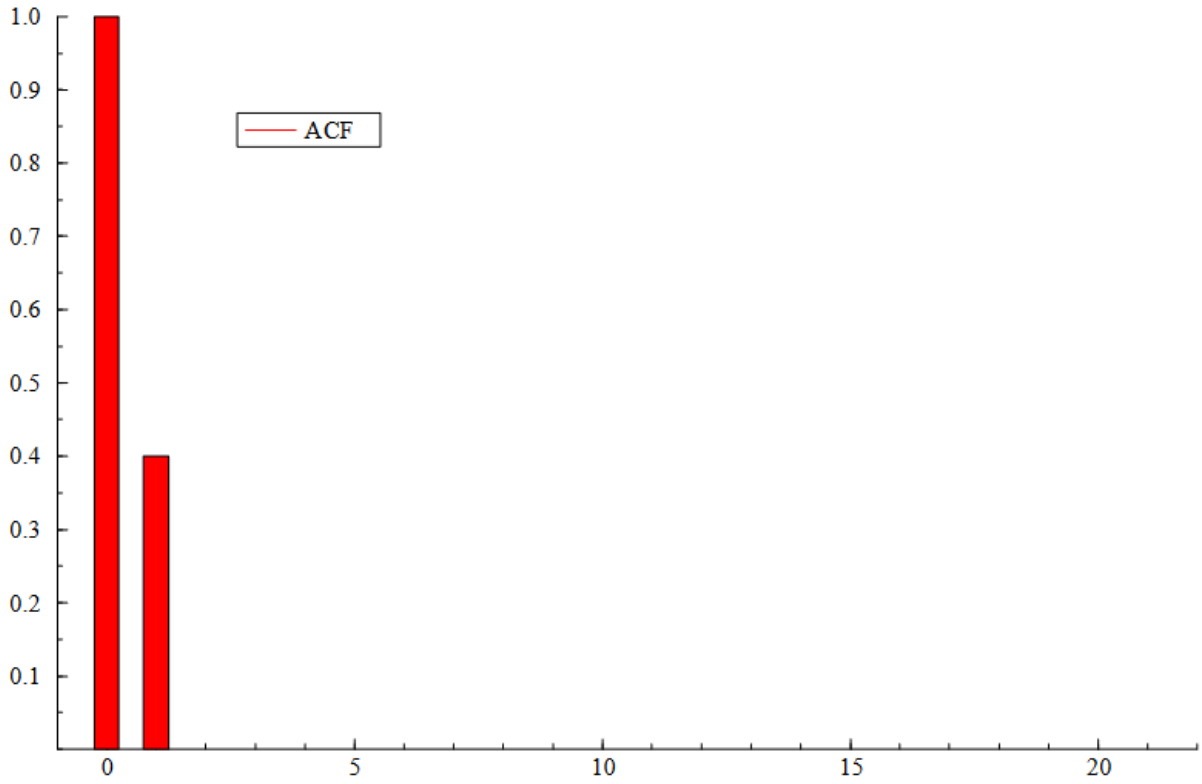


Figure 1: ACF of MA(1) with $\theta = 0.5$

Definition 1.2.2: Lag operator

$$Ly_t = y_{t-1}$$
$$L^\tau = y_{t-\tau}$$

Definition 1.2.3: Forward operator

$$Fy_t = y_{t+1}$$
$$F^\tau = y_{t+\tau}$$

$$L^{-\tau} y_{t+\tau} = F^\tau$$

Definition 1.2.4: First difference operator

$$\Delta = 1 - L$$

Thus $\Delta y_t = y_t - y_{t-1}$

Definition 1.2.5: AR(1) Process

$$y_t = \mu + \phi(y_{t-1} - \mu) + \varepsilon_t = \delta + \phi y_{t-1} + \varepsilon_t$$

By weak stationarity when $\phi < 1$ we can say $\mathbb{E}[y_t] = \mathbb{E}[y_{t-1}]$. Thus

$$\mathbb{E}y_t = \mu + \phi\mathbb{E}y_t - \phi\mu + \mathbb{E}\varepsilon_t \Rightarrow (1-\phi)\mathbb{E}y_t = (1-\phi)\mu \Rightarrow \mathbb{E}y_t = \mu$$

Further we can say that $Var[y_t] = Var[y_{t-1}]$. Thus

$$Var(y_t) = \phi^2 Var(y_t) + Var(\varepsilon_t) \Rightarrow Var(y_t) = \frac{\sigma^2}{1-\phi^2}$$

$$\begin{aligned}
y_t y_{t-\tau} &= \mu y_{t-\tau} + \phi y_{t-1} y_{t-\tau} - \phi\mu y_{t-\tau} + \varepsilon_t y_{t-\tau} \\
\mathbb{E}[y_t y_{t-\tau}] &= \mu\mathbb{E}[y_{t-\tau}] + \phi\mathbb{E}[y_{t-1} y_{t-\tau}] - \phi\mu\mathbb{E}[y_{t-\tau}] + \mathbb{E}[\varepsilon_t y_{t-\tau}] \\
\mathbb{E}[y_t y_{t-\tau}] - \mu^2 &= \phi(\mathbb{E}[y_{t-1} y_{t-\tau}] - \mu^2) \\
\gamma(\tau) &= \phi\gamma(\tau - 1)
\end{aligned}$$

We can solve this first order difference equation to find:

$$\begin{aligned}
\gamma(\tau) &= \phi\gamma(\tau - 1) \\
&= \phi^2\gamma(\tau - 2) \\
&= \dots \\
&= \phi^\tau\gamma(0) \\
&\Rightarrow \rho(\tau) = \phi^\tau
\end{aligned}$$

Since $|\phi| < 1$, the autocorrelation function is a decreasing exponential function of τ.

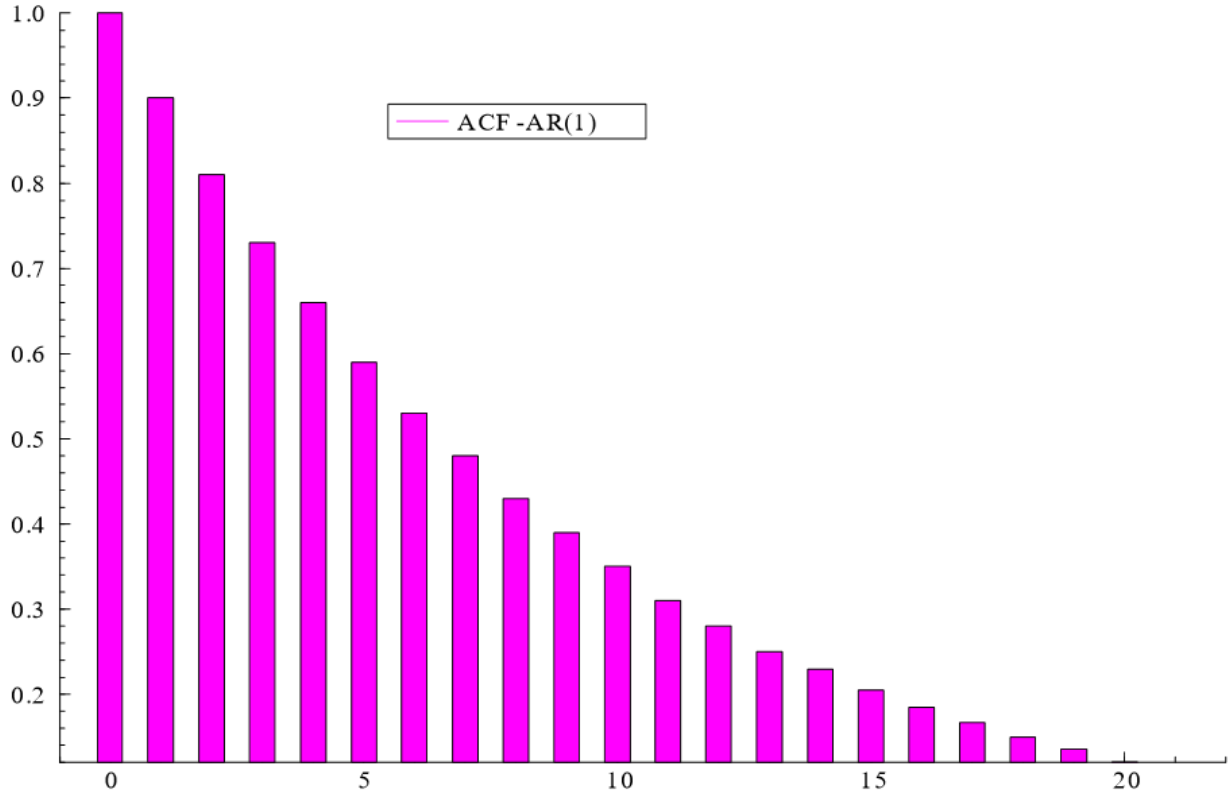


Figure: ACF of AR(1) with $\phi = 0.9$

Definition 1.2.6: Linear process

A time series is a linear process when it can be expressed as

$$y_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}$$

where ε_t is serially uncorrelated, with mean zero and variance σ^2, and $\sum_{j=0}^{\infty} |\psi_j| < \infty$. Such a sequence is said to be absolutely convergent.

A linear process is stationary and its properties may be expressed in terms of the autocovariance function. These properties can be approximated, to any desired level of accuracy, by using an autoregressive-moving average model of order (p,q).

Definition 1.2.7: ARMA Process

$$y_t = \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t$$

Also written as $y_t \sim ARMA(p, q)$.

An ARMA(p,q) can be written more concisely by defining polynomials in the lag operator:

$$\phi(L) = 1 - \phi_1 L - \cdots - \phi_p L^p \theta(L) = 1 + \theta_1 L + \cdots + \theta_q L^q$$

Which allows us to express the ARMA as:

$$\phi(L) y_t = \theta(L) \varepsilon_t$$

Example (ARMA(1,1)). $y_t = \phi y_{t-1} + \varepsilon_t + \theta \varepsilon_{t-1}$
We can write this as $(1 - \phi L) y_t = (1 + \theta L) \varepsilon_t$ and divide both sides by $1 - \phi L$ to get

$$\begin{aligned}
y_t &= \frac{\varepsilon_t}{1 - \phi L} + \frac{\theta \varepsilon_{t-1}}{1 - \phi L} \\
&= \sum_{j=0}^{\infty} (\phi L)^j \varepsilon_t + \theta \sum_{j=0}^{\infty} (\phi L)^j \varepsilon_{t-1} \\
&= \sum_{j=0}^{\infty} (\phi)^j \varepsilon_{t-j} + \theta \sum_{j=0}^{\infty} (\phi)^j \varepsilon_{t-j-1} \\
&= \varepsilon_t + \sum_{j=1}^{\infty} ((\phi)^j + \theta(\phi)^{j-1}) \varepsilon_{t-j} \\
&= \varepsilon_t + (\phi + \theta) \sum_{j=1}^{\infty} (\phi)^{j-1} \varepsilon_{t-j}
\end{aligned}$$

Thus **we have stationarity if** $|\phi| < 1$, the weights decline sufficiently rapidly for the process to have finite variance and for the autocovariances to exist.
To derive the autocovariance function, we multiply both sides by $y_{t-\tau}$ and take expectations:

$$\begin{aligned}
y_t y_{t-\tau} &= \phi y_{t-1} y_{t-\tau} + \varepsilon_t y_{t-\tau} + \theta \varepsilon_{t-1} y_{t-\tau} \\
\mathbb{E}[y_t y_{t-\tau}] &= \phi \mathbb{E}[y_{t-1} y_{t-\tau}] + \mathbb{E}[\varepsilon_t y_{t-\tau}] + \theta \mathbb{E}[\varepsilon_{t-1} y_{t-\tau}] \\
\gamma(\tau) &= \phi \gamma(\tau - 1) + \mathbb{E}[\varepsilon_t y_{t-\tau}] + \theta \mathbb{E}[\varepsilon_{t-1} y_{t-\tau}]
\end{aligned}$$

The last two terms are zero for $\tau > 1$. For $\tau = 1$ the first remains zero, but the second

becomes

$$\mathbb{E}[\varepsilon_{t-1}y_{t-1}] = \mathbb{E}[\varepsilon_{t-1}(\phi y_{t-2} + \varepsilon_{t-1} + \theta\varepsilon_{t-2})] = \phi\mathbb{E}[\varepsilon_{t-1}y_{t-2}] + \sigma^2 = \sigma^2$$

When $\tau = 0$, both expectations are non-zero, and given by $\mathbb{E}[\varepsilon_t y_t] = \sigma^2$ and

$$\mathbb{E}[\varepsilon_{t-1}y_t] = \mathbb{E}[\varepsilon_{t-1}(\phi y_{t-1} + \varepsilon_t + \theta\varepsilon_{t-1})] = \phi\mathbb{E}[\varepsilon_{t-1}y_{t-1}] + \theta\mathbb{E}[\varepsilon_{t-1}^2] = \phi\sigma^2 + \theta\sigma^2$$

Thus we have the autocovariance function:

$$\begin{aligned}\gamma(0) &= \phi\gamma(1) + \sigma^2 + \theta\phi\sigma^2 + \theta^2\sigma^2\\ \gamma(1) &= \phi\gamma(0) + \theta\sigma^2\\ \gamma(\tau) &= \phi\gamma(\tau-1) \quad \tau > 1\end{aligned}$$

Solving the the first two equations gives:

$$\begin{aligned}\gamma(0) &= \frac{1+\theta^2+2\phi\theta}{1-\phi^2}\sigma^2\\ \gamma(1) &= \frac{(1+\phi\theta)(\phi+\theta)}{1-\phi^2}\sigma^2\end{aligned}$$

Which gives us the ACF:

$$\begin{aligned}\rho(1) &= \frac{(1+\phi\theta)(\phi+\theta)}{1+\theta^2+2\phi\theta}\\ \rho(\tau) &= \phi\rho(\tau-1) \quad \tau > 1\end{aligned}$$

The autocorrelations display negative decay for $\tau > 1$, with oscillations when ϕ is negative (as in AR(1)). However $\rho(1)$ depends on both ϕ and θ, with its sign determined by the sign of $phi + \theta$.

1.3 Prediction

Given a set of observations for y_t, $t = 1, \ldots, T$, the expected value of $y_{t+\ell}$ conditional on the information at time $t = T$ is

$$\tilde{y}_{T+\ell|T} = \mathbb{E}[y_{T+\ell}|Y_T] = \mathbb{E}_T[y_{T+\ell}]$$

where Y_T is the information set. The conditional expectation is an optimal predictor in the sense that it minimises the mean square error.

Proof. The estimation error can be written as:

$$y_{T+\ell} - \hat{y}_{T+\ell|T} = [y_{T+\ell} - \mathbb{E}_T[y_{T+\ell}]] + [\mathbb{E}_T[y_{T+\ell}] - \hat{y}_{T+\ell|T}]$$

Since the second term is fixed at T, on squaring this term and taking expectations the cross terms disappear leaving

$$\begin{aligned}MSE(\hat{y}_{T+\ell|T}) &= \mathbb{E}[(y_{T+\ell} - \mathbb{E}_T[y_{T+\ell}])^2] + \mathbb{E}[(\mathbb{E}_T[y_{T+\ell}] - \hat{y}_{T+\ell|T})^2]\\ &= Var(y_{T+\ell}|Y_T) + [\hat{y}_{T+\ell|T} - \mathbb{E}(y_{T+\ell}|Y_T)]^2\end{aligned}$$

The dirts term (the conditional variance of $y_{T+\ell}$) does not depend on $\hat{y}_{T+\ell|T}$. Hence the minimum mean square estimate (MMSE) is given by the conditional mean. □

A predictor is linear if it is a linear combination of past observations. For MSE we use the infinite MA representation. Any such predictor can be written as

$$\hat{y}_{T+\ell|T} = \sum_{j=0}^{\infty} \psi_t + j^* \varepsilon_{T-j}$$

where the ψ_j^*'s are pre-specified weights. The predictor is unbiased in the sense that the unconditional expectation of the predictor

$$y_{T+\ell} - \hat{y}_{T+\ell|T} = \varepsilon_{T+\ell} + \psi_1 \varepsilon_{T+\ell-1} + \cdots + \psi_{\ell-1}\varepsilon_{T+1} + (\psi_\ell - \psi_\ell^*)\varepsilon_T + (\psi_{\ell+1} - \psi_{\ell+1}^*)\varepsilon_{T-1} + \ldots$$

A sufficient condition for ARMA predictions to be linear is that the disturbances are independent (uncorrelatedness not enough).
How do we construct an MMSE of a future observation from an ARMA model? We assume the parameters are known and the disturbances are serially independent with mean zero and constant variance σ^2. An ARMA(p,q) at time $T + \ell$ can be written as

$$y_{T+\ell} = \phi_1 y_{T+\ell-1} + \cdots + \phi_p y_{T+\ell-p} + \theta_1 \varepsilon_{T+\ell-1} + \cdots + \theta_q \varepsilon_{T+\ell-q} + \varepsilon_{T+\ell}$$

The MMSE of a future observation is its expectation conditional on information at T. Since ε_t are independent they cannot be predicted beyond T, thus have zero conditional expectation. This yields

$$\tilde{y}_{T+\ell|T} = \phi_1 \tilde{y}_{T+\ell-1|T} + \cdots + \phi_p \tilde{y}_{T+\ell-p|T} + \tilde{\varepsilon}_{T+\ell|T} + \cdots + \theta_q \tilde{\varepsilon}_{T+\ell-q|T}$$

where $\tilde{y}_{T+j|T} = y_{T+j}$ for $j \leq 0$ and

$$\tilde{\varepsilon}_{T+j|T} = \begin{cases} 0 & \text{if } j > 0 \\ \varepsilon_{T+j} & \text{if } j \leq 0 \end{cases}$$

This provides a recursion for computing optimal predictions of future observations.

Example (AR(1)). For an AR(1) process $\tilde{y}_{T+\ell|T} = \phi \tilde{y}_{T+\ell-1|T}$. The starting value is $\tilde{y}_{T|T} = y_T$. Thus and so

$$\tilde{y}_{T+\ell|T} = \phi^\ell y_T \quad \ell = 1, 2, \ldots$$

The predicted values decline exponentially towards zero; the forecast function has exactly same form as the autocovariance function. This makes sense intuitively: less correlation means less ability to forecast.

Example (MA(1)). For an MA(1) process at time T+1 $y_{T+1} = \varepsilon_{T+1} + \theta\varepsilon_T$. Thus taking conditional expectations $\tilde{y}_{T+1|T} = \theta\varepsilon_T$. For $\ell > 1$ $\tilde{y}_{T+\ell|T} = 0$ and so the current observation is of no help predicting more than one period ahead.

To find the prediction MSE, we use the infinite MA representation

$$y_{T+\ell} = \sum_{j=0}^{\ell-1} \psi_j \varepsilon_{T+\ell-j} + \sum_{j=0}^{\infty} \psi_{\ell+j} \varepsilon_{T-j}, \quad IID(0, \sigma^2)$$

The first term is unknown at t, whilst the second term is known. Taking expectations conditional on Y_T shows that the second term gives an expression for the MMSE of $y_{T+\ell}$:

$$\tilde{y}_{T+\ell|T} = \sum_{j=0}^{\infty} \psi_{\ell+j} \varepsilon_{T-j}$$

Thus the first term of the infinite MA expansion is the error in predicting ℓ steps ahead, and its variance is the conditional variance of $y_{T+\ell}$ (which is also the prediction MSE).

$$MSE(\tilde{y}_{T+\ell|T}) = Var_T(y_{T+\ell}) = (1 + \psi_1^2 + \cdots + \psi_{\ell-1}^2)\sigma^2$$

Example. For the AR(1) process, $\psi_{\ell+j} = \phi^{\ell+j}$ and solve

$$\tilde{y}_{T+\ell|T} = \sum_{j=0}^{\infty} \phi^{\ell+j}\varepsilon_{T-j} = \phi^{\ell}\sum_{j=0}^{\infty}\phi^{j}\varepsilon_{T-j} = \phi^{\ell}y_T$$

The MSE is

$$MSE(\tilde{y}_{T+\ell|T}) = (1+\phi^2+\cdots+\phi^{2(\ell-1)})\sigma^2 = \frac{1-\phi^{2\ell}}{1-\phi^2}\sigma^2$$

As $\ell \to \infty$ the MSE converges to $\sigma^2/(1-\phi^2)$, the unconditional variance of the process.

Making the additional assumption that the shocks are normal (thus the conditional distribution of $y_{t+\ell}$ is normal), a 95% confidence interval is given by

$$CI(y_{T+\ell})_{0.95} = \tilde{y}_{T+\ell|T} \pm 1.96\sigma\sqrt{1+\sum_{j=1}^{\ell-1}\psi_j^2}$$

1.4 Skip sampling and temporal aggregation

Suppose a model for $y_t^\dagger$ is defined for $t = 1, 2, \ldots, T$ but observations are only available at times $t = \delta, 2\delta, \ldots, T$. For example a model may be formulated at the monthly level but observations are only observed quarterly.

Example. Let $y_t^\dagger$ be an AR(1) model and assume sample size T/δ is an integer. Then

$$\begin{aligned} y_t^\dagger &= \phi y_{t-1}^\dagger + \varepsilon_t \\ &== \phi^2 y_{t-2}^\dagger + \phi\varepsilon_{t-1} + \varepsilon_t \\ &= \ldots \\ &= \phi^\delta y_{t-\delta}^\dagger + \sum_{j=0}^{\delta-1}\phi^j\varepsilon_{t-j} \end{aligned}$$

The observation model is then:

$$\begin{aligned} y_\tau &= \phi^\delta y_{\tau-1} + \sum_{j=0}^{\delta-1}\phi^j\varepsilon_{\tau-j} \\ &:= \phi^\delta y_{\tau-1} + \eta_\tau \end{aligned}$$

where $y_\tau = y_{\delta\tau}^\dagger$ and η_τ is serially uncorrelated with zero mean and

$$Var(\eta_\tau) = \frac{1-\phi^{2\delta}}{1-\phi^2}$$

The observation model is still AR(1) with the same unconditional variance, however the autocorrelations become smaller as δ increases.

Question 1

Show that an $MA(q)$ model for $y_t^\dagger$ is WN when $\delta > q$.

Solution:-

$$y_t^\dagger = \varepsilon_t + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j}$$

$$y_{t+\delta}^\dagger = \varepsilon_{t+\delta} + \sum_{j=1}^{q} \theta_j \varepsilon_{t+\delta-j}$$

When $\delta > q$ the two observations are independent (since there are no common terms); we can write the observation model as

$$y_\tau = \eta_\tau$$

where $\eta_\tau = \varepsilon_t + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j}$ which is white noise since the ε_t are independent. Thus

$$\eta_\tau \sim WN(0, \sigma^2(1 + \theta_1^2 + \cdots + \theta_q^2))$$

Note:-

This result is proved generally in Brewer (1973) who shows that when $y_t^\dagger$ is an ARMA(p,q) process the observation model y_τ is $ARMA(p, [\delta^{-1}(p(\delta-1)+q)])$ where [] denotes the floor function. As δ grows the model converges to an $ARMA(p,p)$ process for $q \geq p$ and to an $ARMA(p, p-1)$ process for $q < p$.

The effect of temporal aggregation when

$$y_\tau = \sum_{j=0}^{\delta-1} y_{\delta\tau-j}^\dagger$$

is such that when $y_t^\dagger$ is $ARMA(p,q)$, the corresponding model for y_τ is $ARMA(p, [\delta^{-1}((p+1)(\delta-1)+q)])$.

Example. An aggregated AR(1) process is ARMA(1,1)
PROVE LATER

1.5 Tests on Sample Autocorrelations

The sample autocorrelations are $r(\tau) = c(\tau)/c(0)$ where $c(\tau) = T^{-1}\sum_{t=\tau+1}^{T}(y_t - \bar{y})(y_{t-\tau} - \bar{y})$ is the sample autocovariance.
They are asymptotically normal with mean zero and variance $1/T$ when the observations are independent. DOES THIS NEED PROVING?
Deriving this test under the assumption of uncorrelatedness is more difficult, hence tests are based on independence. We run a test on the sample autocorrelation at a particular lag τ by treating $T^{1/2}r(\tau)$ as a standard normal variate.

Note:-

A test on a particular sample autocorrelation is only valid if the lag is specified in advance. This implies some prior knowledge of the nature of the series. For example, with quarterly data a test of the significance of r(4) would clearly be relevant. However, seasonality aside, such prior knowledge is likely to be the exception rather than the rule, and formal tests on single autocorrelations are generally restricted to r(1).

At the 5% level we reject the null hypothesis of no autocorrelation if $|T^{1/2}r(\tau)| > 1.96$, thus it is common to plot two lines on the sample correlogram at $\pm 2\sqrt{T}$.

Definition 1.5.1: von Neumann ratio

$$VNR = \frac{T}{T-1}\left[\frac{\sum_{t=2}^{T}(y_t - y_{t-1})^2}{\sum_{t=1}^{T}(y_t - \bar{y})^2}\right]$$

where $\bar{y}$ is the sample mean.

Claim 1.5.1.

$$VNR \approx 2[1 - r(1)]$$

Proof. We first note that $(y_t - y_{t-1})^2 = (y_t - \bar{y} - (y_{t-1} - \bar{y}))^2$. Thus

$$\begin{aligned}
VNR &= \frac{T}{T-1}\left[\frac{\sum_{t=2}^{T}(y_t - y_{t-1})^2}{\sum_{t=1}^{T}(y_t - \bar{y})^2}\right] \\
&= \frac{T}{T-1}\left[\frac{\sum_{t=2}^{T}(y_t - \bar{y})^2 + \sum_{t=2}^{T}(y_{t-1} - \bar{y})^2 - 2\sum_{t=2}^{T}(y_t - \bar{y})(y_{t-1} - \bar{y})}{\sum_{t=1}^{T}(y_t - \bar{y})^2}\right] \\
&= \frac{T}{T-1}\frac{\sum_{t=1}^{T}(y_t - \bar{y})^2 - (y_1 - \bar{y})^2 + \sum_{t=1}^{T}(y_{t-1} - \bar{y})^2 - (y_T - \bar{y})^2}{\sum_{t=1}^{T}(y_t - \bar{y})^2} \\
&- 2\frac{T}{T-1}\frac{\sum_{t=1}^{T}(y_t - \bar{y})(y_{t-1} - \bar{y}) - (y_1 - \bar{y})(y_0 - \bar{y})}{\sum_{t=1}^{T}(y_t - \bar{y})^2} \\
&= \frac{T}{T-1}\left[2 - 2r(1) + \frac{(y_1 - \bar{y})^2 + (y_T - \bar{y})^2 + (y_1 - \bar{y})(y_0 - \bar{y})}{\sum_{t=1}^{T}(y_t - \bar{y})^2}\right] \\
&\approx 2[1 - r(1)]
\end{aligned}$$

The approximation arises because of the treatment of the end point. Its effect is negligible in moderate or large samples □

When $y_t \sim NID(0, \sigma^2)$, the small sample distribution of VNR is known. When $r(1) = 0$, VNR is approximately equal to two, but as r(1) tends towards one, VNR tends towards zero.

Definition 1.5.2: Durbin-Watson statistic

When a static regression is fitted, the Durbin-Watson statistic is often used to test for serial correlation. It is defined in terms of the residuals as

$$DW = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2}$$

Definition 1.5.3: Portmanteau test statistic

$$Q(p) = T\sum_{\tau=1}^{p} r(\tau)^2$$

Definition 1.5.4: Ljung-Box test statistic

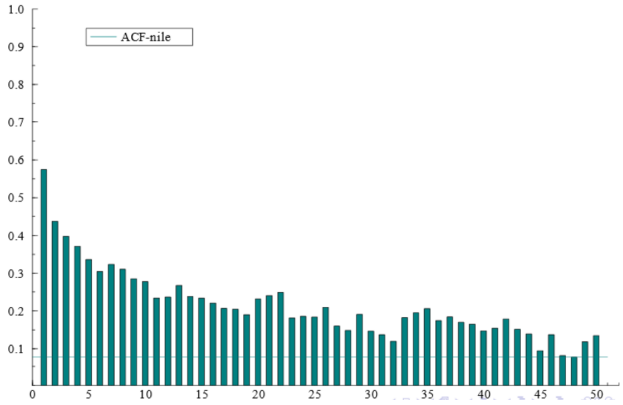
$$Q_{LB}(P) = T(T+2)\sum_{\tau=1}^{p}\frac{r(\tau)^2}{T-\tau}$$

Both of the above are asymptotically distributed as $\chi^2(p)$ when the observations are independent, however the LB test approximates a χ^2 better in finite sample.

Note:-

The choice of P is somewhat arbitrary, when P is large the stat captures potentially high autocorrelations at greater lags, but at the expense of power.

1.6 Long memory



This is a correlogram of the lowest annual water level of the river Nile (fuck me, bro does not shut up about the Nile). It displays a sharp fall initially followed by a much slower, persistent decline. This pattern is called long memory, which can be produced by a **fractionally integrated** process.
ARIMA(p,d,q) has $(1-L)^d$ generated by an ARMA(p,q), but suppose d is not an integer:

$$(1-L)^d y_t = \varepsilon_t$$

where we define the fractional difference operator by the binomial expansion (for any real $d > 1$)

$$\Delta^d = (1-L)^d = 1 - dL - \frac{1}{2}d(1-d)L^2 - \cdots$$

Thus we can write the process in infinite AR form:

$$y_t = dy_{t-1} + \frac{1}{2}d(1-d)y_{t-2} + \cdots + \varepsilon_t$$

The coefficients die away very slowly, so a large number of lags is needed to approximate y_t with an AR. We can still use this model for forecasting however.
The process is stationary if $d < 1/2$. Then

$$\rho(\tau) = \frac{\Gamma(1-d)\Gamma(\tau+d)}{\Gamma(d)\Gamma(\tau+1-d)}$$

1.7 Maximum Likelihood

The joint density can be broken down into a set of one-step ahead predictive distributions by writing

$$p(y;\psi) = \prod_{t=1}^{T} p(y_t|Y_{t-1})$$

where Y_{t-1} is the information set at time $t-1$ and $p(y_1)$ is interpreted as $p(y_1)$, the unconditional density of y_1. For example with $T = 3$:

$$p(y_3, y_2, y_1) = p(y_3|y_2, y_1)p(y_2|y_1)p(y_1)$$

The conditional densities are independent of each other, enabling ML to be used in the same way as for independent observations.

1.7.1 Autoregressive models

In the stationary Gaussian AR(1) model (with zero mean), the disturbance of y_t, conditional on y_{t-1}, is normal with mean ϕy_{t-1} and variance σ^2. Thus $p(y_t|Y_{t-1})$ is $N(\phi y_{t-1}, \sigma^2)$. The log-likelihood function can thus be derived:

$$\begin{aligned} p(y_t;\phi,\sigma^2) &= \left(\prod_{t=2}^{T}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}(y_t-\phi y_{t-1})^2\right\}\right)p(y_1) \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{T-1}{2}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{t=2}^{T}(y_t-\phi y_{t-1})^2\right\}p(y_1) \\ \Rightarrow \ln p(y_t;\phi,\sigma^2) &= -\frac{T-1}{2}\ln(2\pi)-\frac{T-1}{2}\ln(\sigma^2)-\frac{1}{2\sigma^2}\sum_{t=2}^{T}(y_t-\phi y_{t-1})^2+\ln p(y_1) \end{aligned}$$

The last term is the unconditional distribution of y_1, which is normal with mean zero and variance $\sigma^2/(1-\phi^2)$. Thus

$$\begin{aligned} p(y_1) &= \frac{1}{\sqrt{2\pi\sigma^2/(1-\phi^2)}}\exp\left\{-\frac{y_1^2}{2\sigma^2/(1-\phi^2)}\right\} \\ \Rightarrow \ln p(y_1) &= -\frac{1}{2}\ln(2\pi)-\frac{1}{2}\ln(\sigma^2)+\frac{1}{2}\ln(1-\phi^2)-\frac{1}{2\sigma^2}(1-\phi^2)y_1^2 \end{aligned}$$

Thus

$$\ln p(y_t;\phi,\sigma^2) = -\frac{T}{2}\ln(2\pi)-\frac{T}{2}\ln(\sigma^2)-\frac{1}{2}\ln(1-\phi^2)-\frac{1}{2\sigma^2}(1-\phi^2)y_1^2-\frac{1}{2\sigma^2}\sum_{t=2}^{T}(y_t-\phi y_{t-1})^2$$

The ML estimator of ϕ is not linear because the likelihood function is a cubic equation in ϕ. On the other hand, if the first observation is treated as though it were fixed, the third and fourth terms can be dropped and T replaced by T-1 in the first two, i.e. we set $p(y_1) = 1$:

$$\ln p(y_t;\phi,\sigma^2) = -\frac{T-1}{2}\ln(2\pi)-\frac{T-1}{2}\ln(\sigma^2)-\frac{1}{2\sigma^2}\sum_{t=2}^{T}(y_t-\phi y_{t-1})^2$$

The resulting ML estimator of ϕ is then given by a regression of y_t on y_{t-1}. Since the first two terms of the above log-likelihood are independent of ϕ, maximising the likelihood function is equivalent to minimising the sum of squares:

$$S(\phi) = \sum_{t=2}^{T}(y_t-\phi y_{t-1})^2$$

The ML estimator is thus given by

$$\tilde{\phi} = \frac{\sum_{t=2}^{T} y_t y_{t-1}}{\sum_{t=2}^{T} y_{t-1}^2}$$

and the variance is estimated by

$$avar(\tilde{\phi}) = \frac{\tilde{\sigma}^2}{\sum_{t=2}^{T} y_{t-1}^2}, \quad \text{where } \tilde{\sigma}^2 = \frac{1}{T}\sum_{t=2}^{T}(y_t-\tilde{\phi}y_{t-1})^2$$

Asymptotics

$$\tilde{\phi} = \frac{\sum_{t=2}^{T} y_t y_{t-1}}{\sum_{t=2}^{T} y_{t-1}^2} = \frac{\sum_{t=2}^{T} \phi y_{t-1}^2 + \varepsilon_t y_{t-1}}{\sum_{t=2}^{T} y_{t-1}^2} = \phi + \frac{\sum_{t=2}^{T} \varepsilon_t y_{t-1}}{\sum_{t=2}^{T} y_{t-1}^2}$$

Thus

$$\sqrt{T}(\tilde{\phi} - \phi) = \frac{\frac{1}{\sqrt{T}} \sum_{t=2}^{T} \varepsilon_t y_{t-1}}{\frac{1}{T} \sum_{t=2}^{T} y_{t-1}^2}$$

Recall that $\mathbb{E}[y_t^2] = Var(y_t) = \sigma^2/(1-\phi^2)$.

$$\frac{1}{T} \sum_{t=2}^{T} y_{t-1}^2 \xrightarrow{p} \frac{\sigma^2}{1-\phi^2}$$

[1]Furthermore

$$\frac{1}{\sqrt{T}} \sum_{t=2}^{T} \varepsilon_t y_{t-1} \xrightarrow{d} N(\mathbb{E}(\varepsilon_t y_{t-1}), Var(\varepsilon_t y_{t-1}))$$

By independence of ε_t across time, $\mathbb{E}[\varepsilon_t y_{t-1}] = 0$. Further,

$$Var(\varepsilon_t y_{t-1}) = \mathbb{E}[\varepsilon_t^2 y_{t-1}^2] = \mathbb{E}[\varepsilon_t^2]\mathbb{E}[y_{t-1}^2] = \sigma^2 \frac{\sigma^2}{1-\phi^2} = \frac{\sigma^4}{1-\phi^2}$$

where we can split the expectation because ε_t is independent of y_{t-1}. Thus

$$\frac{1}{\sqrt{T}} \sum_{t=2}^{T} \varepsilon_t y_{t-1} \xrightarrow{d} N(0, \frac{\sigma^4}{1-\phi^2})$$

Applying the CMT:

$$\begin{aligned}\sqrt{T}(\tilde{\phi} - \phi) &= \frac{\frac{1}{\sqrt{T}} \sum_{t=2}^{T} \varepsilon_t y_{t-1}}{\frac{1}{T} \sum_{t=2}^{T} y_{t-1}^2} \\ &\xrightarrow{d} \frac{N(0, \frac{\sigma^4}{1-\phi^2})}{\frac{\sigma^2}{1-\phi^2}} \sim N\left(0, \frac{\sigma^4}{1-\phi^2} \frac{(1-\phi^2)^2}{(\sigma^2)^2}\right) \sim N(0, 1-\phi^2)\end{aligned}$$

In other words, $\tilde{\phi}$ is asymptotically normal with mean ϕ and variance $Avar(\tilde{\phi}) = \frac{1-\phi^2}{T}$.

1.7.2 Moving Average models

In the MA(1) model,

$$y_t = \varepsilon_t + \theta \varepsilon_{t-1}, \quad \varepsilon_t \sim IID(0, \sigma^2), quadt = 1, \ldots, T$$

the distribution of y_t conditional on the disturbance in the previous period is normal with mean $\theta\varepsilon_{t-1}$ and variance σ^2. The full set of disturbances can be computed as $\varepsilon_t = y_t - \theta\varepsilon_{t-1}$, however without knowledge of ε_0 we compute a set of residuals

$$\varepsilon_t(\theta; \varepsilon_0) = y_t - \theta \varepsilon_{t-1}(\theta; \varepsilon_0), \quad t = 1, \ldots, T$$

with $\varepsilon_0(\theta)$ taken to be fixed at zero. The log likelihood takes the same form as the AR(1), where we seek to minimise a sum of squares:

$$S(\theta) = \sum_{t=1}^{T} \varepsilon_t^2(\theta; \varepsilon_0)$$

This is known as the **conditional sum of squares (CSS)** estimator since it is conditional on setting $\varepsilon_0 = 0$.

[1]Source: Trust me bro

The likelihood equations are non-linear because the derivative of $\varepsilon_t(\theta;\varepsilon_0)$ depends on θ. This is in contrast to the AR(1) model where the derivative of the residual with respect to ϕ is $-y_{t-1}$. We thus need a different method to minimise $S(\theta)$.
Only one parameter is involved here, so we could just do a grid search over (-1,1). [2] However, for more general models this may not be viable, here we use Gauss-Newton iteration. For the MA(1) differentiating the error expression gives

$$\frac{\partial\varepsilon_t(\theta;\varepsilon_0)}{\partial\theta} = -\theta\frac{\partial\varepsilon_{t-1}(\theta;\varepsilon_0)}{\partial\theta} - \varepsilon_{t-1}(\theta;\varepsilon_0), \quad t = 1,\ldots,T$$

Because $\varepsilon_0(\theta)$ is fixed, it follows that $\frac{\partial\varepsilon_0(\theta)}{\partial\theta} = 0$. Thus the derivatives are produced by a recursion running parallel to the computation of the residual set. The algorithm proceeds by updating an estimate of θ, $\hat{\theta}$, from a regression of $\varepsilon_t(\theta;\varepsilon_0)$ on $z_t(\hat{\theta}) = \frac{\partial\varepsilon_t(\theta;\varepsilon_0)}{\partial\theta}$:

$$\theta^* = \hat{\theta} + \frac{\sum_{t=1}^{T}\varepsilon_t(\hat{\theta};\varepsilon_0)z_t(\hat{\theta})}{\sum_{t=1}^{T}z_t(\hat{\theta})^2}$$

Asymptotics
Provided the model is both stationary and invertible, the CSS estimator of θ has the same asymptotic distribution as the (infeasible) estimator based on knowledge of initial disturbances. For the MA(1) model:

$$\frac{\partial\varepsilon_t}{\partial\theta} = -\theta\frac{\partial\varepsilon_{t-1}}{\partial\theta} - \varepsilon_{t-1}, \quad t = \ldots,-1,0,1,\ldots,T$$

By writing $z_t = \frac{\partial\varepsilon_t}{\partial\theta}$, it becomes clear we have an AR(1): $z_t = -\theta z_{t-1} - \varepsilon_{t-1}$. Hence $\mathbb{E}[z_t^2] = Var(z_t) = \frac{\sigma^2}{1-\theta^2}$ leading to the result that $\tilde{\theta}$ is asymptotically normal with mean θ and variance $Avar(\tilde{\theta}) = \frac{1-\theta^2}{T}$.

[2] We don't necessarily need to constrain it, we could solve for the global max of the likelihood. If we get something outside (-1,1) we can just take reciprocal.

2 Unobserved Components. State Space Models. Kalman Filter.

2.1 Unobserved Components

A time series model may sometimes be formulated in terms of components. These components are not observed directly, but are assumed to have ARMA representations. A simple model, consisting of an AR(1) embedded in white noise, is:

$$\begin{aligned} y_t &= \mu_t + \varepsilon_t, \quad \varepsilon_t \sim \text{NID}(0, \sigma_\varepsilon^2), \quad t = 1, \ldots, T \\ \mu_t &= \phi\mu_{t-1} + \eta_t, \quad \eta_t \sim \text{NID}(0, \sigma_\eta^2), \quad |\phi| < 1 \end{aligned}$$

where η_t and ε_t are independent. When $|\phi| < 1$, the process is stationary, when $|\phi| = 1$ gives a random walk + noise. An important value $q = \frac{\sigma_\eta^2}{\sigma_\varepsilon^2}$ is the signal-to-noise ratio.
From lecture 1, we know the autocovariances of the AR(1) process:

$$\gamma_\mu(\tau) = \phi^\tau \gamma_\mu(0) = \phi^\tau \frac{\sigma_\eta^2}{1-\phi^2}$$

and because $\gamma_\varepsilon(\tau) = 0$ for $\tau \neq 0$, we have:

$$\begin{aligned} \gamma_y(\tau) &= \gamma_\mu(\tau) + \gamma_\varepsilon(\tau) \\ \Rightarrow \rho_y(\tau) &= \frac{\gamma_y(\tau)}{\gamma_y(0)} \\ &= \frac{\gamma_\mu(\tau)}{\gamma_\mu(0) + \gamma_\varepsilon(0)} \\ &= \frac{\phi^\tau \sigma_\eta^2 (1-\phi^2)^{-1}}{\sigma_\eta^2(1-\phi^2)^{-1} + \sigma_\varepsilon^2} \quad \tau = 1, 2, \ldots \end{aligned}$$

We can substitute ϕy_{t-1} to derive the reduced form:

$$\begin{aligned} y_t &= \mu_t + \varepsilon_t \\ &= \phi\mu_{t-1} + \eta_t + \varepsilon_t \\ &= \phi(y_{t-1} - \varepsilon_{t-1}) + \eta_t + \varepsilon_t \\ y_t - \phi y_{t-1} &= \eta_t + \varepsilon_t - \phi\varepsilon_{t-1} \end{aligned}$$

This is equivalent to an ARMA(1,1) process:

$$y_t - \phi y_{t-1} = \nu_t + \theta\nu_{t-1}.$$

The MA parameter can be founded by equating the autocovariances of the right hand side of the two processes:

Reduced Form

$$\begin{aligned} \gamma_y(0) &= Var(\eta_t) + Var(\varepsilon_t) + \phi^2 Var(\varepsilon_t) \\ &= \sigma_\eta^2 + \sigma_\varepsilon^2(1+\phi^2) \\ \gamma_y(1) &= Cov(-\phi\varepsilon_{t-1}, \varepsilon_{t-1}) \\ &= -\phi\sigma_\varepsilon^2 \end{aligned}$$

ARMA(1,1)

$$\begin{aligned} \gamma_y(0) &= Var(\nu_t) + \theta^2 Var(\nu_t) \\ &= \sigma_\nu^2(1+\theta^2) \\ \gamma_y(1) &= Cov(\nu_t + \theta\nu_{t-1}, \nu_{t-1} + \theta\nu_{t-2}) \\ &= \theta\sigma_\nu^2 \end{aligned}$$

Equating the expression for $\gamma_y(1)$ gives:

$$\sigma_\nu^2 = -\frac{\phi\sigma_\varepsilon^2}{\theta}$$

Define the signal to noise ratio $q = \frac{\sigma_\eta^2}{\sigma_\varepsilon^2}$, then equate the expressions for $\gamma_y(0)$:

$$\begin{aligned}
\sigma_\eta^2 + \sigma_\varepsilon^2(1+\phi^2) &= \sigma_\nu^2(1+\theta^2) \\
\sigma_\eta^2 + \sigma_\varepsilon^2(1+\phi^2) &= -\frac{\phi\sigma_\varepsilon^2}{\theta}(1+\theta^2) \\
\frac{\sigma_\eta^2}{\sigma_\varepsilon^2} + 1 + \phi^2 &= -\frac{\phi}{\theta}(1+\theta^2) \\
\Rightarrow 0 &= \phi\theta^2 + (q+1+\phi^2)\theta + \phi \\
\Rightarrow \theta &= \frac{-(q+1+\phi^2) \pm \sqrt{(q+1+\phi^2)^2 - 4\phi^2}}{2\phi}
\end{aligned}$$

Drop the negative root, since this implies a $|\theta| > 1$. Thus

$$\theta = -\frac{q+1+\phi^2}{2\phi} + \frac{\sqrt{(q+1+\phi^2)^2 - 4\phi^2}}{2\phi}$$

Thus $\phi \leq \theta \leq 0$, but when $\theta = -\phi$, which corresponds to $q = 0$, the ARMA(1,1) has a common factor and reduces to white noise. When $\phi = 1$

$$\theta = -\frac{q+2}{2} + \frac{\sqrt{q^2+4q}}{2}$$

Alternatively we can write express q as a function of ϕ and θ:

$$q = -\frac{\phi\theta^2 + \phi}{\theta} - \phi^2 - 1 = -\frac{\phi + \phi\theta^2 + \theta + \theta\phi^2}{\theta} = -\frac{(\phi+\theta)(1+\phi\theta)}{\theta}$$

2.2 Filtering and Prediction

Prediction requires filtering the observations so as to give the best estimate of μ_t at the end of the sample. This estimate is then extrapolated. The filtered estimate at the end of the sample is known as the **nowcast**.

Example (Random Walk + Noise). The conditional distribution of $y_{t+\ell}$ is obtained by writing:

$$y_{T+\ell} = \mu_T + \sum_{j=1}^{\ell} \eta T + j + \varepsilon_{T+\ell}$$

Thus the MMSE predictor is $\tilde{y}_{T+\ell|T} = \mathbb{E}(\mu_{T+\ell}|Y_t) = \mathbb{E}(\mu_T|Y_t)$, and so the forecast function is a horizontal straight line.
To find the MSE write:

$$\begin{aligned}
y_{T+\ell} &= \mu_T + \sum_{j=1}^{\ell} \eta T + j + \varepsilon_{T+\ell} \\
&= \mathbb{E}(\mu_T|Y_t) + (\mu_T - \mathbb{E}(\mu_T|Y_t)) + \sum_{j=1}^{\ell} \eta T + j + \varepsilon_{T+\ell}
\end{aligned}$$

The conditional variance of $y_{T+\ell}$ is:

$$\sigma_{T+\ell|T}^2 = MSE(\tilde{y}_{T+\ell|T}) = Var(\mu_T|Y_t) + \ell\sigma_\eta^2 + \sigma_\varepsilon^2$$

The MSE increases linearly with the forecast horizon, with $Var(\mu_T|Y_t)$ being the price paid for not knowing μ_T. Given σ_η^2 and σ_ε^2, a 95% prediction interval for $y_{T+\ell}$ is:

$$\tilde{y}_{T+\ell|T} \pm 1.96\sigma_{T+\ell|T}$$

How do we filter for AR(1) + noise? Filtering is a two step process, firstly we obtain the prediction equations, then we update the predictions using new observations.

Prediction

Suppose we know the mean and variance of μ_{t-1} conditional on Y_{t-1}: $\mu_{t-1}|Y_{t-1} \sim N(\mu_{t-1|t-1}, p_{t-1|t-1})$. We can calculate the conditional mean and variance of μ_t as follows:

$$\begin{aligned}\mu_{t|t-1} &= \mathbb{E}(\mu_t|Y_{t-1}) \\ &= \mathbb{E}(\phi\mu_{t-1} + \eta_t|Y_{t-1}) \\ &= \phi\mathbb{E}(\mu_{t-1}|Y_{t-1}) \\ &= \phi\mu_{t-1|t-1}\end{aligned}$$

We can write $\mu_t = \phi\mu_{t-1|t-1} + \phi(\mu_{t-1} - \mu_{t-1|t-1}) + \eta_t$. Thus:

$$\mu_t - \mu_{t|t-1} = \mu_t - \phi\mu_{t-1|t-1} = \phi(\mu_{t-1} - \mu_{t-1|t-1}) + \eta_t$$

and we can derive the conditional variance as:

$$\begin{aligned}p_{t|t-1} &= Var(\mu_t|Y_{t-1}) \\ &= \mathbb{E}[(\mu_t - \mu_{t|t-1})^2|Y_{t-1}] \\ &= \mathbb{E}[(\phi(\mu_{t-1} - \mu_{t-1|t-1}) + \eta_t)^2|Y_{t-1}] \\ &= \phi^2\mathbb{E}[(\mu_{t-1} - \mu_{t-1|t-1})^2|Y_{t-1}] + \sigma_\eta^2, \quad \text{since } \eta_t \perp Y_{t-1} \\ &= \phi^2 p_{t-1|t-1} + \sigma_\eta^2 \\ &\Rightarrow \mu_t|Y_{t-1} \sim N(\phi\mu_{t-1|t-1}, \phi^2 p_{t-1|t-1} + \sigma_\eta^2)\end{aligned}$$

The conditional distribution of $y_t = \mu_t + \varepsilon_t$ is normal with mean and variance:

$$\begin{aligned}\mathbb{E}(y_t|Y_{t-1}) &= \mathbb{E}(\mu_t|Y_{t-1}) + \mathbb{E}(\varepsilon_t|Y_{t-1}) \\ &= \phi\mu_{t-1|t-1} \\ Var(y_t|Y_{t-1}) &= \mathbb{E}[(y_t - \mu_{t|t-1})^2|Y_{t-1}] \\ &= \mathbb{E}[(\mu_t - \mu_{t|t-1})^2|Y_{t-1}] + \mathbb{E}[\varepsilon_t^2|Y_{t-1}] \\ &= \phi^2 p_{t-1|t-1} + \sigma_\eta^2 + \sigma_\varepsilon^2 \\ \Rightarrow y_t|Y_{t-1} &\sim N(\phi\mu_{t-1|t-1}, \phi^2 p_{t-1|t-1} + \sigma_\eta^2 + \sigma_\varepsilon^2)\end{aligned}$$

Updating

Having obtained the prediction equations, the next step is to update the estimates of the mean and variance of μ_t to take account of the information in the new observation. I.e., we want to find $\mu_t|Y_t \sim N(\mu_{t|t}, p_{t|t})$.

The current information is contained in the conditional distribution of $(\mu_t, y_t)'$ which is bivariate normal:

$$\begin{aligned}\mathbb{E}\mu_t|Y_{t-1} = \mu_{t|t-1}, &\quad \mathbb{E}y_t|Y_{t-1} = \mu_{t|t-1} \\ Var(\mu_t|Y_{t-1}) = p_{t|t-1}, &\quad Var(y_t|Y_{t-1}) = p_{t|t-1} + \sigma_\varepsilon^2 \\ Cov(\mu_t, y_t|Y_{t-1}) &= Cov(\mu_t, \mu_t + \varepsilon_t|Y_{t-1}) = p_{t|t-1} \\ \Rightarrow \begin{bmatrix}\mu_t \\ y_t\end{bmatrix} |Y_{t-1} &\sim N\left(\begin{bmatrix}\mu_{t|t-1} \\ \mu_{t|t-1}\end{bmatrix}, \begin{bmatrix}p_{t|t-1} & p_{t|t-1} \\ p_{t|t-1} & p_{t|t-1} + \sigma_\varepsilon^2\end{bmatrix}\right)\end{aligned}$$

This updated distribution is the distribution of μ_t is the distribution of μ_t conditional on Y_{t-1}, and y_t. To calculate $\mu_{t|t}$ and $p_{t|t}$, we invoke the following lemma:

Lemma 2.2.1 (Marginal and conditional normal distributions).

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

Consider two distributions:

$$x_1 \sim N(\mu_1, \Sigma_{11}) \quad x_2 \sim N(\mu_2, \Sigma_{22})$$

The conditional distribution is given by:

$$x_1 | x_2 \sim N(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$$

Applying this lemma we can find the conditional distribution of μ_t given y_t:

$$\begin{aligned} \mu_{t|t} &= \mu_{t|t-1} + \frac{Cov(\mu_t, y_t|Y_{t-1})}{Var(y_t|Y_{t-1})}(y_t - \mu_{t|t-1}) \\ &= \mu_{t|t-1} + \frac{p_{t|t-1}}{p_{t|t-1} + \sigma_\varepsilon^2}(y_t - \mu_{t|t-1}) \\ p_{t|t} &= p_{t|t-1} - \frac{Cov(\mu_t, y_t|Y_{t-1})^2}{Var(y_t|Y_{t-1})} \\ &= p_{t|t-1} - \frac{p_{t|t-1}^2}{p_{t|t-1} + \sigma_\varepsilon^2} \end{aligned}$$

The updating and prediction equations can thus be written together as a single set of recursions:

$$\mu_{t+1|t} = \phi\mu_{t|t} = \phi\mu_{t|t-1} + \phi\frac{p_{t|t-1}}{p_{t|t-1} + \sigma_\varepsilon^2}(y_t - \mu_{t|t-1})$$

$$p_{t+1|t} = \phi^2 p_{t|t} + \sigma_\eta^2 = \phi^2 p_{t|t-1} - \phi^2\frac{p_{t|t-1}^2}{p_{t|t-1} + \sigma_\varepsilon^2} + \sigma_\eta^2$$

These equations can be repeatedly applied as new observations become available. This is a special case of the Kalman filter (**???**). Indeed the first equation can be written as:

$$\mu_{t+1|t} = \phi\mu_{t|t-1} + k_t(y_t - \mu_{t|t-1}) \quad \text{where} \quad k_t = \phi\frac{p_{t|t-1}}{p_{t|t-1} + \sigma_\varepsilon^2}$$

is known as the **Kalman gain**.
How do we start the filter?
When $|\phi| < 1$, we can use the unconditional distribution of μ_0: $\mu_0 \sim N(0, \sigma_\eta^2/(1-\phi^2))$ and so $\mu_{0|0} = 0$ and $p_{0|0} = \sigma_\eta^2/(1-\phi^2)$.
When $|\phi| = 1$ we start with $\mu_1|y_1 \sim N(y_1, \sigma_\varepsilon^2)$. This is equivalent (**???**) to starting with $\mu_0 \sim N(0, \kappa)$ where $\kappa \to \infty$ (Diffuse prior). In both cases the filter converges to a steady state with $p_t = p > 0$.

Explanation. Indeed we can actually set an arbitrary value for μ in $\mu_0 \sim N(\mu, \kappa)$, to see this we first set up the predictive equation for p_1:

$$p_{t+1|t} = p_{t|t} + \sigma_\eta^2 \Rightarrow p_{1|0} = p_{0|0} + \sigma_\eta^2 = \kappa + \sigma_\eta^2$$

Now we update given y_1:

$$\begin{aligned} p_{1|1} &= p_{1|0} - \frac{p_{1|0}^2}{p_{1|0} + \sigma_\varepsilon^2} \\ &= \kappa + \sigma_\eta^2 - \frac{(\kappa + \sigma_\eta^2)^2}{\kappa + \sigma_\eta^2 + \sigma_\varepsilon^2} \\ &= (\kappa + \sigma_\eta^2)\left(1 - \frac{\kappa + \sigma_\eta^2}{\kappa + \sigma_\eta^2 + \sigma_\varepsilon^2}\right) \\ &= (\kappa + \sigma_\eta^2)\left(\frac{\sigma_\varepsilon^2}{\kappa + \sigma_\eta^2 + \sigma_\varepsilon^2}\right) \\ &= \frac{\kappa + \sigma_\eta^2}{\kappa + \sigma_\eta^2 + \sigma_\varepsilon^2}\sigma_\varepsilon^2 \\ \Rightarrow \lim_{\kappa\to\infty} p_{1|1} &= \sigma_\varepsilon^2 \end{aligned}$$

Now we can update μ:

$$\begin{aligned} \mu_{1|1} &= \mu 1|0 + \frac{p_{1|0}}{p_{1|0} + \sigma_\varepsilon^2}(y_1 - \mu_{1|0}) \\ &= mu + \frac{\kappa + \sigma_\eta^2}{\kappa + \sigma_\eta^2 + \sigma_\varepsilon^2}(y_1 - \mu) \\ \lim_{\kappa\to\infty} \mu_{1|1} &= \mu + y_1 - \mu = y_1 \end{aligned}$$

Since the conditional distribution is also normal, we get the distribution of $\mu_1|y_1 \sim N(y_1, \sigma_\varepsilon^2)$ as required. □

For a random walk the filter is an exponentially weighted moving average:

Example ($\phi = 1$)**.** First recall the updating and prediction equation of the variance:

$$p_{t+1|t} = \phi^2 p_{t|t-1} - \phi^2 \frac{p_{t|t-1}^2}{p_{t|t-1} + \sigma_\varepsilon^2} + \sigma_\eta^2$$

Indeed setting $\phi = 1$ gives:

$$\begin{aligned} p_{t+1|t} &= p_{t|t-1} - \frac{p_{t|t-1}^2}{p_{t|t-1} + \sigma_\varepsilon^2} + \sigma_\eta^2 \\ \sigma_\varepsilon^{-2} p_{t+1|t} &= \sigma_\varepsilon^{-2} p_{t|t-1} - \frac{\sigma_\varepsilon^{-2} p_{t|t-1}^2}{p_{t|t-1} + \sigma_\varepsilon^2} + \sigma_\varepsilon^{-2}\sigma_\eta^2 \\ &= \sigma_\varepsilon^{-2} p_{t|t-1} - \frac{(\sigma_\varepsilon^{-2})^2 p_{t|t-1}^2}{\sigma_\varepsilon^{-2} p_{t|t-1} + 1} + q \quad \text{since } q = \sigma_\varepsilon^{-2}\sigma_\eta^2 \end{aligned}$$

Re-define $p_{t|t-1} = \sigma_\varepsilon^{-2} p_{t|t-1}$ and we obtain the Ricatti equation:

$$p_{t+1|t} = p_{t|t-1} - \frac{p_{t|t-1}^2}{p_{t|t-1} + 1} + q$$

Setting $p_{t+1|t} = p_{t|t-1} = p$ as in a steady state gives the algebraic Riccati equation:

$$p = p - \frac{p^2}{p+1} + q = \frac{p}{p+1} + q$$

Note:-

For a general ϕ we can write:

$$p = \phi^2 \left(p - \frac{p^2}{p+1} \right) + q$$

Solving the quadratic Ricatti equation:

$$\begin{aligned}
&\frac{p^2}{p+1} = q \\
&\Rightarrow 0 = p^2 - qp - q \\
&\Rightarrow p = \frac{q \pm \sqrt{q^2+4q}}{2} \\
&\Rightarrow p = \frac{q + \sqrt{q^2+4q}}{2} \quad \text{since } p > 0
\end{aligned}$$

Recall the updating equation for the state (with $\phi = 1$), we can show that it takes the form of an EWNA:

$$\begin{aligned}
\mu_{t+1|t} &= \mu_{t|t-1} + \frac{p_{t|t-1}}{p_{t|t-1} + \sigma_\varepsilon^2}(y_t - \mu_{t|t-1}) \\
&:= \mu_{t|t-1} + \kappa(y_t - \mu_{t|t-1}) \\
&= (1-\kappa)\mu_{t|t-1} + \kappa y_t
\end{aligned}$$

with

$$\begin{aligned}
\kappa &= \frac{p_{t|t-1}}{p_{t|t-1} + \sigma_\varepsilon^2} \\
&= \frac{p}{p+1} \quad \text{with } p_{t|t-1} = \sigma_\varepsilon^{-2} p_{t|t-1} \text{ and } p_{t|t-1} = p
\end{aligned}$$

Recall

$$p = \frac{q + \sqrt{q^2+4q}}{2}, \quad \theta = -\frac{q+2}{2} + \frac{\sqrt{q^2+4q}}{2}$$

$$\begin{aligned}
\Rightarrow \theta + 1 + q &= p \\
&= \frac{p}{p+1} + q \\
\Rightarrow \theta + 1 &= \frac{p}{p+1}
\end{aligned}$$

Thus

$$\begin{aligned}
\kappa &= \frac{p}{p+1} \\
&= \theta + 1 \\
&= -\frac{q}{2} + \frac{\sqrt{q^2+4q}}{2}
\end{aligned}$$

2.3 Innovations Form

In a signal plus noise model, the update in the Kalman Filter depends on the prediction errors or *innovations*,

$$\nu_t = y_t - \mu_{t|t-1}.$$

The innovations form of the AR(1) + noise model is:

$$\begin{aligned} y_t &= \mu_{t|t-1} + \nu_t \\ \mu_{t+1|t} &= \phi\mu_{t|t-1} + k_t\nu_t \end{aligned}$$

This is different from the structural model since we are examining conditionals ($t|t-1$ not t subscript). Thus the dynamic equations (above) are a function of past observations. The Kalman gain k_t depends on ϕ and q. In the steady state, $k_t = \kappa$ is constant:

$$\begin{aligned} y_t &= \mu_{t|t-1} + \nu_t \\ \mu_{t+1|t} &= \phi\mu_{t|t-1} + \kappa\nu_t \\ \Rightarrow y_t &= \phi\mu_{t-1|t-2} + \kappa\nu_{t-1} + \nu_t \\ &= \phi(y_{t-1} - \nu_{t-1}) + \kappa\nu_{t-1} + \nu_t \\ &= \phi y_{t-1} + (\kappa - \phi)\nu_{t-1} + \nu_t \end{aligned}$$

I.e. we get an ARMA(1,1) with MA parameter $\theta = \kappa - \phi$.

2.4 ML Estimation

The joint density for the full set of T observations, $y_1, \ldots, y_T$ is:

$$p(\mathbf{y}; \mathbf{\Psi}) = \prod_{t=1}^{T} p(y_t|Y_{t-1}; \Psi)$$

where Ψ is the vector of parameters. In a Gaussian model:

$$y_t|Y_{t-1} \sim N(\tilde{y}_{t|t-1}, f_t)$$

where $\tilde{y}_{t|t-1} = \phi\mu_{t-1|t-1}$ and $f_t = p_{t|t-1} + \sigma_\varepsilon^2$. Thus the log-likelihood can be derived:

$$\begin{aligned} L(\phi, \sigma_\eta^2, \sigma_\varepsilon^2) &= \prod_{t=1}^{T} \frac{1}{\sqrt{2\pi f_t}} \exp\left(-\frac{(y_t - \tilde{y}_{t|t-1})^2}{2f_t}\right) \\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^T \left(\prod_{t=1}^{T} \frac{1}{\sqrt{f_t}}\right) \exp\left(-\sum_{t=1}^{T} \frac{(y_t - \tilde{y}_{t|t-1})^2}{2f_t}\right) \\ \ln L(\phi, \sigma_\eta^2, \sigma_\varepsilon^2) &= -\frac{T}{2}\ln 2\pi - \frac{1}{2}\sum_{t=1}^{T} \ln f_t - \frac{1}{2}\sum_{t=1}^{T} \frac{(y_t - \tilde{y}_{t|t-1})^2}{f_t} \end{aligned}$$

The prediction error $y_t - \tilde{y}_{t|t-1}$ is now a residual, whose interpretation is the same as $\varepsilon(\Psi)$ in an ARMA from lecture 1. When f_t is constant, as it is in the steady-state, maximising lnL is equivalent to minimising a sum of squares function, again as in the ARMA.

Diagnostics and testing with the residuals

It follows from $y_t|Y_{t-1} \sim N(\tilde{y}_{t|t-1}, f_t)$ that the prediction errors ν_t are normally distributed and serially independent with mean zero and variance f_t. When a model has been estimated, diagnostic checking can be based on the standardised residuals, $\nu_t/\sqrt{f_t}$.

However, unless q and ϕ are known, they will not be NID. There are two options here, carry out

a modified Box-Pierce test based on χ^2_{P-2} or do an LM test.

Tests on variance parameters

The null hypothesis that $\sigma^2_\varepsilon = 0$ can be tested against $\sigma^2_\varepsilon > 0$ using a likelihood ratio test, the distribution is a mixture of of χ^2's:

$$LR \sim \frac{1}{2}\chi^2_0 + \frac{1}{2}\chi^2_1$$

The χ^2_0 is a degenerate distribution with all it's mass at the origin, thus the size of the test can be set to α by using the 2α quantile of the χ^2_1 distribution. When $\phi = 1$ (random walk) this test is still valid since the model is stationary and invertible in first differences.

The null hypothesis that $\sigma^2_\eta = 0$ vs $\sigma^2_\eta > 0$ does not produce a LR statistic with the previous distribution, because when the null is true, the first differenced observations $\Delta y_t = \Delta\varepsilon_t$ are strictly non-invertible. Such non-standard behaviour is a feature of the situations when the model is stationary under the null, but non-stationary under the alternative.

Note:-

A test of the hypothesis that $\phi = 1$ against the alternative $\phi < 1$ encounters similar problems. Indeed the LR test of $\sigma^2_\eta = 0$ cannot be implemented in the stationary AR(1) + noise since ϕ is not identified when $\sigma^2_\eta = 0$.

2.5 Signal Extraction

Signal extraction is concerned with finding the optimal estimate or estimator of an unobserved component at some point within the sample. Whilst filtering aims to find the expected value of the signal conditional on information at time t, smoothing seeks to account for the information made available after t.

In the AR(1) + noise model, the smoothed estimate of the component of interest μ_t, is its expectation conditional on the full set of observations. When the model is Gaussian, the estimate is liner so:

$$\tilde{\mu}_{t|T} = \mathbb{E}(\mu_t|Y_T) = \sum_{j=t-T}^{t-1} w_{j,t} y_{t-j}$$

where the w's are non-random weights.

2.5.1 Wiener-Kolmogorov (WK) filter

Given an UC model, the WK formulae give expressions for the weights w_j. The assumption of a doubly infinite series is made, meaning there are an infinite number of observations on either side of the point of interest. The weights are constant, and so:

$$\mu_{t|\infty} = \sum_{j=-\infty}^{\infty} w_j y_{t-j} = w(L) y_t$$

where $w(L)$ is a symmetric polynomial in the lag operator. The WK formula for finding the weights in $y_t = \mu_t + \xi_t$ where μ_t and ξ_t are mutually uncorrelated ARMA models is:

$$w(L) = \frac{g_\mu(L)}{g_y(L)} = \frac{g_\mu(L)}{g_\mu(L) + g_\xi(L)}$$

where $g_\mu(L)$ and $g_\xi(L)$ are the ACGF's of μ_t and ξ_t respectively. The ACGF of y is $g_\mu(L) + g_x i(L)$.

Definition 2.5.1: Autocovariance generating function (ACGF)

The ACGF is defined as a polynomial in the lag operator, g(L), such that:

$$g(L) = \sum_{\tau=-\infty}^{\infty} \gamma(\tau) L^{\tau}$$

The coefficient of L^{τ} gives the autocovariance at lag τ.

Recall the ACGF of an ARMA model $y_t = \phi^{-1}(L)\theta(L)\nu_t$ where $\nu_t \sim WN(0, \sigma_\nu^2)$ is:

$$g(L) = \frac{\theta(L)\theta(L^{-1})}{\phi(L)\phi(L^{-1})}\sigma_\nu^2$$

Example. MA(1): $y_t = \nu_t + \theta\nu_{t-1}$

$$\begin{aligned} g(L) &= \sum_{\tau=-\infty}^{\infty} \gamma(\tau) L^{\tau} \\ &= \gamma(0) + \gamma(1)L + \gamma(-1)L^{-1} \\ &= (1+\theta^2)\sigma_\nu^2 + \theta\sigma_\nu^2 L + \theta\sigma_\nu^2 L^{-1} \end{aligned}$$

In the AR(1) + noise model, the reduced form is ARMA(1,1) with MA parameter

$$\theta = -\frac{q+1+\phi}{2\phi} + \frac{\sqrt{(q+1+\phi^2)^2 - 4\phi^2}}{2\phi}$$

Provided q¿0, the WK formula gives:

$$\mu_{t|\infty} = \frac{\frac{\sigma_\eta^2}{(1-\phi L)^2}}{\sigma_\varepsilon^2 + \frac{\sigma_\eta^2}{(1-\phi L)^2}} y_t = \frac{\sigma_\eta^2}{\sigma_\nu^2(1+\theta L)^2} y_t$$

The second equality holds since the reduced form is equated to an ARMA(1,1): $\eta_t + \varepsilon_t - \phi\varepsilon_{t-1} = \nu_t + \theta\nu_{t-1}$. Further, from equating the first autocovariances of the reduced form and the ARMA(1,1) as at the start of the lecture, we get:

$$\sigma_\nu^2 = -\frac{\phi\sigma_\varepsilon^2}{\theta} \Rightarrow \mu_{t|\infty} = -\frac{q\theta}{\phi}\frac{1}{(1+\theta L)^2} y_t$$

On recognising that the second term is the ACGF of an AR(1) model with parameter $-\theta$, it becomes apparent that the weights decline symmetrically and exponentially.

$$w_j = -\frac{q\theta/\phi}{1-\theta^2}(-\theta)^{|j|}$$

By setting L = 1[1] it can be seen that the weights sum to:

$$\begin{aligned}\sum_{j=-\infty}^{\infty} w_j &= -\frac{q\theta/\phi}{1-\theta^2}\sum_{j=-\infty}^{\infty}(-\theta)^{|j|} \\ &= -\frac{q\theta/\phi}{1-\theta^2}\left(2\sum_{j=0}^{\infty}\theta^j - 1\right) \\ &= -\frac{q\theta/\phi}{1-\theta^2}\left(\frac{2}{1-\theta} - 1\right) \\ &= -\frac{q\theta/\phi}{1-\theta^2}\frac{1-\theta}{1+\theta} \\ &= -\frac{q\theta/\phi}{(1+\theta)^2}\end{aligned}$$

Recall from 50 fucking slides ago:

$$q = -\frac{(\phi+\theta)(1+\phi\theta)}{\theta}$$

Thus

$$\begin{aligned}-\frac{q\theta/\phi}{(1+\theta)^2} &= \frac{\frac{(\phi+\theta)(1+\phi\theta)}{\theta}\theta}{\phi(1+\theta)^2} \\ &= \frac{(\phi+\theta)(1+\phi\theta)}{\phi(1+\theta)^2}\end{aligned}$$

In the case of a random walk ($\phi = 1$) this simplifies to unity. The lower q, the closer θ is to one, and the more spread out the weights. The MSE is given by

$$MSE(\mu_{t|\infty}) = \sigma_\varepsilon^2\frac{1+\theta}{1-\theta}$$

The weights for smoothing near the end of a semi-infinite sample can be found by modifying the signal extraction formulae:

$$w_j = \frac{\phi+\theta}{\phi(1-\theta^2)}[(1+\phi\theta)(-\theta)^{|-j|} + (\phi+\theta)(\theta)^{j+2(T-t)+1}], \quad j = t-T,\ldots,0,\ldots,\infty$$

The smoothing weights are obtained when $t << T$. On the other hand setting t=T gives the weights for the filtered estimator of μ_T as

$$\begin{aligned}w_j &= \frac{(\phi+\theta)[(1+\phi\theta)(-\theta)^j + (\phi+\theta)(\theta)^{j+1}]}{\phi(1-\theta^2)} \\ &= \frac{(\phi+\theta)(-\theta)^j}{\phi(1-\theta^2)}[(1+\phi\theta) - (\phi+\theta)\theta] \\ &= \frac{\phi+\theta}{\phi}\frac{1-\theta^2}{1-\theta^2}(-\theta)^j \\ &= \frac{\phi+\theta}{\phi}(-\theta)^j, \quad j = 0,1,2,\ldots\end{aligned}$$

The weights decline exponentially when $\phi = 1$:

$$w_j = (1+\theta)(-\theta)^j, \quad j = 0,1,2,\ldots$$

as in the classic EWMA.

[1] Literally how does the lag operator even work, why can you set it to 1??? It's an OPERATOR wtf is going on

2.5.2 Auxiliary Residuals

Not covered in lectures.

2.6 State Space form

We derived the filter for an AR(1) dynamic equation in the previous section, but what if μ_t is AR(2)? It turns out this is not an issue, since an AR(2) can be expressed in first-order Markov form (a multivariate AR(1)) by writing

$$\begin{bmatrix} \mu_t \\ \mu_{t-1} \end{bmatrix} = \begin{bmatrix} \phi_1 & \phi_2 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \mu_{t-1} \\ \mu_{t-2} \end{bmatrix} + \begin{bmatrix} \eta_t \\ 0 \end{bmatrix}$$

This is known as the transition equation. We can extract the first element as

$$y_t = \mu_t + \varepsilon_t = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} \mu_t \\ \mu_{t-1} \end{bmatrix} + \varepsilon_t$$

known as the measurement equation. The role of the second element is simply to store the information in the second lag. We can generalise this further and write the state space form. The observation is related to an $m \times 1$ vector $alpha_t$, the state vector, through a measurement equation

$$y_t = z_t \alpha_t + d_t + \varepsilon_t$$

where z_t is a $1 \times m$ vector, and d_t is deterministic. The elements of α_t are generated by a first order Markov process, the transition equation:

$$\alpha_t = T_t \alpha_{t-1} + \eta_t$$

where T_t is an $m \times m$ matrix and η_t is a $m \times 1$ vector of serially uncorrelated Gaussian disturbances with $\mathbb{E}\eta_t = 0$ and $Var(\eta_t) = Q_t$ (i.e. multivariate WN). The disturbances η_t and ε_t may be contemporaneously correlated, but $\mathbb{E}\varepsilon_t'\eta_s = 0$ for all $t \neq s$.
The specification of the state space system is completed by assuming that $\mathbb{E}\alpha_0 = \alpha_{0|0}$ and $Var(\alpha_0) = P_{0|0}$ are known and $P_{0|0}$ is positive definite. Further $\mathbb{E}\eta_t\alpha_0' = 0$ and $\mathbb{E}\varepsilon_t\alpha_0' = 0$ for all t. All system matrices and vectors (z_t, d_t, T_t, Q_t) are non-stochastic.
Models can also be set up with what is called the future SSF, where the transition equation is

$$\alpha_{t+1} = T_t \alpha_t + \eta_t$$

2.6.1 ARMA models

Any ARMA(p,q) model can be written in state space form. A transition equation can be set up as follows:

$$\alpha_t = \begin{bmatrix} \phi_1 & 1 & 0 & \cdots & 0 \\ \phi_2 & 0 & 1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ \vdots & 0 & \cdots & 0 & 1 \\ \phi_m & 0 & 0 & \dots & 1 \end{bmatrix} \alpha_{t-1} + \begin{bmatrix} 1 \\ \theta_1 \\ \vdots \\ \theta_{m-1} \end{bmatrix} \varepsilon_t$$

where m = max(p,q+1). The original ARMA can be recovered by noting that the first element of α_t is equal to y_t. The role of the measurement equation is simply to extract the first element of the state vector α_t. Thus

$$y_t = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix} \alpha_t$$

There is no disturbance term.

Example. The MA(1) model, $y_t = \mu_t + \xi_t + \theta\xi_{t-1}$ can be written in state space form as:

$$y_t = [1 \quad 0]\alpha_t + d_t$$

where $d_t = \mu$ and

$$\alpha_t = \begin{bmatrix} y_t \\ \theta\xi_t \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \alpha_{t-1} + \begin{bmatrix} 1 \\ \theta \end{bmatrix} \xi_t$$

The future SSF is:

$$y_t = \mu + \theta\alpha_t + \xi_t, \quad \alpha_{t+1} = \xi_t$$

2.6.2 Two AR1 components

$$y_t = \mu_{1,t} + \mu_{2,t} + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma_\varepsilon^2)$$
$$\mu_i, t = \phi_i \mu_{i,t-1} + \eta_{i,t}, \quad \eta_{i,t} \sim N(0, \sigma_i^2), \quad i = 1, 2$$

where $|\phi_i| < 1$, $i = 1, 2$. Using stationarity in the AR(1)'s:

$$\begin{aligned} Var(\mu_{i,t}) &= \phi_i^2 Var(\mu_{i,t}) + \sigma_i^2 \\ \Rightarrow Var(\mu_{i,t}) &= (1 - \phi_i^2)^{-1}\sigma_i^2, \quad i = 1, 2 \\ \Rightarrow Var(y_t) &= Var(\mu_{1,t}) + Var(\mu_{2,t}) + \sigma_1^2 \\ &= (1 - \phi_1^2)^{-1}\sigma_1^2 + (1 - \phi_2^2)^{-1}\sigma_2^2 + \sigma_\varepsilon^2 \end{aligned}$$

Iterating backwards from t:

$$\begin{aligned} y_t &= \mu_{1,t} + \mu_{2,t} + \varepsilon_t \\ &= \phi_1\mu_{1,t-1} + \phi_2\mu_{2,t-1} + \eta_{1,t} + \eta_{2,t} + \varepsilon_t \\ &= \dots \\ &= \phi_1^\tau \mu_{1,t-\tau} + \phi_2^\tau \mu_{2,t-\tau} + \sum_{i=1}^{\tau} \phi_1^{i-1}\eta_{1,t-i+1} + \sum_{i=1}^{\tau} \phi_2^{i-1}\eta_{2,t-i+1} + \varepsilon_t \\ \Rightarrow \gamma(\tau) &= Cov(\phi_1^\tau \mu_{1,t-\tau} + \phi_2^\tau \mu_{2,t-\tau} + \sum_{i=1}^{\tau} \phi_1^{i-1}\eta_{1,t-i+1} + \sum_{i=1}^{\tau} \phi_2^{i-1}\eta_{2,t-i+1}, \mu_{1,t-\tau} + \mu_{2,t-\tau} + \varepsilon_{t-\tau}) \\ &= \phi_1^\tau Var(\mu_{1,t-\tau}) + \phi_2^\tau Var(\mu_{2,t-\tau}) \\ &= \phi_1^\tau(1 - \phi_1^2)^{-1}\sigma_1^2 + \phi_2^\tau(1 - \phi_2^2)^{-1}\sigma_2^2 \\ \Rightarrow \rho(\tau) &= \frac{\phi_1^\tau(1 - \phi_1^2)^{-1}\sigma_1^2 + \phi_2^\tau(1 - \phi_2^2)^{-1}\sigma_2^2}{(1 - \phi_1^2)^{-1}\sigma_1^2 + (1 - \phi_2^2)^{-1}\sigma_2^2 + \sigma_\varepsilon^2} \quad \tau = 1, 2, \dots \end{aligned}$$

The SSF is:

$$y_t = \mu_{1,t} + \mu_{2,t} + \varepsilon_t = z\alpha_t + \varepsilon_t$$

where $z = [1 \quad 1]$ and

$$\alpha_t = \begin{bmatrix} \mu_{1,t} \\ \mu_{2,t} \end{bmatrix} = \begin{bmatrix} \phi_1 & 0 \\ 0 & \phi_2 \end{bmatrix} \alpha_{t-1} + \begin{bmatrix} \eta_{1,t} \\ \eta_{2,t} \end{bmatrix}, \quad Q = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

The initial conditions are $\alpha_{1|0} = 0$ and

$$P_{1|0} = \begin{bmatrix} \frac{\sigma_1^2}{1-\phi_1^2} & 0 \\ 0 & \frac{\sigma_2^2}{1-\phi_2^2} \end{bmatrix}$$

The KF and associated smoother deliver estimates of $\mu_{1,t}$ and $\mu_{2,t}$ and their respective conditional variances. Using the WK filter for this task is much more difficult.

Long memory and unobserved components

The characteristics of long memory can often be approximated by the sum of two stationary AR(1)'s.

2.7 Missing observations and Data Irregularities

Missing observations pose no problems when handled in the state space framework. If an observation is missing at a particular point in time, $t = \tau$, the Kalman Filter updating equations are redundant. Thus for the AR(1) + noise model, the predicted state at time τ is $\mu_{\tau|\tau-1}$ with MSE $p_{\tau|\tau-1}$. If y_τ is missing, then the update is skipped, or we set $\mu_{\tau|\tau} = \mu_{\tau|\tau-1}$ and $p_{\tau+1|\tau-1} = \phi^2 p_{\tau+1|\tau-1}$, and the prediction equations are applied to give the two-step ahead predictor $\mu_{\tau+1|\tau} = \phi\mu_{\tau|\tau-1}$ and $p_{\tau+1|\tau} = \phi^2 p_{\tau|\tau-1} + \sigma_\eta^2$. If $y_{\tau+1}$ is missing, the prediction equations are just applied again.

The likelihood function is computed in the usual way, by using the innovations from the Kalman filter, except there is no innovation corresponding to the missing observation. This generalises for any linear Gaussian state space model irrespective of the number of missing observations or where they're located. E.G. if m consecutive observations are missing, the mean and variance of an $m+1$ step ahead predictive distribution will enter the likelihood function.

Data Irregularities - Suppose a survey becomes quarterly instead of annually, we can just construct a quarterly model with missing observations for the first part of the series.

4 Trends, Cycles and Seasonality

4.1 Deterministic trends

A deterministic trend model takes the form

$$y_t = \mu(t) + \xi_t$$

where ξ_t is a stationary process (EG some ARMA) and $\mu(t)$ is a deterministic function of time. This function can be modelled as a polynomial in time, where $\mu(t) = \beta_0 + \beta_1 t + \beta_2 t^2 + \ldots + \beta_h t^h$. The linear trend model $\mu(t) = \beta_0 + \beta_1 t$ is obtained by setting $h = 1$.
The drawback with deterministic time trend models is that a time series rarely keeps on the same path indefinitely. Structural time series models achieve this aim directly by allowing the trend parameters to evolve over time as stochastic processes.

4.2 ARIMA models

ARMA(p, q) models may be fitted to nonstationary series by differencing the observations. When forecasting is carried out, the differencing must be reversed; Box and Jenkins (1976) called this operation integration If differencing d times gives a stationary and invertible model, then the original series is said to be *integrated of order d*, or $y_t \sim I(d)$.

Definition 4.2.1

An auto-regressive integrated moving average (ARIMA) model of order (p, d, q) is defined as:

$$\phi(L)\Delta^d y_t = \theta(L)\varepsilon_t \tag{4.1}$$

and denoted as $ARIMA(p, d, q)$.

Explanation. ***When a series has been overdifferenced, it will be stationary, but strictly non-invertible due to the presence of a unit root in the MA polynomial***
Suppose we have a random walk model:

$$y_t = y_{t-1} + \varepsilon_t$$

where $\varepsilon_t \sim WN(0, \sigma^2)$. Clearly the first difference is stationary:

$$\Delta y_t = \varepsilon_t$$

however suppose we over-difference:

$$\Delta^2 y_t = \Delta y_t - \Delta y_{t-1} = \varepsilon_t - \varepsilon_{t-1}.$$

We have introduced a unit root in the MA polynomial: $\Theta(L) = 1 - L$, and the series is no longer invertible. □

4.2.1 Prediction

Prediction from stationary models tend towards the mean as the lead time, ℓ, increases. When ℓ is large, the dynamic structure of the model is irrelevant. This is no longer the case with integrated

models, where the forecast function contains a deterministic component. The term $\phi(L)\Delta^d y_t$ in (4.1) is expanded to give an AR polynomial of order $p+d$:

$$\phi(L)\Delta^d = \varphi(L) = 1 - \varphi_1 L - \varphi_2 L^2 - \ldots - \varphi_{p+d} L^{p+d}$$

and predictions are made from the recursive equation

$$\hat{y}_{T+\ell|T} = \varphi_1 \hat{y}_{T+\ell-1|T} + \ldots + \varphi_{p+d}\hat{y}_{T+\ell-p-d|T} + \hat{\varepsilon}_{T+\ell|T} + \ldots + \theta_q \hat{\varepsilon}_{T+\ell-q|T} \quad \ell = 1, 2, \ldots$$

The eventual forecast function (EG for stationary processes forecasts converge to the mean) is a polynomial of order $d-1$. When the model includes a constant it is a polynomial of order d.

Example (Prediction in ARIMA(0,1,1)).

$$\Delta y_t = \varepsilon_t + \theta \varepsilon_{t-1}$$

$$\begin{cases} \hat{y}_{T+1|T} = y_T + \theta \varepsilon_T \\ \hat{y}_{T+\ell|T} = \hat{y}_{T+\ell-1|T} \end{cases}$$

Thus for all lead times, the forecasts made at time T follow a horizontal line. The disturbance term can be obtained from the recursive equation:

$$\varepsilon_t = y_t - y_{t-1} - \theta \varepsilon_{t-1}$$

with $\varepsilon_1 = 0$. Substitute repeatedly for lagged values of ε_t:

$$\begin{aligned}
\varepsilon_2 &= y_2 - y_1 - \theta\varepsilon_1 = y_2 - y_1 \\
\varepsilon_3 &= y_3 - y_2 - \theta(y_2 - y_1) = y_3 - (1+\theta)y_2 + \theta y_1 \\
\varepsilon_4 &= y_4 - y_3 - \theta(y_3 - (1+\theta)y_2 + \theta y_1) = y_4 - (1+\theta)y_3 + \theta(1+\theta)y_2 - \theta^2 y_1 \\
&\vdots \\
\varepsilon_T &= y_T - (1+\theta)\sum_{j=1}^{T-1}(-\theta)^{j-1} y_{T-j}
\end{aligned}$$

So the expression for the one-step ahead predictor can be written as:

$$\begin{aligned}
\hat{y}_{T+1|T} &= y_T + \theta\varepsilon_T \\
&= y_T + \theta\left(y_T - (1+\theta)\sum_{j=1}^{T-1}(-\theta)^{t-1} y_{T-t}\right) \\
&= (1+\theta)\sum_{j=0}^{T-1}(-\theta)^j y_{T-j}
\end{aligned}$$

This predictor is an EWMA, with weights that decrease geometrically with time.

Example (Prediction in ARIMA(0,1,1) + constant). When the model above includes a constant:

$$\Delta y_t = \theta_0 + \varepsilon_t + \theta\varepsilon_{t-1}$$

the forecasts are

$$\hat{y}_{T+1|T} = y_T + \theta_0 + \theta\varepsilon_T$$

and

$$\hat{y}_{T+\ell|T} = y_T + \ell\theta_0 + \theta\varepsilon_T$$

The eventual forecast function is therefore a linear time trend.

Example (ARIMA(1,1,0)).

$$\varphi(L) = (1-\phi L)(1-L) = 1-(1+\phi)L+\phi L^2$$

and forecasts are constructed from the difference equation:

$$\hat{y}_{T+\ell|T} = (1+\phi)\hat{y}_{T+\ell-1|T} - \phi\hat{y}_{T+\ell-2|T}$$

The corresponding forecast function:

$$\hat{y}_{T+\ell|T} = y_T + (y_T - y_{T-1})\frac{\phi(1-\phi^\ell)}{1-\phi} \quad \ell = 1, 2, \ldots$$

As $\ell \to \infty$, $\hat{y}_{T+\ell|T} \to y_T + (y_T - y_{T-1})\frac{\phi}{1-\phi}$, a horizontal line. In contrast to the $ARIMA(0,1,1)$ model where the eventual forecast function is also horizontal, but depends on all past values, the eventual forecast function here depends only on the last two observations.

Note:-

Model Selection: First plot the series in levels and differences, judge which show trending movements. Checking the correlogram is also useful, for stationary processes the autocorrelations tend towards zero as the lag increases, whilst non-stationary processes do not damp so quickly. Keep differencing the model until a correlogram having the characteristics of a stationary model is obtained.

4.3 Structural Time Series Models

4.3.1 Local level model

Definition 4.3.1: Local level model

The local level model consists of a random walk plus noise:

$$\begin{cases} y_t = \mu_t + \varepsilon_t & \varepsilon_t \sim NID(0, \sigma_\varepsilon^2) \\ \mu_t = \mu_{t-1} + \eta_t & \eta_t \sim NID(0, \sigma_\eta^2) \end{cases}$$

where $\mathbb{E}(\varepsilon_t\eta_t) = 0$.

When σ_η^2 is zero the level is constant and the process is just noise around a mean. The signal noise ratio $q = \sigma_\eta^2/\sigma_\varepsilon^2$ plays a key role in determining how observations should be weighted for prediction and signal extraction. The higher q, the more past observations are discounted in forecasting.
By taking first differences of y_t:

$$\begin{aligned} \Delta y_t &= \mu_t - \mu_{t-1} + \varepsilon_t \\ &= \eta_t + \varepsilon_t \end{aligned}$$

we can see that the reduced form is an ARIMA(0,1,1). Setting the autocovariances of the first

differences equal to those of an MA(1) model $[y_t = \xi_t + \theta_{t-1}]$ yields:

$$\begin{aligned}
\gamma_1 &= -\sigma_\varepsilon^2 := \theta\sigma_\xi^2 \\
&\Rightarrow \sigma_\xi^2 = -\frac{\sigma_\varepsilon^2}{\theta} \\
\gamma_0 &= \sigma_\eta^2 + 2\sigma_\varepsilon^2 := (1+\theta^2)\sigma_\xi^2 \\
&\Rightarrow \sigma_\eta^2 + \sigma_\varepsilon^2 = (1+\theta^2)\left(-\frac{\sigma_\varepsilon^2}{\theta}\right) \\
&\Rightarrow q + 2 = -\frac{1+\theta^2}{\theta} \\
&\Rightarrow \theta = \frac{-q-2 \pm \sqrt{q^2+4q}}{2}
\end{aligned}$$

4.3.2 Local linear trend model

A linear time trend of the form $\mu(t) = \beta_0 + \beta_1 t$ can be constructed recursively from $\mu_t = \mu_{t-1} + \beta_1$ with $\beta_0 = \mu_0$. It can then be made stochastic by adding a disturbance to the level μ_t (random walk + drift), or the slope itself can be made stochastic - resulting in the local linear trend model:

Definition 4.3.2: Local linear trend model

$$\begin{aligned}
y_t &= \mu_t + \varepsilon_t \quad \varepsilon_t \sim NID(0, \sigma_\varepsilon^2) \\
\mu_t &= \mu_{t-1} + \beta_{t-1} + \eta_t \quad \eta_t \sim NID(0, \sigma_\eta^2) \\
\beta_t &= \beta_{t-1} + \zeta_t \quad \zeta_t \sim NID(0, \sigma_\zeta^2)
\end{aligned}$$

where all disturbances are mutually independent.

Only when $\sigma_\zeta^2 = 0$ the slope is fixed and the trend reduces to random walk plus drift. Allowing $\sigma_\zeta^2 > 0$, but setting $\sigma_\eta^2 = 0$ makes μ_t an integrated random walk.
The SSF of the local linear trend model is:

$$y_t = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} \mu_t \\ \beta_t \end{bmatrix} + \varepsilon_t = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \mu_{t-1} \\ \beta_{t-1} \end{bmatrix} + \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} \eta_t \\ \zeta_t \end{bmatrix} + \varepsilon_t$$

where the state vector is $\alpha_t = \begin{bmatrix} \mu_t \\ \beta_t \end{bmatrix}$ and covariance matrix is $Q = \begin{bmatrix} \sigma_\eta^2 & 0 \\ 0 & \sigma_\zeta^2 \end{bmatrix}$.

Claim 4.3.1. The reduced form of the local linear trend model is an ARIMA(0,2,2) model.

Proof. Begin with the trend, and difference until we find a stationary expression for y_t:

$$\begin{aligned}
\beta_t &= \beta_{t-1} + \zeta_t \\
\Rightarrow \Delta\beta_t &= \zeta_t \\
\mu_t &= \mu_{t-1} + \beta_{t-1} + \eta_t \\
\Rightarrow \Delta\mu_t &= \beta_{t-1} + \eta_t \\
\Rightarrow \Delta^2\mu_t &= \Delta\beta_{t-1} + \eta_t + \eta_{t-1} = \zeta_{t-1} + \eta_t + \eta_{t-1} \\
y_t &= \mu_t + \varepsilon_t \\
\Rightarrow \Delta^2 y_t &= \Delta^2\mu_t + (1-L)^2\varepsilon_t = \zeta_{t-1} + \eta_t + \eta_{t-1} + \varepsilon_t - 2\varepsilon_{t-1} + \varepsilon_{t-2}
\end{aligned}$$

The ACF is thus:

$$\begin{aligned}\gamma(0) &= \sigma_\zeta^2 + 2\sigma_\eta^2 + 6\sigma_\varepsilon^2\\ \gamma(1) &= -\sigma_\eta^2 - 4\sigma_\varepsilon^2\\ \gamma(2) &= \sigma_\varepsilon^2\\ \gamma(\tau) &= 0 \quad \tau > 2\end{aligned}$$

Which we could match to the ACF of an ARIMA(0,2,2) model and solve for the parameters. □

Question 1

Show that the reduced form of the modified damped trend model with a mean is ARIMA(1,1,2):

$$\begin{aligned}y_t &= \mu_t + \varepsilon_t\\ \mu_t &= \mu_{t-1} + \beta_{t-1} + \eta_t\\ \beta_t &= (1-\rho)\beta + \rho\beta_{t-1} + \zeta_t\end{aligned}$$

In other words the slope is an AR(1) with mean β.

Solution:-

$$\begin{aligned}\beta_t &= (1-\rho)\beta + \rho\beta_{t-1} + \zeta_t\\ \Rightarrow (1-\rho L)\beta_t &= (1-\rho)\beta + \zeta_t\\ \Rightarrow \beta_t &= \beta + \frac{\zeta_t}{1-\rho L} \quad \text{since } \beta \text{ constant}\\ \mu_t &= \mu_{t-1} + \beta_{t-1} + \eta_t\\ \Rightarrow \Delta\mu_t &= \beta_{t-1} + \eta_t\\ &= \beta + \frac{\zeta_{t-1}}{1-\rho L} + \eta_t\\ y_t &= \mu_t + \varepsilon_t\\ \Rightarrow \Delta y_t &= \Delta\mu_t + \varepsilon_t - \varepsilon_{t-1}\\ &= \beta + \frac{\zeta_{t-1}}{1-\rho L} + \eta_t + \varepsilon_t - \varepsilon_{t-1}\\ \Rightarrow (1-\rho L)\Delta y_t &= (1-\rho)\beta + \zeta_{t-1} + (1-\rho L)(\eta_t + \varepsilon_t - \varepsilon_{t-1})\end{aligned}$$

The left hand side is an AR(1) on the differenced series, and the RHS is a mean + an MA(2) component. Thus it is an ARIMA(1,1,2) model.

4.4 Trend-cycle models

We consider stochastic cycles with frequency λ_c and damping term ρ.

Definition 4.4.1: Stochastic cycle

The stochastic cycle model is given by:

$$\begin{bmatrix} \psi_t \\ \psi_t^* \end{bmatrix} = \rho \begin{bmatrix} \cos\lambda_c & \sin\lambda_c \\ -\sin\lambda_c & \cos\lambda_c \end{bmatrix} \begin{bmatrix} \psi_{t-1} \\ \psi_{t-1}^* \end{bmatrix} + \begin{bmatrix} \kappa_t \\ \kappa_t^* \end{bmatrix} \tag{4.2}$$

where κ_t is a white noise process with the same variance as κ_t^* and $0 \leq \rho \leq 1$.

Observe that the model collapses to an AR(1) model when the cycle has frequency zero or π (giving positive and negative AR coefficients respectively[1]).

Definition 4.4.2: Trend-cycle models

The trend plus cycle plus noise model is:

$$\begin{aligned} y_t &= \mu_t + \psi_t + \varepsilon_t \\ \mu_t &= \mu_{t-1} + \beta_{t-1} + \eta_t \\ \beta_t &= \beta_{t-1} + \zeta_t \\ \begin{bmatrix} \psi_t \\ \psi_t^* \end{bmatrix} &= \rho \begin{bmatrix} \cos\lambda_c & \sin\lambda_c \\ -\sin\lambda_c & \cos\lambda_c \end{bmatrix} \begin{bmatrix} \psi_{t-1} \\ \psi_{t-1}^* \end{bmatrix} + \begin{bmatrix} \kappa_t \\ \kappa_t^* \end{bmatrix} \end{aligned}$$

Claim 4.4.1. The reduced form of the trend-cycle model is an ARIMA(2,2,4) model.

Proof. Recall that the reduced form of the local linear trend model is an ARIMA(0,2,2), thus the second difference of y_t has the form:

$$\Delta^2 y_t = \zeta_{t-1} + \eta_t + \eta_{t-1} + \varepsilon_t - 2\varepsilon_{t-1} + \varepsilon_{t-2} + \Delta^2\psi_t$$

The stochastic seasonal is defined via the recursions in (4.2):

$$\begin{aligned} \psi_t &= \rho\cos\lambda_c\psi_{t-1} + \rho\sin\lambda_c\psi_{t-1}^* + \kappa_t \\ \Rightarrow (1-\rho\cos\lambda_c L)\psi_t &= \rho\sin\lambda_c L\psi_t^* + \kappa_t \\ \psi_t^* &= -\rho\sin\lambda_c\psi_{t-1} + \rho\cos\lambda_c\psi_{t-1}^* + \kappa_t^* \\ \Rightarrow (1-\rho\cos\lambda_c L)\psi_t^* &= -\rho\sin\lambda_c L\psi_t + \kappa_t^* \end{aligned}$$

We can thus express the evolution of the cycle ψ_t as:

$$\begin{aligned} (1-\rho\cos\lambda_c L)\psi_t &= \rho\sin\lambda_c L\left(\frac{-\rho\sin\lambda_c L\psi_t + \kappa_t^*}{1-\rho\cos\lambda_c L}\right) + \kappa_t \\ (1-\rho\cos\lambda_c L)^2\psi_t &= \rho\sin\lambda_c L\left(-\rho\sin\lambda_c L\psi_t + \kappa_t^*\right) + (1-\rho\cos\lambda_c L)\kappa_t \\ (1-2\rho\cos\lambda_c L + (\rho L)^2(\cos^2\lambda_c + \sin^2\lambda_c))\psi_t &= \kappa_t + \rho L(\sin\lambda_c\kappa_t^* - \cos\lambda_c\kappa_t^*) \\ \psi_t &= \frac{\kappa_t + \rho L(\sin\lambda_c\kappa_t^* - \cos\lambda_c\kappa_t^*)}{1-2\rho\cos\lambda_c L + \rho^2 L^2} \end{aligned}$$

This has the structure of an ARMA(2,1) model. We can take second differences and substitute

[1] See lecture on Spectral Analysis

into the expression for $\Delta^2 y_t$:

$$\begin{aligned}\Delta^2 y_t &= \zeta_{t-1} + \eta_t + \eta_{t-1} + \varepsilon_t - 2\varepsilon_{t-1} + \varepsilon_{t-2} + \Delta^2\psi_t \\ &= \zeta_{t-1} + \eta_t + \eta_{t-1} + \varepsilon_t - 2\varepsilon_{t-1} + \varepsilon_{t-2} + \frac{(1-L)^2(\kappa_{t-1} + \rho L(\sin\lambda_c\kappa^*_{t-1} - \cos\lambda_c\kappa^*_{t-1}))}{1 - 2\rho\cos\lambda_c L + \rho^2 L^2}\end{aligned}$$

We can rearrange to show:

$$\begin{aligned}(1 - 2\rho\cos\lambda_c L + \rho^2 L^2)\Delta^2 y_t = (1 - 2\rho\cos\lambda_c L + \rho^2 L^2)(\zeta_{t-1} + \eta_t + \eta_{t-1} + \varepsilon_t - 2\varepsilon_{t-1} + \varepsilon_{t-2}) \\ + (1-L)^2(\kappa_{t-1} + \rho L(\sin\lambda_c\kappa^*_{t-1} - \cos\lambda_c\kappa_{t-1})^*)\end{aligned}$$

The left hand side is an AR(2) on the second difference of the series, and the right hand side is an MA(4) (the deepest lag is $\rho^2\varepsilon_{t-4}$). Thus the reduced form of the trend-cycle model is an ARIMA(2,2,4) model. □

The state space form of the model is:

$$y_t = \begin{bmatrix}1 & 0 & 1 & 0\end{bmatrix}\alpha_t + \varepsilon_t$$

$$\alpha_t = \begin{bmatrix}\mu_t \\ \beta_t \\ \psi_t \\ \psi^*_t\end{bmatrix} = \begin{bmatrix}1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \rho\cos\lambda_c & \rho\sin\lambda_c \\ 0 & 0 & -\rho\sin\lambda_c & \rho\cos\lambda_c\end{bmatrix}\begin{bmatrix}\mu_{t-1} \\ \beta_{t-1} \\ \psi_{t-1} \\ \psi^*_{t-1}\end{bmatrix} + \begin{bmatrix}\eta_t \\ \zeta_t \\ \kappa_t \\ \kappa^*_t\end{bmatrix}$$

Note that this matrix is block diagonal, this is because the trend and cycle are uncorrelated.

Definition 4.4.3: Cyclical trend model

The cyclical trend model incorporates the cycle into the slope by moving it into the level equation:

$$\begin{aligned}y_t &= \mu_t + \varepsilon_t \\ \mu_t &= \mu_{t-1} + \beta_{t-1} + \eta_t \\ \beta_t &= \rho\beta_{t-1} + \zeta_t\end{aligned}$$

We can write this in state space form:

$$y_t = \begin{bmatrix}1 & 0 & 0 & 0\end{bmatrix}\alpha_t + \varepsilon_t$$

$$\alpha_t = \begin{bmatrix}\mu_t \\ \beta_t \\ \psi_t \\ \psi^*_t\end{bmatrix} = \begin{bmatrix}1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \rho\cos\lambda_c & \rho\sin\lambda_c \\ 0 & 0 & -\rho\sin\lambda_c & \rho\cos\lambda_c\end{bmatrix}\begin{bmatrix}\mu_{t-1} \\ \beta_{t-1} \\ \psi_{t-1} \\ \psi^*_{t-1}\end{bmatrix} + \begin{bmatrix}\eta_t \\ \zeta_t \\ \kappa_t \\ \kappa^*_t\end{bmatrix}$$

and see that the trend and cycle are now correlated!

Claim 4.4.2. The cyclical trend model, where the cycle appears in the slope by moving it into the level equation:

$$\mu_t = \mu_{t-1} + \beta_{t-1} + \psi_{t-1} + \eta_t$$

has reduced form ARIMA(2,2,4).

Proof. Fuck off □

Question 2

Show that the damped trend model:

$$\begin{aligned} y_t &= \mu_t + \varepsilon_t \\ \mu_t &= \mu_{t-1} + \beta_{t-1} + \eta_t \\ \beta_t &= \rho\beta_{t-1} + \zeta_t \end{aligned}$$

Is a special case of the cyclical trend model.

Solution:-

Let us suppose our cycles have frequency zero, then $\lambda_c = 0$. Thus $\sin \lambda_c = 0$ and the second part of the recursion is redundant. Using $\cos 0 = 1$:

$$\psi_t = \rho\psi_{t-1} + \kappa_t$$

Consider the special case of the cyclical trend model where β_t does not appear in the level equation:

$$\begin{aligned} y_t &= \mu_t + \varepsilon_t \\ \mu_t &= \mu_{t-1} + \psi_{t-1} + \eta_t \end{aligned}$$

However this is exactly the same as the damped trend model, just with a change of notation! We can see this clearly by writing the structural forms side by side:

Damped Trend Model	**Cyclical Trend Model**
$y_t = \mu_t + \varepsilon_t$	$y_t = \mu_t + \varepsilon_t$
$\mu_t = \mu_{t-1} + \beta_{t-1} + \eta_t$	$\mu_t = \mu_{t-1} + \psi_{t-1} + \eta_t$
$\beta_t = \rho\beta_{t-1} + \zeta_t$	$\psi_t = \rho\psi_{t-1} + \kappa_t$

These are the same, just interchanging the roles of β_t and ψ_t.

4.5 Unit roots and Stationarity tests

Two classes of tests are available to help distinguish between models that are integrated of different orders. In stationarity tests, nonstationarity appears under the alternative hypothesis. In unit root tests, the situation is reversed, with nonstationarity corresponding to the null hypothesis. This duality between the two sets of tests can also be seen in terms of AR and MA models. For example, the local level model is equivalent to an MA(1) in first differences, and the test of a zero variance driving the random walk corresponds to a test that the MA parameter is strictly noninvertible.

4.5.1 Unit root tests: Dickey-Fuller

The basic unit root test is of the null hypothesis that a series is a random walk against the alternative that it's a stationary AR(1). Thus in

$$y_t = \phi y_{t-1} + \varepsilon_t$$

with y_0 fixed, the null hypothesis is $\phi = 1$ and the alternative is $\phi < 1$. The Dickey-Fuller test is a one sided test based on the 't-statistic' obtained from the OLS estimate of ϕ. This is not asymptotically normal, but critical values are tabulated by Dickey and Fuller (1979). There are different critical values when the model includes a constant and/or a time trend:

$$y_t = \alpha + \beta t + \phi y_{t-1} + \varepsilon_t \quad \varepsilon_t \sim NID(0, \sigma^2)$$

Note:-

If the true model is a random walk plus drift and a constant, but no time trend ($y_t = \alpha + y_{t-1} + \varepsilon_t$) then the t statistic is asymptotically normal. See Hamilton (1994, p 495-7) if you're feeling brave.

We can subtract y_{t-1} from both sides to obtain the dickey fuller test:

Definition 4.5.1: Dickey-Fuller test

$$\Delta y_t = \rho y_{t-1} + \varepsilon_t \quad \rho = \phi - 1$$

$H_0 : \rho = 0$ and $H_1 : \rho < 0$ using Dickey-Fuller tables.

More generally we can consider the augmented Dickey-Fuller test, based on an AR(p):

$$y_t = \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} + \varepsilon_t$$

Definition 4.5.2: Augmented Dickey-Fuller test

$$\Delta y_t = \rho y_{t-1} + \sum_{i=1}^{p-1} \gamma_i \Delta y_{t-i} + \varepsilon_t$$

where

$$\gamma_i = -\sum_{j=i+1}^{p} \phi_j \quad \text{and} \quad \rho = \sum_{i=1}^{p} \phi_i - 1$$

$H_0 : \rho = 0$ and $H_1 : \rho < 0$ using Dickey-Fuller tables.

The distribution of $\hat{\rho}$ is unaffected by the lagged differences since it converges at a faster rate. Thus we use the standard DF tables. That $\rho = 0$ implies a unit root follows because when it holds the polynomial associated with the AR(p) factorises as $\phi(L) = (1 - L)\phi^*(L)$. Setting L=1 then gives $\phi(1) = 0$ and the equation for ρ given above.

4.5.2 Stationarity tests

Consider the random walk plus noise model (or local level model):

$$y_t = \mu_t + \varepsilon_t \quad \mu_t = \mu_{t-1} + \eta_t$$

When $\sigma_\eta^2 = 0$ the random walk becomes a constant level. Thus a nonstationary model becomes stationary and standard asymptotic theory no longer applies to a test of $H_0 : \sigma_\eta^2 = 0$ against $H_1 : \sigma_\eta^2 > 0$.

Definition 4.5.3: Nyblom-Makelainen test

$$NM = \frac{1}{T^2 s^2} \sum_{i=1}^{T} \left[\sum_{t=1}^{i} e_t \right]^2 > c$$

where $e_t = y_t - \bar{y}$, $s^2 = T^{-1} \sum_{t=1}^{T} e_t^2$ and c is a critical value.

Asymptotically NM has a Cramer-von Mises distribution under the null when the irregular's are iid. This model can be extended such that ε_t follows any indeterministic stationary process, giving us the KPSS test.

Definition 4.5.4: KPSS test

Replace s^2 in the NM statistic with an estimator of the long run variance of ε_t:

$$\sigma_L^2(m) = \hat{\gamma}(0) + 2\sum_{\tau=1}^{m} w(\tau, m)\hat{\gamma}(\tau)$$

where $w(\tau, m)$ is a weighting function, such as $w(\tau, m) = 1 - |\tau|/(m+1)$, and $\hat{\gamma}(\tau)$ is the sample autocovariance at lag τ. The lag length m is specified by the researcher.

4.6 Seasonality

4.6.1 Deterministic seasonality

A seasonal component, γ_t may be added to a model consisting of a trend and irregular to give the *basic structural model* (BSM):

$$y_t = \mu_t + \gamma_t + \varepsilon_t$$

where the components evolve as before. The signal noise ratio associated with the seasonal component is $q_\omega = \sigma_\omega^2/\sigma_\varepsilon^2$.
a fixed seasonal pattern may be modelled as:

$$\gamma_t = \sum_{j=1}^{s} \gamma_j z_{jt}$$

where s is the number of seasons and the dummy variables z_{jt} take the value 1 in the jth season and 0 otherwise. In order to not confound trend with seasonality, we constrain $\sum_{j=1}^{s} \gamma_j = 0$.
Of course seasonal patterns are not fixed in time, but the requirement that the terms sum to zero must be maintained when the seasonal pattern changes.

4.6.2 Stochastic seasonality

Let γ_{jt} denote the effect of season j at time t and define $\gamma_t = [\gamma_{1t}, \ldots, \gamma_{st}]^T$. The seasonals evolve according to a multivariate random walk:

$$\gamma_t = \gamma_{t-1} + \omega_t \quad \equiv \quad \begin{bmatrix} \gamma_{1t} \\ \vdots \\ \gamma_{st} \end{bmatrix} = \begin{bmatrix} \gamma_{1,t-1} \\ \vdots \\ \gamma_{s,t-1} \end{bmatrix} + \begin{bmatrix} \omega_{1t} \\ \vdots \\ \omega_{st} \end{bmatrix}$$

where $\omega_t = [\omega_{1t}, \ldots, \omega_{st}]^T$ is a zero-mean disturbance with

$$Var(\omega_t) = \sigma_\omega^2 \left(I - \frac{ii'}{s} \right) = \sigma^2 \begin{bmatrix} 1 - \frac{1}{s} & -\frac{1}{s} & \ldots & -\frac{1}{s} \\ -\frac{1}{s} & 1 - \frac{1}{s} & \ldots & -\frac{1}{s} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{s} & -\frac{1}{s} & \ldots & 1 - \frac{1}{s} \end{bmatrix}$$

This structure is necessary to ensure that the seasonal components always sum to zero. The sum of the elements of γ_t can be written as $i'\gamma_t$. From this structure we can see that $Var(i'\omega_t) =$

$\sigma^2_\omega(i' - \frac{i'ii'}{s}) = \sigma^2_\omega(i' - i') = 0$. Combining this with an initial condition setting the sum 0 at t=0, we have that the expectation and variance of $i'\gamma_t$ are always zero.
Note that here although all s seasonal components are continually changing, only one affects the observation at time t, that is $\gamma_t = \gamma_{jt}$ when j is prevailing at t.
We can write this in state-space form:

$$y_t = \begin{bmatrix} 1 & 0 & z_t' \end{bmatrix} \alpha_t + \varepsilon_t$$

$$\alpha_t = \begin{bmatrix} \mu_t \\ \beta_t \\ \gamma_t \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & I_s \end{bmatrix} \begin{bmatrix} \mu_{t-1} \\ \beta_{t-1} \\ \gamma_{t-1} \end{bmatrix} + \begin{bmatrix} \eta_t \\ \zeta_t \\ \omega_t \end{bmatrix}$$

where the bottom row in the transition equation is just rewriting the multivariate random walk.
We can write the reduced form of the dummy variable stochastic seasonal model as:

$$\gamma_t = -\sum_{j=1}^{s-1} \gamma_{t-j} + \omega_t$$

which is simply rearranging the fact that the sum of the seasonals equals a random disturbance.

Claim 4.6.1. The reduced form of the BSM with stochastic seasonality is $\Delta\Delta_s y_t \sim MA(s+1)$

Proof.

$$\begin{aligned} y_t &= \mu_t + \gamma_t + \varepsilon_t \\ \mu_t &= \mu_{t-1} + \beta_{t-1} + \eta_t \\ \beta_t &= \beta_{t-1} + \zeta_t \\ \gamma_t &= -\sum_{j=1}^{s-1} \gamma_{t-j} + \omega_t \end{aligned}$$

We can make the stochastic trend component stationary by second differencing as usual:

$$\Delta^2 \mu_t = \Delta\beta_{t-1} + \eta_t - \eta_{t-1} = \zeta_{t-1} + \eta_t - \eta_{t-1}$$

The seasonal component can be made stationary by applying the seasonal summation operator

Definition 4.6.1: Seasonal summation operator

$$S(L) = 1 + L + L^2 + \ldots + L^{s-1}$$

which is a special case of the summation operator.

Thus $S(L)\gamma_t = \omega_t$ (by rearranging the reduced form, we see the sum of seasonals is ω_t). The stationary form of the BSM is obtained by multiplying through by $\Delta^2 S(L) = \Delta\Delta_s$. This follows from

$$(1 - L)S(L) = 1 - L^s = \Delta_s \Rightarrow \Delta^2 S(L) = \Delta\Delta_s$$

The result is:

$$\Delta\Delta_s y_t = \Delta_s \eta_t + S(L)\zeta_{t-1} + \Delta^2 \omega_t + \Delta\Delta_s \varepsilon_t$$

Which is clearly an Ma(s+1) from the deepest lag on epsilon. □

Note:-

If the slope were deterministic in the above, we would not need to difference out the trend and the observations are rendered stationary by the seasonal difference alone, giving $\Delta_s y_t \sim MA(s+1)$ plus a constant trend β.

4.7 Seasonal ARIMA models

Definition 4.7.1: SARIMA(p,d,q)(P,D,Q)$_s$

The seasonal ARIMA model is given by:

$$\phi(L)\Phi(L^s)\Delta^d\Delta_s^D y_t = \theta_0 + \theta(L)\Theta(L^s)\varepsilon_t$$

where θ_0 is a constant, while $\Phi(L^s)$ and $\Theta(L^s)$ are polynomials in the seasonal lag operator L^s:

$$\Phi(L^s) = 1 - \Phi_1 L^s - \ldots - \Phi_P L^{sP} \quad \Theta(L^s) = 1 - \Theta_1 L^s - \ldots - \Theta_Q L^{sQ}$$

The most common implementation of this very broad class of models is known as the *airline model*, which is a SARIMA$(0,1,1)(0,1,1)_s$ model:

$$\Delta\Delta_s y_t = (1+\theta L)(1+\Theta L^s)\varepsilon_t$$

Claim 4.7.1. With a deterministic seasonal and no trend, the airline model is equivalent to the BSM when $\Theta = -1$.

Proof. The BSM is:

$$\begin{aligned} y_t &= \mu_t + \gamma_t + \varepsilon_t \\ \mu_t &= \mu_{t-1} + \eta_t \\ \gamma_t &= \sum_{j=1}^{s} \gamma_j z_{jt} \end{aligned}$$

We can write the seasonally differenced series as:

$$\begin{aligned} \Delta_s y_t &= \Delta_s \mu_t + \Delta_s \gamma_t + \Delta_s \varepsilon_t \\ &= \mu_t - \mu_{t-s} + \varepsilon_t - \varepsilon_{t-s} \end{aligned}$$

Since $\gamma_t - \gamma_{t-s} = 0$. We now iterate on the level equation:

$$\begin{aligned} \mu_t &= \mu_{t-1} + \eta_t \\ &= \mu_{t-2} + \eta_{t-1} + \eta_t \\ &= \ldots \\ &= \mu_{t-s} + \sum_{j=1}^{s} \eta_{t-j} \end{aligned}$$

Thus we can write the seasonally differenced series as:

$$\begin{aligned} \Delta_s y_t &= \sum_{j=1}^{s} \eta_{t-j} + \varepsilon_t - \varepsilon_{t-s} \\ &= S(L)\eta_t + \Delta_s \varepsilon_t \end{aligned}$$

Now we first difference the series:

$$\begin{aligned}\Delta\Delta_s y_t &= \Delta S(L)\eta_t + \Delta\Delta_s\varepsilon_t \\ &= \Delta_s\eta_t + \Delta\Delta_s\varepsilon_t \quad \text{since } \Delta S(L) = \Delta_s \\ &= (1-L^s)(\eta_t + \varepsilon_t - \varepsilon_{t-1})\end{aligned}$$

This is an MA(s+1) process, and we can compare it to the airline model:

$$\Delta\Delta_s y_t = (1+\Theta L^s)(\varepsilon_t + \theta\varepsilon_{t-1}) \quad \Delta\Delta_s y_t = (1-L^s)(\eta_t + \varepsilon_t - \varepsilon_{t-1})$$

Thus the airline model is equivalent to the BSM when $\Theta = -1$, since both final terms are MA(1) processes.
The airline model thus provides a good approximation to the reduced form when the slope and seasonal are close to being deterministic. □

4.8 Model Selection

Structural models aim to capture the most obvious parts of a time series; when there are obvious trends or seasonal patterns to capture. ARIMA models are more parsimonious, however they may miss some of the more subtle features of the data.
ARIMA selection is based on the premise that ACFs are stable over time. Even if this is the case, it is still difficult to identify complex models from the ACF alone, and we may miss important features of the series. In practice sampling error in the correlogram can make it difficult to identify even simple ARIMA models. STMs are more robust as the choice of model is not dependent on correlograms.
Model selection for ARIMAs is normally straightforward, unit root tests determine the degree of differencing and lags are included according to some information criterion or statistical significance. However, these tests often have poor size and power properties.

7 Nonlinear Models

Non-linearities can enter a model in a variety of ways. Some common examples are:

- ARMA models with multiplicative terms:

$$, y_t = \phi y_{t-1} + \varepsilon_t + \beta \varepsilon_{t-1} y_{t-1}, \quad \varepsilon_t \sim NID(0, \sigma^2)$$

- Nonlinear functional forms
- Non-Gaussian, EG disturbances follow a t-distribution
- Dynamics in scale rather than location (e.g. GARCH)
- Switching regimes

There are 2 broad classes of models, parameter-driven and observation-driven.
An observation driven model is set up in terms of conditional distribution for the t-th observation: $p(y_t|Y_{t-1}|\phi)$. THe likelihood function is immediately available.
A parameter driven model typically does not allow for a likelihood function, where we have some link function: $y_t = \mu e^{\beta_t} + \varepsilon_t, \quad \beta_t = \phi\beta_{t-1} + \eta_t$, here it is exponential.

7.1 Nonlinear modelling and white noise

7.1.1 Law of iterated expectations

Definition 7.1.1: Law of iterated expectations (LIE)

$$\mathbb{E}[y] = \mathbb{E}_x[\mathbb{E}[y|x]]$$

Proof.

$$\mathbb{E}_x[\mathbb{E}[y|x]] = \int \left[\int yp(y|x)dy \right] p(x)dx = \int \int yp(y,x)dydx = \int yp(y)dy = \mathbb{E}[y]$$

□

This is useful, since we can find a sequence of one step ahead expectations:

$$\mathbb{E}_{t-j}[g(y_t)] = \mathbb{E}_{t-j} \cdots \mathbb{E}_{t-1}[g(y_t)]$$

where the unconditional expectation is found by letting $j \to \infty$.

Exercise 7.1.1. Show

$$Var[y] = \mathbb{E}_x[Var[y|x]] + Var_x[\mathbb{E}[y|x]]$$

Solution:-

$$\begin{aligned} Var[y] &= \mathbb{E}[y^2] - \mathbb{E}[y]^2 \\ &= \mathbb{E}_x[\mathbb{E}[y^2|x]] - \mathbb{E}_x[\mathbb{E}[y|x]]^2 \\ &= \mathbb{E}_x[Var[y|x] + \mathbb{E}[y|x]^2] - \mathbb{E}_x[\mathbb{E}[y|x]]^2 \\ &= \mathbb{E}_x[Var[y|x]] + \mathbb{E}_x[\mathbb{E}[y|x]^2] - \mathbb{E}_x[\mathbb{E}[y|x]]^2 \\ &= \mathbb{E}_x[Var[y|x]] + Var[\mathbb{E}[y|x]] \end{aligned}$$

7.1.2 White noise

White noise is uncorrelated, i.e. $\mathbb{E}[y_t y_s] = 0, \quad t \neq s$ with constant variance (and zero mean).
Strict white noise is stronger, we require independence, not just uncorrelatedness[1].
Martingale Difference has a zero (or constant) conditional expectation:

$$\mathbb{E}_{t-1}[y_t] = 0$$

and thus is uncorrelated with any function of past observations:

$$\mathbb{E}[y_t f(Y_{t-1})|Y_{t-1}] = f(Y_{t-1})\mathbb{E}[y_t|Y_{t-1}] = 0$$

Example.

$$y_t = \varepsilon_t + \beta\varepsilon_{t-1}\varepsilon_{t-2} \quad \varepsilon_t \sim NID(0, \sigma^2)$$

The autocovariance at lag τ can be derived as:

$$\begin{aligned} \mathbb{E}(y_t y_{t-\tau}) &= \mathbb{E}(\varepsilon_t + \beta\varepsilon_{t-1}\varepsilon_{t-2})(\varepsilon_{t-\tau} + \beta\varepsilon_{t-\tau-1}\varepsilon_{t-\tau-2}) \\ &= \mathbb{E}(\varepsilon_t\varepsilon_{t-\tau}) + \beta\mathbb{E}(\varepsilon_t\varepsilon_{t-\tau-1}\varepsilon_{t-\tau-2}) + \beta\mathbb{E}(\varepsilon_{t-1}\varepsilon_{t-\tau}\varepsilon_{t-\tau-1}) + \beta^2\mathbb{E}(\varepsilon_{t-1}\varepsilon_{t-2}\varepsilon_{t-\tau}\varepsilon_{t-\tau-1}) \\ &= 0 \quad \text{if} \quad \tau \neq 0 \end{aligned}$$

Since all observations are uncorrelated the series is white noise, however the observations are not independent, the conditional mean is:

$$\mathbb{E}_{t-1}[y_t] = \mathbb{E}_{t-1}[\varepsilon_t] + \beta\mathbb{E}_{t-1}[\varepsilon_{t-1}\varepsilon_{t-2}] = \beta\varepsilon_{t-1}\varepsilon_{t-2}$$

so the series is not a martingale difference.

Example (ARCH).

$$\begin{aligned} y_t &= \sigma_{t|t-1}\varepsilon_t, \quad \varepsilon_t \sim NID(0,1) \\ \sigma^2_{t|t-1} &= \gamma + \alpha y^2_{t-1} \end{aligned}$$

This is a Martingale difference since

$$\mathbb{E}_{t-1}[y_t] = \sigma_{t|t-1}\mathbb{E}_{t-1}[\varepsilon_t] = 0$$

implying it is also white noise.

[1] These are the same with Gaussian noise, since the distribution is fully defined by the first 2 moments

7.1.3 Linearity and Prediction

When disturbances in an ARMA are IID, the MMSE predictor is the conditional mean. It is linear in the observations and disturbances.
Assuming instead that the disturbances are MDs with mean zero and constant variance, the MMSE predictor is again the conditional expectation by the LIE.
When disturbances are WN (not strict WN) the MMSE = MMSLE, however if disturbances are not independent the MMSE is not the MMSLE.

Example.

$$y_t = \varepsilon_t + \beta\varepsilon_{t-1}\varepsilon_{t-2}, \quad \varepsilon_t \sim NID(0, \sigma^2)$$

The MMSLE of future observations is zero, with MSE equal to the variance of future observations:

$$\begin{aligned} Var(y_t) &= \mathbb{E}[\varepsilon_t^2] + 2\beta\mathbb{E}[\varepsilon_t\varepsilon_{t-1}\varepsilon_{t-2}] + \beta^2\mathbb{E}[\varepsilon_{t-1}^2\varepsilon_{t-2}^2] \\ &= \sigma^2 + 0 + \beta^2\sigma^4 \end{aligned}$$

However the MMSE (the conditional mean) is:

$$\mathbb{E}[y_t|Y_{t-1}] = \beta\varepsilon_{t-1}\varepsilon_{t-2}$$

which has MSE:

$$\mathbb{E}[y_t - \beta\varepsilon_{t-1}\varepsilon_{t-2}]^2 = \mathbb{E}[\varepsilon_t]^2 = \sigma^2$$

We can use the LIE to compute multi-step predictions:

$$\mathbb{E}_T[y_{T+\ell}] = \begin{cases} \mathbb{E}_T[\beta\varepsilon_T\varepsilon_{T-1}] = \beta\varepsilon_T\varepsilon_{T-1} & \ell = 1 \\ \mathbb{E}_T[\beta\varepsilon_{T+1}\varepsilon_T] = 0 & \ell = 2 \\ 0 & \ell > 2 \end{cases}$$

7.2 Stationarity

Theorem 7.2.1 (Krengel's Theorem). If y_t is strictly stationary and ergodic (SE) then a continuous transformation $g(y_t, y_{t-1}, \dots)$ is also SE.

Weak stationarity of g doesn't follow from weak stationarity of y_t since the moments may not exist. E.g. if $y_t \sim t\nu$ then $g(y_t) = e^{y_t}$ has no finite moments.

Definition 7.2.1: Linear stochastic recurrence equation

$$y_{t+1} = x_t y_t + z_t$$

where x_t and z_t are strictly stationary and ergodic.

Theorem 7.2.2. The conditions:

1. $\mathbb{E}(\max(0, \ln|z_t|)) = \mathbb{E}(\ln^+|z_t|) < \infty$
2. $\mathbb{E}(\ln|x_t|) < 0$

are sufficient for the existence and uniqueness of a strictly stationary solution for y_t.

Condition 1 usually holds, it is the second condition that is important. It is known as the **contraction condition** and can be interpreted as saying that the x_t's are on average smaller than 1

(in absolute value). We can see this by applying Jensen's inequality:

$$\mathbb{E}(\ln|x_t|) \leq \ln \mathbb{E}(|x_t|) \Rightarrow \mathbb{E}(|x_t|) < 1$$

Example. If x_t is lognormal, $lnx_t \sim N(\mu, \sigma^2)$ the contraction condition is $\mathbb{E}(\ln x_t) = \mu < 0$, whereat $\mathbb{E}(x_t) = e^{\mu+\sigma^2/2}$, thus $\ln \mathbb{E}(x_t) = \mu + \sigma^2/2 > \mu$ is stronger than needed.

Exercise 7.2.1. What is the stationarity condition for the bilinear model?

$$y_t = \phi y_{t-1} + \varepsilon_t + \beta \varepsilon_{t-1} y_{t-1} \tag{7.1}$$

Solution:-

We can write the model as:

$$y_t = (\phi + \beta \varepsilon_{t-1}) y_{t-1} + \varepsilon_t$$

which is a SRE with $x_t = \phi + \beta \varepsilon_{t-1}$ and $z_t = \varepsilon_t$. The contraction condition is:

$$\mathbb{E}(\ln|x_t|) = \mathbb{E}(\ln|\phi + \beta \varepsilon_{t-1}|) < 0$$

7.2.1 Asymptotic Stationarity

Consider a non-linear generalisation of the linear SRE above:

$$y_{t+1} = \varphi(y_t, \mathbf{z}_t, \psi)$$

where $\mathbf{z}_t$ is a vector of SE variables[2] and ψ is a vector of parameters.

Theorem 7.2.3 (Bougerol's Theorem). If

- There exists y_1 such that $\mathbb{E}(\ln^+|\varphi(y_1, \mathbf{z}_1|)) < \infty$
- $\mathbb{E}(\ln \sup_y |\frac{\partial \varphi(y, \mathbf{z})}{\partial y}|) < 0$

then for any starting value y_1, a series y_t converges exponentially and almost surely to a unique SE solution. In other words $|y_t(y_1, \psi) - y_t(\psi)| \overset{e.a.s.}{\to} 0$ as $t \to \infty$.

Intuitively this just means that it doesn't matter where we start, we will always converge to the same solution.

Example. Consider this cursed AR(1) model:

$$y_t = \phi \frac{e^{y_{t-1}} - 1}{e^{y_{t-1}} + 1} + \varepsilon_t \quad \varepsilon_t \sim NID(0, \sigma^2)$$

Clearly

$$-1 < \frac{e^{y_{t-1}} - 1}{e^{y_{t-1}} + 1} < 1$$

[2] When $\mathbf{z_t}$ is a vector of IID variables, the equation is known as a Markov system

so the first condition holds for any finite y_1. We now examine the second condition:

$$\begin{aligned}\left|\frac{\partial}{\partial y}\left(\phi\frac{e^y-1}{e^y+1}+\varepsilon_t\right)\right| &= \left|\phi\frac{(e^y+1)e^y-(e^y-1)e^y}{(e^y+1)^2}\right| \\ &= |\phi|\left|\frac{2e^y}{(e^y+1)^2}\right| \\ &= |\phi|\frac{2e^y}{(e^y+1)^2}\end{aligned}$$

To solve for the supremum we take the derivative and set it to zero:

$$\begin{aligned}0 &= \frac{d}{dy}\frac{2e^y}{(e^y+1)^2} \\ \Rightarrow 0 &= (e^y+1)^2e^y - 2e^y(e^y+1)e^y \\ \Rightarrow e^y+1 &= 2 \\ \Rightarrow y &= 0 \quad \text{with} \quad \frac{2e^y}{(e^y+1)^2} = \frac{1}{2}\end{aligned}$$

Thus:

$$\mathbb{E}(\ln\sup_y|\frac{\partial\varphi(y,\mathbf{z})}{\partial y}|) = \mathbb{E}(\ln\frac{|\phi|}{2}) = \mathbb{E}(\ln|\phi|) - \ln 2 < 0 \equiv |\phi| < 2$$

7.3 Distributions

Definition 7.3.1: Survival function

$$S(y) = P(Y > y) = 1 - F(y)$$

Example (Exponential).

$$F(y) = 1 - e^{-y/\theta} \Rightarrow S(y) = e^{-y/\theta}$$

Definition 7.3.2: Probability integral transform

The PIT of Y is the standard uniform, i.e. $F(Y) \sim U(0,1)$

$$f(F(Y)) = f(y)\frac{dy}{dF(y)} = 1$$

Thus we can generate any distribution by transforming a standard uniform.

Definition 7.3.3: t-distribution

$$f(y) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})}\left(1+\frac{(y-\mu)^2}{\nu\phi^2}\right)^{-\frac{\nu+1}{2}}$$

where μ is median and ϕ is scale.

The location-dispersion model is:

$$y_t = \mu + \psi\varepsilon_t$$

where ε_t has mean zero, and ψ is called the dispersion for y_t. For non-negative variables a location + scale model is needed:

$$y_t = \psi\varepsilon_t$$

where ε_t has mean one. When $y_t > 0$ taking logarithms gives:

$$\ln y_t = \ln \psi + \ln \varepsilon_t$$

so $\ln \psi$ is now a location parameter.

Definition 7.3.4: Gamma distribution

A gamma(ψ, γ) distribution has density:

$$f(y) = \psi^{-\gamma} \frac{y^{\gamma-1} e^{-y/\psi}}{\Gamma(\gamma)}$$

Where ψ is the scale parameter and γ is the shape parameter. The mean is $\gamma\psi$ and the variance is $\gamma\psi^2$.

The chi-squared distribution is gamma$(2, \nu/2)$, setting $\gamma = 1$ gives an exponential distribution.

Definition 7.3.5: Log-logistic distribution

A log-logistic(ψ, γ) distribution has density:

$$f(y) = \frac{\nu \frac{y}{\psi}^{\nu-1}}{\psi \left(1 + \left(\frac{y}{\psi}\right)^{\nu}\right)^2}$$

Definition 7.3.6: Quantiles

The α-quantile of a distribution is the value y_α such that $F(y_\alpha) = \alpha$. The median is the 0.5-quantile.

Definition 7.3.7: Heavy tails

A distribution is said to be heavy tailed if

$$\lim_{y \to \infty} \exp(y/\alpha) S(y) = \infty \quad \forall \alpha > 0$$

Example (Exponential distribution). $S(y) = e^{-y/\alpha}$ so

$$\exp(y/\alpha) S(y) = e^{y/\alpha} e^{-y/\alpha} = 1$$

thus it is not heavy tailed.

Definition 7.3.8: Fat tails

A distribution is said to have fat tails if

$$S(y) = cL(y) y^{-\alpha} \quad \alpha > 0$$

where $L(y)$ is slowly varying, i.e. $\lim_{y \to \infty} L(\lambda y)/L(y) = 1$ for all $\lambda > 0$ and c is a non-negative constant.

Claim 7.3.1. Fat tailed ⇒ heavy tailed, but not the reverse.

7.4 Nonlinear state space models

Parameter driven models may be nonlinear in the measurement equation, the transition equation or both. The basic model is:

$$y_t = f(\theta_t, \varepsilon_t|\varphi)$$
$$\theta_{t+1} = \psi(\theta_t, \eta_t|\varphi)$$

where θ is the signal, φ parameters and ε, η disturbances with specified distributions.

Example (AR1 dynamic equation)**.** Consider

$$y_t = \mu \exp(\beta_t) + \varepsilon_t \quad \varepsilon_t \sim NID(0, \sigma_\varepsilon^2)$$
$$\beta_{t+1} = \phi\beta_t + \eta_t \quad \eta_t \sim NID(0, \sigma_\eta^2) \quad |\phi| < 1$$

Clearly the state equation follows an AR1 with parameter less than 1, it is SE. We can thus apply Krengel's theorem to show that y_t is also SE.

Example (Non-negativity)**.** When y_t is non-negative, any model must be non-linear. Consider the measurement equation below, with time varying mean μ_t:

$$y_t = \mu_t\varepsilon_t \quad 0 \leq y_t < \infty$$

where ε_t has a gamma distribution with mean 1. We can use an exponential link function to model the logarithm of mu_t to ensure mu_t remains positive:

$$\ln \mu_{t+1} = \delta + \phi \ln \mu_t + \alpha\eta_t$$

The restriction $|phi| < 1$ guarantees the stationarity of $\ln \mu_t$, and hence of y_t.

We can also consider conditionally Gaussian state space models, allowing the possibility of feedback from past observations to the system matrices that would otherwise be a linear SSM. We can derive the Kalman filter exactly as before and construct the likelihood function since the system matrixes are fixed at time $t-1$. It is not usually possible to predict more than one-step ahead however.

Example. An example of a conditionally Gaussian process is

$$y_t = \phi_t y_{t-1} + \varepsilon_t$$

$$\phi_{t+1} = \phi(1-\alpha) + \alpha\phi_t + \eta_t$$

where α is a fixed parameter. When ϕ_t is regarded as the state, we have a conditionally Gaussian SSM with $z = y_{t-1}$. When $|\alpha| < 1$ the model is SE.

7.5 Observation driven models

An observation driven model is set up to give a conditional distribution for each observation, that is:

$$p(y_t|Y_{t-1}|\varphi)$$

where Y_{t-1} denotes all past observations. We can then construct the likelihood in the usual way.

Example (Bilinear model).

$$y_t = \phi y_{t-1} + \varepsilon_t + \beta \varepsilon_{t-1} y_{t-1}$$

We can compute the ML estimate by minimising the SSR. Setting up the model in SSF:

$$y_t = \mu_{t|t-1} + \varepsilon_t$$
$$\mu_{t+1|t} = \phi y_t + \beta \varepsilon_t y_t = (\phi + \beta \varepsilon_t)\mu_{t|t-1} + \phi \varepsilon_t + \beta \varepsilon_t^2$$

More general observation models can be written:

$$p(y_t | \alpha_{t|t-1}, Y_{t-1} | \varphi)$$

with the transition equation becoming a filtering equation

$$\alpha_{t+1|t} = g(\alpha_{t|t-1}, y_t, Y_{t-1} | \varphi)$$

The signal θ is the dynamic parameter in the conditional distribution. It depends on $\alpha_{t+1|t}$ so

$$\theta_{t|t-1} = h(\alpha_{t|t-1}, y_t, Y_{t-1} | \varphi)$$

and we can write

$$p(y_t | h(\alpha_{t|t-1}, Y_{t-1} | \varphi))$$

Commonly it is useful to keep the state variable α positive, so we can use an exponential link function:

$$\alpha_{t+1|t} = e^{\theta_{t+1|t}}$$

7.5.1 Moment-driven and score-driven models

In a moment-driven model the dynamic equation for the state is driven by observed y_t, E.G.:

$$\mu_{t+1|t} = \delta + \beta \mu_{t|t-1} + \alpha y_t$$

The score-driven model is driven by the score of the conditional distribution, that is:

$$\mu_{t+1|t} = \lambda + \phi \mu_{t|t-1} + \kappa \left(k \frac{\partial \ln p(y_t | \mu_{t|t-1})}{\partial \mu_{t|t-1}} \right)$$

Consider the linear Gaussian AR1N model. The finite sample Kalman filter begins at $t = 1$ taking account of the uncertainty with a diffuse prior. However we might consider a model based on the steady-state (constant Kalman gain), but initialised at $t = 1$. We write this as:

$$y_t = \mu + \mu_{t|t-1} + \nu_t$$

$$\mu_{t+1|t} = \phi \mu_{t|t-1} + \kappa \nu_t$$

where the innovations ν_t are NID and $\mu_{1|0}$ is fixed. We can write the filter as

$$\begin{aligned} \mu_{t+1|t} &= \phi \mu_{t|t-1} + \kappa \nu_t \\ &= \phi \mu_{t|t-1} + \kappa (y_t - \mu - \mu_{t|t-1}) \\ &= (\phi - \kappa)\mu_{t|t-1} + \kappa (y_t - \mu) \end{aligned}$$

Substituting repeatedly gives

$$\mu t + 1|t = \kappa \sum_{j=0}^{t-1} (\phi - \kappa)^j (y_{t-j} - \mu) + (\phi - \kappa)^t \mu_{1|0}$$

Although the filter is typically started with $\mu_{1|0} = 0$, as long as $|\phi - \kappa| < 1$ the filtered level will converge to the same value for any starting value.
In a score-driven model the dynamic equation for the above is driven by a variable that is proportional to the score of the conditional distribution, that is:

$$\mu_{t+1|t} = \phi\mu_{t|t-1} + \kappa\left(k\frac{\partial \ln p(y_t|\mu_{t|t-1})}{\partial \mu_{t|t-1}}\right)$$

The inverse of the information matrix is a common choice for k.

Example (Non-negativity - Gamma distribution)**.** For non-negative variables we use a location-scale model. Suppose $p(y_t|\mu_{t|t-1}, \psi)$ is gamma with shape parameter γ and evolving scale parameter $\mu_{t|t-1}$.
Moment-driven Model
A filter for the scale can be written as:

$$\mu_{t+1|t} = \delta + \beta\mu_{t|t-1} + \alpha y_t$$

We can write the likelihood function as

$$L = \prod_{t=1}^{T} \frac{1}{\Gamma(\gamma)}\left(\frac{\gamma}{\mu_{t|t-1}}\right)^{\gamma} y_t^{\gamma-1} e^{-\frac{\gamma y_t}{\mu_{t|t-1}}}$$

and thus the log-likelihood is:

$$\ln L = -T\ln\Gamma(\gamma) + T\gamma\ln\gamma - \gamma\sum_{t=1}^{T}\ln\mu_{t|t-1} + (\gamma-1)\sum_{t=1}^{T}\ln y_t - \gamma\sum_{t=1}^{T}\frac{y_t}{\mu_{t|t-1}}$$

We can use MLE to find the optimal values of δ, β, α.
Score-driven Model
The scale evolves according to:

$$\mu_{t+1|t} = \lambda + \phi\mu_{t|t-1} + \kappa u_t$$

Differentiating the likelihood with respect to $\mu_{t|t-1}$ gives the score:

$$\frac{\partial \ln L}{\partial \mu_{t|t-1}} = -\frac{\gamma}{\mu_{t|t-1}} + \frac{\gamma y_t}{\mu_{t|t-1}^2} = \frac{\gamma(y_t - \mu_{t|t-1})}{\mu_{t|t-1}^2}$$

To normalise the score we first need to find the Fisher information, i.e. the expectation of the square of the score:

$$\mathbb{E}\left(\frac{\partial \ln L}{\partial \mu_{t|t-1}}^2\right) = \frac{\gamma}{\mu_{t|t-1}^4}\mathbb{E}(y_t - \mu_{t|t-1})^2$$

However note that this is just the variance of y_t which follows a gamma distribution, so the Fisher information is:

$$\mathbb{E}\left(\frac{\partial \ln L}{\partial \mu_{t|t-1}}^2\right) = \frac{\gamma}{\mu_{t|t-1}^4}\mu_{t|t-1}^2 = \frac{\gamma}{\mu_{t|t-1}^2}$$

Thus we can compute the standardised score as:

$$\hat{u}_t = \frac{\partial \ln L}{\partial \mu_{t|t-1}}\left(I(\mu_{t|t-1})\right)^{-1} = y_t - \mu_{t|t-1}$$

Equivalence
By substituting the score back into the filter we can see it takes the same form as the moment-driven model, this is an exceptional case - generally the two models are not equivalent.

$$\begin{aligned}\mu_{t+1|t} &= \lambda + \phi\mu_{t|t-1} + \kappa u_t \\ &= \lambda + \phi\mu_{t|t-1} + \kappa(y_t - \mu_{t|t-1}) \\ &= \lambda + (\phi - \kappa)\mu_{t|t-1} + \kappa y_t\end{aligned}$$

7.5.2 Multiplicative Error Models (MEMs)

Models for location/scale based on the moment-driven filter were called MEMs by Engle (2002). These are called as such since we can write the model as:

$$y_t = \mu_{t|t-1}\varepsilon_t$$

where ε_t follows the same distribution as y_t but with unit scale. The dynamic equation for the scale is thus:

$$\mu_{t+1|t} = \delta + \beta\mu_{t|t-1} + \alpha\varepsilon_t\mu_{t|t-1}$$

7.5.3 Stationarity

We can write the above as:

$$\mu_{t+1|t} = \delta + (\beta + \alpha\varepsilon_t)\mu_{t|t-1} \coloneqq \delta + x_t\mu_{t|t-1}$$

Contraction condition is: $\mathbb{E}(\ln|x_t|) = \mathbb{E}(\ln|\beta + \alpha\varepsilon_t|) < 0$. Thus strict stationarity is guaranteed if $\beta + \alpha < 1$. However we have used Jensen's here, so it is possible to have a stationary solution even if $\beta + \alpha \geq 1$.
If this holds, we know $\mu_{t|t-1}$ is stationary, and thus y_t is also stationary by Krengel's theorem.

7.5.4 Invertibility

The model is invertible if the signal can be recovered from the observations, a property we need to forecast. Denote the state as $\theta_{t|t-1}$.

Definition 7.5.1: Lipschitz Coefficient

The stochastic Lipschitz coefficient is:

$$\Lambda_t(\psi) = \sup_{\theta_{t|t-1}} \left|\frac{\partial\theta_{t+1|t}}{\partial\theta_{t|t-1}}\right|$$

When $\theta_{t+1|t}$ and hence y_t is strictly stationary and ergodic, invertibility is guaranteed by the Lyapunov condition:

Definition 7.5.2: Lyapunov Condition

$$\sup_{\psi}\mathbb{E}\left(\ln\Lambda_t(\psi)\right) < 0$$

Example. Consider the gamma example from before, since observations are taken as given in

the moment driven filter we have:

$$\frac{\partial \mu_{t+1|t}}{\partial \mu_{t|t-1}} = \beta$$

giving us a Lipschitz coefficient of $\Lambda_t(\psi) = |\beta|$. The Lyapunov condition is thus satisfied if $|\beta| < 1$.

7.5.5 Moments

Given the SRE: $\mu_{t+1|t} = \delta + x_t \mu_{t|t-1}$ where $x_t = \alpha + \beta\varepsilon_t$ we can find the moments of $\mu_{t|t-1}$ with stationarity.

Mean:

$$\begin{aligned}
\mathbb{E}_{t-1}[\mu_{t+1|t}] &= \delta + \mathbb{E}_{t-1}[x_t \mu_{t|t-1}] = \delta + (\alpha+\beta)\mu_{t|t-1} \\
\mathbb{E}_{t-2}\mathbb{E}_{t-1}[\mu_{t+1|t}] &= \delta + (\alpha+\beta)(\delta + (\alpha+\beta)\mu_{t-1|t-2}) = \delta + (\alpha+\beta)\delta + (\alpha+\beta)^2\mu_{t-1|t-2} \\
&= \cdots \\
\mathbb{E}\mu_{t+1|t} &= \frac{\delta}{1-\alpha-\beta}
\end{aligned}$$

Provided that $|\alpha + \beta| < 1$. Obviously if we are happy to assume the unconditional mean exists, just take expectations of the SRE.

Variance: Squaring both sides of the SRE gives us:

$$\begin{aligned}
\mu^2_{t+1|t} &= \delta^2 + 2\delta x_t \mu_{t|t-1} + x_t^2 \mu^2_{t|t-1} \\
\mathbb{E}_{t-1}[\mu^2_{t+1|t}] &= \delta^2 + 2\delta \mathbb{E}_{t-1}[x_t \mu_{t|t-1}] + \mathbb{E}_{t-1}[x_t^2 \mu^2_{t|t-1}] \\
&= \delta^2 + 2\delta(\alpha+\beta)\mu_{t|t-1} + ((\alpha+\beta)^2 + \alpha^2\sigma_\varepsilon^2)\mu^2_{t|t-1}
\end{aligned}$$

We can iterate as above, or just take unconditional expectations:

$$\begin{aligned}
\mathbb{E}[\mu^2_{t+1|t}] &= \frac{\delta^2 + 2\delta^2 \frac{\alpha+\beta}{1-\alpha-\beta}}{1-(\alpha+\beta)^2 - \alpha^2\sigma_\varepsilon^2} \\
\Rightarrow Var[\mu_{t+1|t}] &= \mathbb{E}[\mu^2_{t+1|t}] - \mathbb{E}[\mu_{t+1|t}]^2 \\
&= \frac{\delta^2}{(1-\alpha-\beta)^2} \frac{(1-\beta^2-2\alpha\beta)\sigma_\varepsilon^2}{(1-(\alpha+\beta)^2-\alpha^2\sigma_\varepsilon^2)}
\end{aligned}$$

ACF

We can rewrite the model in innovations form by letting $v_t = y_t - \mu_{t|t-1}$ so

$$\mu_{t+1|t} = \delta + (\alpha+\beta)\mu_{t|t-1} + \kappa v_t$$

We can see that v_t is a martingale difference, since $\mathbb{E}v_t = \mu_{t|t-1}\mathbb{E}(\varepsilon_t - 1) = 0$. It is also WN since MD + constant variance $\Rightarrow$ WN:

$$\mathbb{E}v_t^2 = \mathbb{E}\mu^2_{t|t-1}(\varepsilon_t - 1)^2 = \sigma_\varepsilon^2 \mathbb{E}\mu^2_{t|t-1}$$

which we know is constant from variance proof above. Subbing the filter into the definition of the innovation we can see:

$$\begin{aligned}
\mu_{t+1|t} &= \delta + (\alpha+\beta)\mu_{t|t-1} + \kappa v_t \\
&= \frac{\delta + \kappa v_t}{1-(\alpha+\beta)L} \\
\Rightarrow y_t &= \frac{\delta + \kappa v_{t-1}}{1-(\alpha+\beta)L} + v_t
\end{aligned}$$

which follows an ARMA(1,1) and the ACF has autocorrelations from $\tau = 1$ decaying exponentially.

7.5.6 Prediction

Multi-step predictions can be found using the Law of Iterated Expectations. Taking conditional expectations in the the filter gives:

$$\begin{aligned}
\mu_{T+\ell|T+\ell-1} &= \delta + (\beta + \alpha\varepsilon_{T+\ell-1})\mu_{T+\ell-1|T+\ell-2} \\
\Rightarrow \mathbb{E}_{t+\ell-2}\mu_{T+\ell|T+\ell-1} &= \delta + (\beta + \alpha)\mu_{T+\ell-1|T+\ell-2} \\
&= \cdots \\
\Rightarrow \mathbb{E}_T\mu_{T+\ell|T+\ell-1} &= \frac{1-(\alpha+\beta)^{\ell-1}}{1-\alpha-\beta}\delta + (\alpha+\beta)^{\ell-1}\mu_{T+1|T}
\end{aligned}$$

Using the MEM we can see:

$$\begin{aligned}
y_{T+\ell} &= \mu_{T+\ell|T}\varepsilon_{T+\ell} \\
\mathbb{E}_T y_{T+\ell} &= \mu_{T+\ell|T} \\
\lim_{\ell\to\infty}\mathbb{E}_T y_{T+\ell} &= \frac{\delta}{1-\alpha-\beta}
\end{aligned}$$

As the forecast becomes infinitely distant, we revert to the unconditional mean.

7.6 Log-Logistic example

The choice of link function is important for the invertibility and stationarity of a model. With a direct link function stationarity requires the contraction condition, and the MEM is invertible by construction. However, the invertibility of a model with an exponential link function is not easy to check. We give an example for a fat-tailed log-logistic distribution with an exponential link function.

$$\begin{aligned}
p(y_t|\nu,\psi_{t|t-1}) &= (\nu\psi_{t|t-1})(y_t\psi_{t|t-1})^{\nu-1}(1+(y_t\psi_{t|t-1})^{\nu})^{-2} \\
\psi_{t|t-1} &= e^{\lambda_{t|t-1}} \\
\lambda_{t+1|t} &= \delta + \phi\lambda_{t|t-1} + \kappa u_t
\end{aligned}$$

We can derive the score as usual by setting up the log-likelihood and differentiating with respect to $\psi_{t|t-1}$:

$$\begin{aligned}
\ln p(y_t|\nu,\psi_{t|t-1}) &= \ln\nu + \ln\psi_{t|t-1} + (\nu-1)\ln y_t + (\nu-1)\ln\psi_{t|t-1} - 2\ln(1+(y_t\psi_{t|t-1})^{\nu}) \\
\frac{d\ln p(y_t)}{d\psi_{t|t-1}} &= \frac{1}{\psi_{t|t-1}} + \frac{\nu-1}{\psi_{t|t-1}} - \frac{2\nu y_t^{\nu}\psi_{t|t-1}^{\nu-1}}{1+(y_t\psi_{t|t-1})^{\nu}} \\
&= \frac{\nu}{\psi_{t|t-1}} - \frac{2\nu y_t^{\nu}\psi_{t|t-1}^{\nu-1}}{1+(y_t\psi_{t|t-1})^{\nu}} \\
\psi_{t|t-1} = e^{\lambda_{t|t-1}} &\Rightarrow \frac{d\psi_{t|t-1}}{d\lambda_{t|t-1}} = \psi_{t|t-1} \quad \text{From definition of exponential link} \\
\Rightarrow \frac{d\ln p(y_t)}{d\lambda_{t|t-1}} &= \frac{d\ln p(y_t)}{d\lambda_{t|t-1}}\frac{d\psi_{t|t-1}}{d\lambda_{t|t-1}} \\
&= \frac{d\ln p(y_t)}{d\lambda_{t|t-1}}\psi_{t|t-1} \\
u_t &\coloneqq \nu - \frac{2\nu(y_t\psi_{t|t-1})^{\nu}}{1+(y_t\psi_{t|t-1})^{\nu}}
\end{aligned}$$

We can write the model in MEM form, where $y_t = \psi_{t|t-1}\varepsilon_t$ and ε_t is a log-logistic random variable with shape parameter ν and scale parameter 1. Thus we can rewrite the score in terms of ε_t:

$$u_t = \nu - \frac{2\nu\varepsilon_t^\nu}{1+\varepsilon_t^\nu}$$

We know the distribution of ε_t: $p(\varepsilon_t) = \nu\varepsilon_t^{\nu-1}(1+\varepsilon_t^\nu)^{-2}$, we just need to find the distribution of $b_t := \frac{\varepsilon_t^\nu}{1+\varepsilon_t^\nu}$ to find the variance of u_t (to derive Fisher Information).

Lemma 7.6.1. $p(f(x)) = p(x)\frac{dx}{df(x)}$

Thus we can write the distribution of b_t as:

$$\begin{aligned}
p(b_t) &= p(\varepsilon_t)\frac{d\varepsilon_t}{db_t} \\
&= p(\varepsilon_t)\left(\frac{\nu\varepsilon_t^{\nu-1}}{(1+\varepsilon_t^\nu)^2}\right)^{-1} \\
&= \nu\varepsilon_t^{\nu-1}(1+\varepsilon_t^\nu)^{-2}\left(\frac{\nu\varepsilon_t^{\nu-1}}{(1+\varepsilon_t^\nu)^2}\right)^{-1} \\
&= 1
\end{aligned}$$

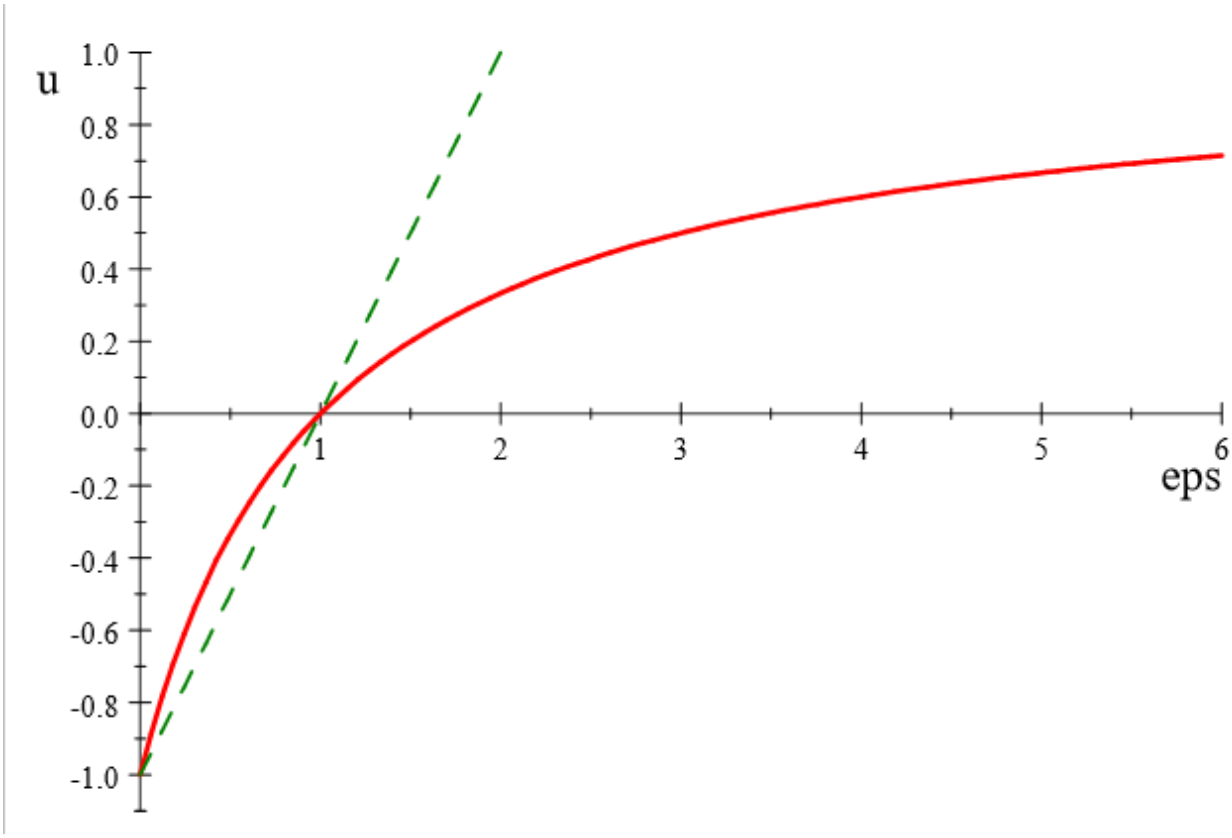
Thus b_t is a standard uniform random variable, with mean 0.5 and variance $\frac{1}{12}$. We can now find the Fisher Information:

$$\begin{aligned}
Var(u_t) &= Var\left(\nu - \frac{2\nu\varepsilon_t^\nu}{1+\varepsilon_t^\nu}\right) \\
&= 4\nu^2 Var(b_t) \\
&= \frac{4\nu^2}{12} \\
&= \frac{\nu^2}{3}
\end{aligned}$$

We can now find the standardised score:

$$\hat{u}_t = \frac{3-6b_t}{\nu} \in \left[-\frac{3}{\nu}, \frac{3}{\nu}\right]$$

The bounded score gives us a nicer invertibility condition, as well as making the model more robust to outliers, plotted below:



7.6.1 Stationarity and Invertibility

Stationarity
Stationarity of the filter follows from $\phi < 1$, Krengel's theorem implies stationarity of y_t.
Invertibility
The Lipschitz coefficient is

$$\begin{aligned}
\Lambda_t(\psi) &= \sup_{\lambda_{t|t-1}} \left| \frac{\partial \lambda_{t+1|t}}{\partial \lambda_{t|t-1}} \right| \\
&= \sup_{\lambda_{t|t-1}} \left| \phi + \frac{\partial \hat{u}_t}{\partial \lambda_{t|t-1}} \right| \\
&= \sup_{\lambda_{t|t-1}} |\phi - 6\kappa b_t(1-b_t)| \\
&= \left| \phi - \frac{3}{2}\kappa \right| \quad \text{since } x(1-x) \text{ is maximised at 0.5, with value 0.25}
\end{aligned}$$

The Lyapunov condition is thus satisfied if:

$$\begin{aligned}
\sup_{\nu} \mathbb{E} \ln \Lambda_t(\psi) = \mathbb{E} \ln \left| \phi - \frac{3}{2}\kappa \right| < 0 \\
\left| \phi - \frac{3}{2}\kappa \right| < 1 \quad \text{by Jensen's} \\
\left| \frac{3}{2}\kappa \right| - |\phi| < 1 \quad \text{by triangle inequality} \\
|\kappa| < \frac{2}{3}(1 + |\phi|)
\end{aligned}$$