Econ 623 Econometrics II
Topic 2: Stationary Time Series

1 Introduction

- In the regression model we can model the error term as an autoregression – AR(1) process. That is, we can use the past value of the error term to explain the current value of the error term. This idea can be generalized to any variable, such as, $Y_t = \phi Y_{t-1} + e_t$
- Time series model: a series is modeled only in terms of its own past values, time trend and some disturbance. Often, a series which is modeled only in terms of it own past values and some disturbance is referred to as a time series model.
- Why time series? Difference between regression models and time series models?
  - Regression: assume one variable is related to another variable. Once we know the relationship, we know how to make inference and we know how to use one variable(s) to predict another variable. In the regression world, all different variables are related to each other according to economic theories, intuition or expectation.

- Time series: assume one variable is related to its history. There is something which triggers the series to evolve over time. We are not making any attempt to discover the relationships between different variables probably because the relationships are too complicated or the underlying economic theory is not so clear. However, we try to find out the underlying dynamic behavior of one variable over time (time series pattern triggered by any system). The mechanism which produces an economic varible over time is referred to as the **data generating process** (DGP). If we know DGP, we know everything about the time series properties of any variable. Hence we can make inferences about the variable and use the realised values of this variable to forecast future values of this variable since we believe the new series will be also generated by the same DGP.
- GDP=f(monetary policy, fiscal policy, inflation, real interest, import, export,...)
- GDP_t=f(GDP_{t-1},GDP_{t-2}, ...)

• $X_t = f(X_{t-1}, X_{t-2}, ..., \varepsilon_t, \varepsilon_{t-1}, ...)$

2 Important Concepts

- A stochastic process is a set of random variable, typically indexed by time.

- $\{X_t\}_{t=1}^{T}$ is a Gaussian white noise procesess if $X_t \sim N(0, \sigma^2)$

- A realization is a set of T observations, say $\{X_t^{(k)}\}_{t=1}^{T}$. If we have K independent realizations, ie $\{X_t^{(k)}\}_{t=1}^{T}, k = 1, ..., K$, then $X_t^{(1)}, ..., X_t^{(K)}$ can be described as a sample of K realizations of random variable X_t. This random variable has some density which is typically called the marginal density or unconditional density $(f(x))$. From the unconditional density, one can obtained **unconditional moments**. For example, the unconditional mean is $E(X_t) = \int\limits_{-\infty}^{+\infty} x f(x) dx \equiv \mu_t$. By SLLN for an iid sequence, $\text{plim} \frac{1}{K} \sum\limits_{k=1}^{K} X_t^{(k)} = E(X_t)$. Similarly, one can define unconditional variance.

- Given a particular realization $\{X_t^{(1)}\}_{t=1}^{T}$, we construct a vector $Y_t^{(1)}$ which contains the $[j+1]$ most recent observations on X: $Y_t^{(1)} = \begin{bmatrix} X_t^{(1)} & \cdots & X_{t-j}^{(1)} \end{bmatrix}'$. As before, if we have K independent realizations, for each vector, we have K independent realizations, $Y_t^{(k)}, k = 1, ..., K$. This random vector has a joint density $(f(x_t, ..., x_{t-j}))$. From the joint density, one can obtained **autocovariance**. For example, jth autocovariance of X_t is
  $\gamma_{jt} = \int\limits_{-\infty}^{+\infty} \cdots \int\limits_{-\infty}^{+\infty} (x_t - E(X_t))(x_{t-j} - E(X_{t-j})) f(x_t, ..., x_{t-j}) dx_t \cdots dx_{t-j} = E(X_t - \mu_t)(X_{t-j} - \mu_{t-j})$
  By SLLN, $\text{plim} \frac{1}{K} \sum\limits_{k=1}^{K} (X_t^{(k)} - \mu_t)(X_{t-j}^{(k)} - \mu_{t-j}) = E(X_t - \mu_t)(X_{t-j} - \mu_{t-j})$.

- $\{X_t\}_{t=1}^T$ is a **covariance stationary process** if

    1. $E(X_t) = \mu < \infty \;\; \forall \; t$
    2. $Var(X_t) = \sigma^2 < \infty \;\; \forall \; t$
    3. $Cov(X_t, X_{t-j}) = \gamma_j \; \forall \; t, \; \forall \; j \neq 0.$

- What happens when we only work with one realization? For example, what are the properties of the time series average $\frac{1}{T}\sum_{t=1}^{T} X_t^{(1)}$. Note that in general SLLN for the iid sequence is not applicable any more since $X_t^{(1)}$ is not independent over t. Whether $\frac{1}{T}\sum_{t=1}^{T} X_t^{(1)}$ converges to μ for a stationary process has to do with **ergodicity**. A covariance stationary process is said to be ergodic for the mean if $\text{plim}\frac{1}{T}\sum_{t=1}^{T} X_t^{(1)} = \mu$ as $T \to \infty$. It requires the autocovariance γ_j goes to zero sufficiently quickly as j becomes large (for example $\sum_{j=0}^{\infty} |\gamma_j| < \infty$ for covariance stationary processes). Similarly, a covariance stationary process is said to be ergodic for second moments if $\text{plim}\frac{1}{T-j}\sum_{t=j+1}^{T} (X_t^{(1)} - \mu)(X_{t-j}^{(1)} - \mu) = \gamma_j$.

- Partial correlation coefficient
  - Three variables: X_1, X_2, X_3. Let ρ_{12} be the correlation coefficient between X_1 and X_2. Similarly we define ρ_{13} and ρ_{23}.
  - Partial correlation coefficient between X_1 and X_3, with the influence of variable X_2 removed, is defined as $\frac{\rho_{13}-\rho_{12}\rho_{23}}{\sqrt{1-\rho_{12}^2}\sqrt{1-\rho_{23}^2}}$
  - In fact it can be estimated by the OLS estimator of β_2 in the following regression model

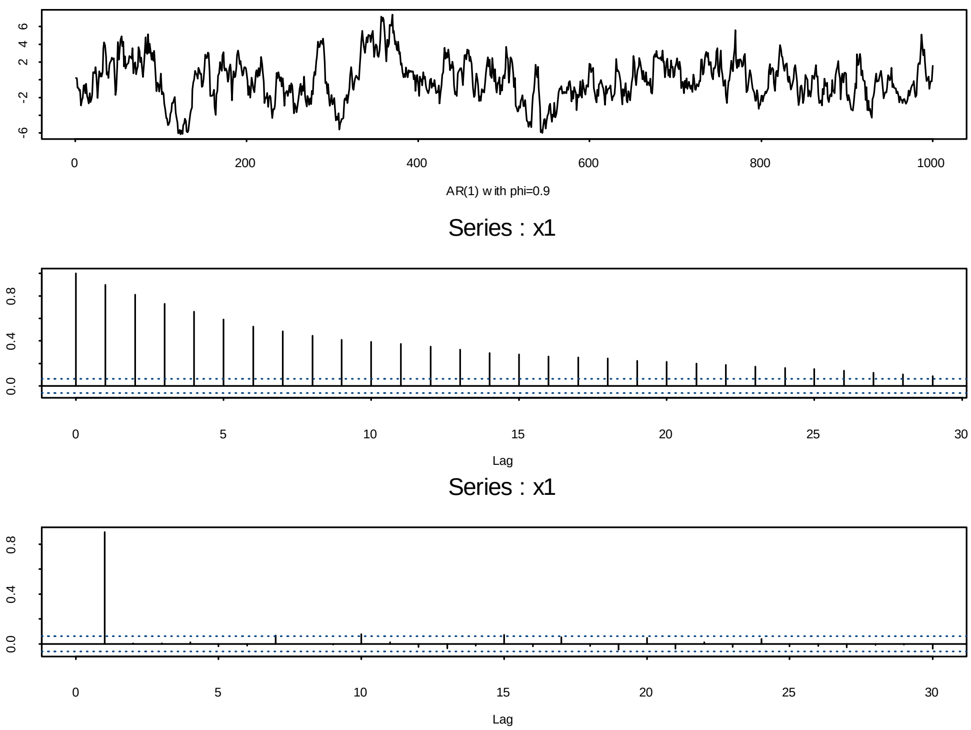
    $X_t = \beta_0 + \beta_1 X_{t-1} + \beta_2 X_{t-2} + e_t$

- ACF (autocorrelation function): a function of h.
  - $\rho_j = \gamma_j/\gamma_0$, where $\gamma_0 = Var(X_t)$
  - $\rho_0 = 1$
  - The graphical representation is called "correlogram".
  - ACF at j can be estimated by the sample counterpart, $\frac{\sum_{t=j+1}^{T}(X_t-\overline{X})(X_{t-j}-\overline{X})}{\sum_{t=1}^{T}(X_t-\overline{X})^2}$. This is a consistent estimator of ACF.
- PACF (partial ACF): a function of k.
  - PACF at k is the correlation between X_t and X_{t-k}, with the influence of variable $X_{t-1}, ..., X_{t-k+1}$ removed
  - PACF at k can be estimated by running a regression and obtaining the coefficient of X_{t-k}: $\widehat{X}_t = \alpha_0(k) + \alpha_1(k)X_{t-1} + ... + \alpha_k(k)X_{t-k}$. $\hat{\alpha}_k(k)$ is a consistent estimator of PACF.
  - The graphical representation is called "partial correlogram".

3 Autoregressive and Moving Average (ARMA) Model

- AR(1): $X_t = \phi X_{t-1} + e_t \quad e_t \sim iidN(0, \sigma_e^2)$
  - $X_t = e_t + \phi e_{t-1} + \phi^2 e_{t-2} + \cdots$ if $|\phi| < 1$ (*)
  - $E(X_t) = 0, Var(X_t) = \frac{\sigma_e^2}{1-\phi^2}$ if $|\phi| < 1$
  - $\rho_j = \phi^j, \forall j$. ACF geometrically decreases with the number of the lags
  - Since $\sum_{j=0}^{\infty} |\gamma_j| < \infty$ when $|\phi| < 1$, it is ergodic for the mean.
  - $\alpha_1(1) = \phi, \alpha_k(k) = 0 \forall k > 1$. PACF cuts off after the first lag

- AR(1) with a drift: $X_t = \mu + \phi X_{t-1} + e_t \quad e_t \sim iidN(0, \sigma_e^2)$
  - $(1 - \phi L)X_t = \mu + e_t$
  - $X_t = \frac{\mu}{1-\phi} + e_t + \phi e_{t-1} + \phi^2 e_{t-2} + \cdots$ if $|\phi| < 1$
  - covariance stationary if $|\phi| < 1$
  - not covariance stationary if $|\phi| \geq 1$
  - a unit root process if $\phi = 1$. A more general unit root process is $X_t = \mu + X_{t-1} + u_t$ where u_t is stationary.
  - explosive if $\phi > 1$.
  - $E(X_t) = \frac{\mu}{1-\phi}, Var(X_t) = \frac{\sigma_e^2}{1-\phi^2}$ if stationary
  - $\rho_j = \phi^j, \forall j$. ACF geometrically decreases with the number of the lags
  - Since $\sum_{j=0}^{\infty} |\gamma_j| < \infty$ when $|\phi| < 1$, it is ergodic for the mean.
  - $\alpha_1(1) = \phi, \alpha_k(k) = 0 \forall k > 1$. PACF cuts off after the first lag



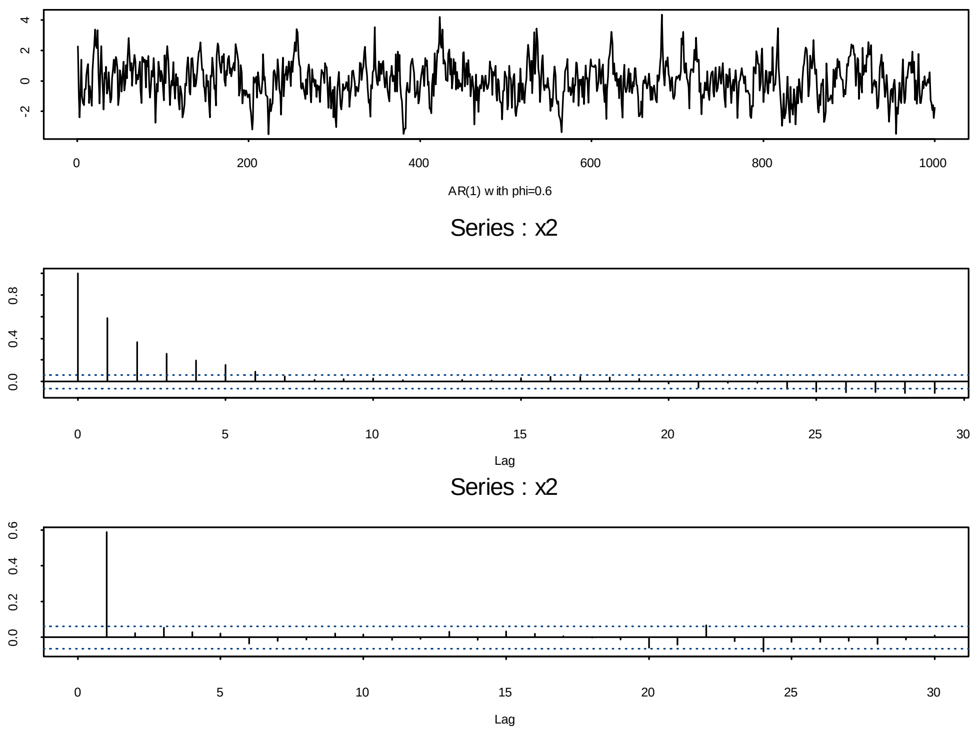
– Estimation of ϕ in $Y_t = \phi Y_{t-1} + e_t \quad e_t \sim iidN(0, \sigma_e^2)$

1. OLS = conditional maximum likelihood (ML) under Gaussianity.

2. Conditional ML or exact ML if we know the distribution function of e. $X_{t+1}|X_t \sim N(\phi X_{t-1}, \sigma_e^2)$.

3. $\widehat{\phi}_{ols}$ has no exact distribution function even under normality assumption on e_t. Phillips (1977, Econometrica) found an edgeworth expansion to approximate the finite sample distribution of $\widehat{\phi}_{ols}$. $E(\widehat{\phi}_{ols})$ does not have a closed-form expression. White (1961, Biometrika) showed that $E(\widehat{\phi}_{ols}) \approx \phi - \frac{2\phi}{T}$

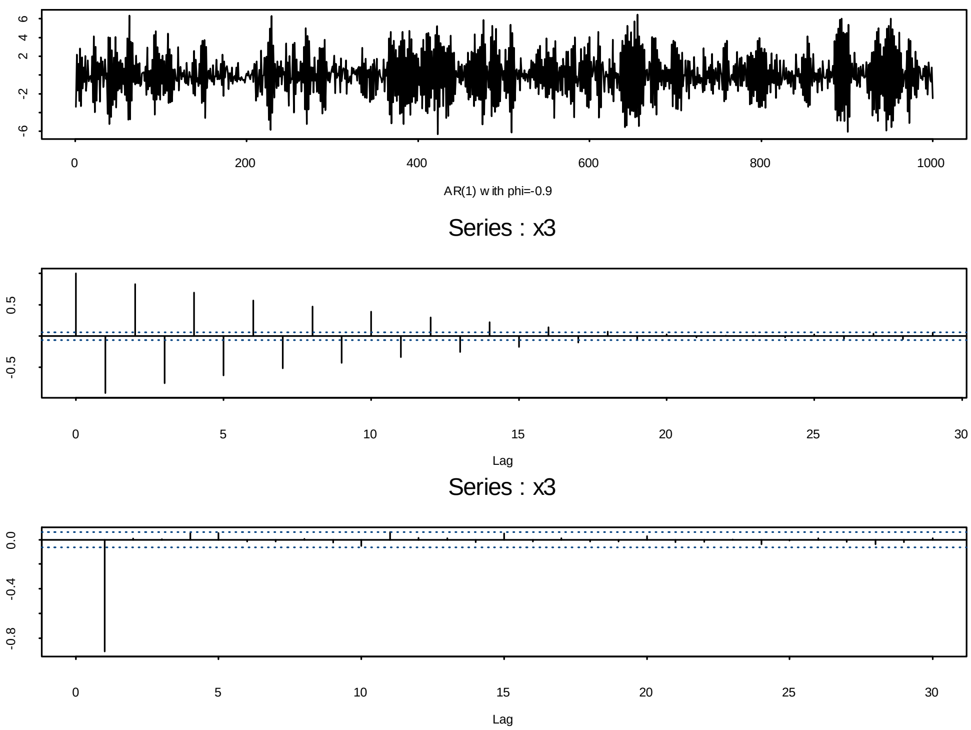
4. $\widehat{\phi}_{ols}$ is consistent. We have to use the asymptotic distribution to test a hypothesis.



– Estimation of ϕ in $X_t = \mu + \phi X_{t-1} + e_t \quad e_t \sim iidN(0, \sigma_e^2)$

1. OLS = conditional maximum likelihood (ML) under Gaussianity.
2. Conditional ML or exact ML if we know the distribution function of e.

   $X_{t+1}|X_t \sim N(\mu + \phi X_{t-1}, \sigma_e^2)$
3. $\widehat{\phi}_{ols}$ has no exact distribution function even under normality assumption on e_t. Tanaka (1983, Econometrica) found an edgeworth expansion to approximate the finite sample distribution of $\widehat{\phi}_{ols}$. $E(\widehat{\phi}_{ols})$ does not have a closed-form expression. Kendall (1954, Biometrika) showed that $E(\widehat{\phi}_{ols}) \approx \phi - \frac{1+3\phi}{T}$
4. $\widehat{\phi}_{ols}$ is consistent. $\widehat{\phi}_{ols} = \widehat{\phi}_{mle}$. $\sqrt{n}\left(\widehat{\phi}_{mle} - \phi\right) \xrightarrow{d} N(0, 1-\phi^2)$ which can be used to test a hypothesis.



– Deriving the finite sample bias of $\widehat{\phi}_{ols}$ in the AR(1) model without intercept

1. From Equation (5.3.19) in Priestley (1981), we have

$$E(X_t X_{t+r} X_s X_{s+r+v}) = E(X_t X_{t+r})E(X_s X_{s+r+v}) + E(X_t X_s)E(X_{t+r} X_{s+r+v}) + E(X_t X_{s+r+v})E(X_{t+r} X_s).$$

2. $\sum\sum \phi^{|t-s|} = T\frac{1+\phi}{1-\phi} + \frac{2\phi^{T+1}-2\phi}{(1-\phi)^2}, \sum\sum \phi^{|t-s|+|t-s-1|} = T\frac{2\phi}{1-\phi^2} + \frac{(1+\phi^2)(\phi^{2T+1}-\phi)}{(1-\phi^2)^2}$

3. Let $\widehat{\phi}_{ols} = \frac{\frac{1}{T}\sum X_t X_{t-1}}{\frac{1}{T}\sum X_{t-1}^2} \equiv \frac{U_T}{V_T}$,with $U_T = \frac{1}{T}\sum X_t X_{t-1}$ and $V_T = \frac{1}{T}\sum X_{t-1}^2$.

4. Defining $Var(X_t)$ by σ_X^2, we have $Cov(X_t, X_{t+i}) = \phi^{|i|}\sigma_X^2, E(U_T) = \phi\sigma_X^2$ and $E(V_T) = \sigma_X^2$.

5. We also have

$$\begin{aligned} Cov(U_T, V_T) &= \frac{1}{T^2} \sum \sum Cov(X_t^2, X_s X_{s+1}) \\ &= \sigma_X^4 \frac{2}{T^2} \sum \sum \phi^{|t-s|+|t-s-1|} \\ &= \frac{2\sigma_X^4}{T^2} \left[T \frac{2\phi}{1-\phi^2} + \frac{(1+\phi^2)(\phi^{2T+1}-\phi)}{(1-\phi^2)^2} \right], \end{aligned} \tag{1}$$

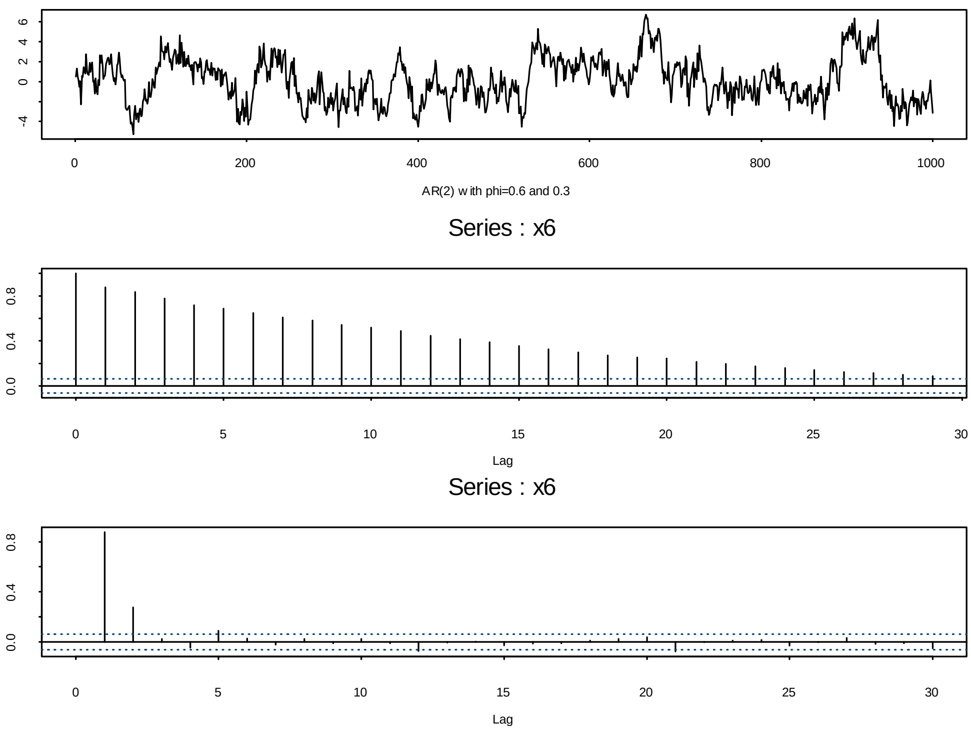
$$\begin{aligned} Var(V_T) &= \frac{1}{T^2} \sum \sum Cov(X_t^2, X_s^2) \\ &= \frac{2\sigma_X^4}{T^2} \sum \sum \phi^{2|t-s|} \\ &= \frac{2\sigma_X^4}{T^2} \left[T \frac{1+\phi^2}{1-\phi^2} + \frac{2\phi^{2T+2} - 2\phi^2}{(1-\phi^2)^2} \right], \end{aligned} \tag{2}$$

6. Taking the Taylor expansion to the first two terms, we have

$$E\left(\widehat{\phi}_{ols}\right) = E\left(\frac{U_T}{V_T}\right) = \frac{E(U_T)}{E(V_T)} - \frac{Cov(U_T, V_T)}{E^2(V_T)} + \frac{E(U_T)Var(V_T)}{E^3(V_T)} + o(T^{-1}) \tag{3}$$

$$= \phi - \frac{2\phi}{T} + o(T^{-1})$$

- AR(p): $X_t = \mu + \sum_{i=1}^{p} \phi_i X_{t-i} + e_t$, $e_t \sim iidN(0, \sigma_e^2)$
  - covariance stationary if the roots of $1 - \phi_1 z - \cdots - \phi_p z^p = 0$ lie outside the unit circle (ie the absolute values of the roots are greater than unity).
  - $E(X_t) = \frac{\mu}{1 - \sum_{i=1}^{p} \phi_i}$
  - $\alpha_k(k) = 0 \forall k > p$. PACF cuts off after the p^{th} lag
  - ACFs for the first p lags are more complicated than those in the AR(1) process. After the first p lags, however, the ACF geometrically decreases as the number of the lags increases
  - $\sum_{i=1}^{p} \phi_i = 1$ implies a unit root.



- Estimation

    1. OLS = conditional ML under Gaussianity.
    2. Conditiona ML or exact ML if we know the distribution function of e.
    3. $\widehat{\phi}$'s are consistent. In AR(2),

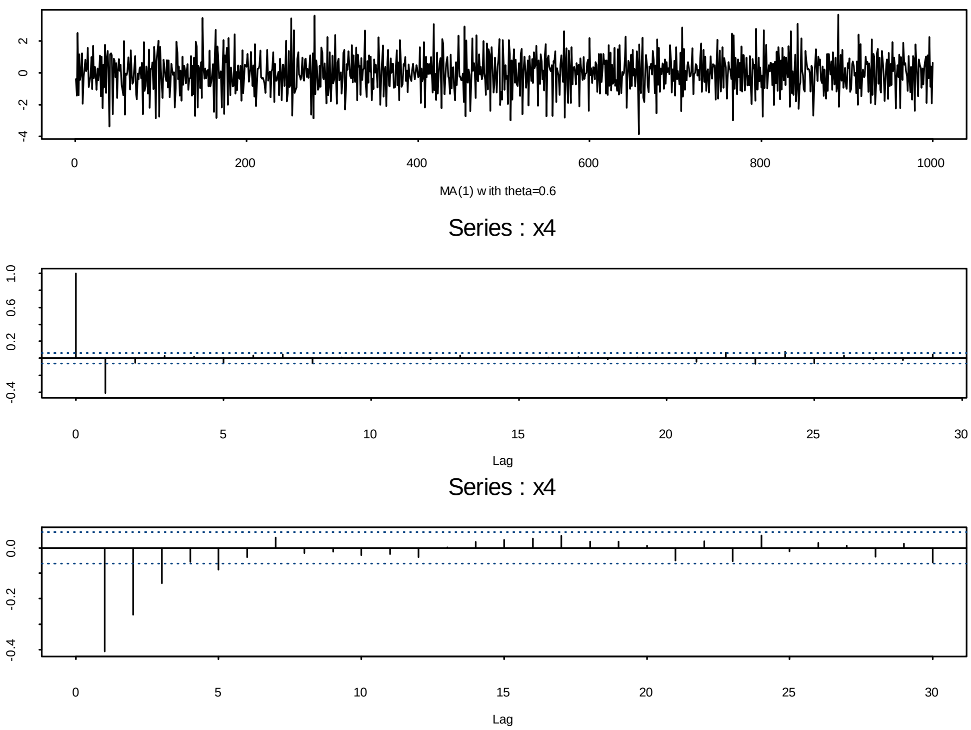
$$\sqrt{n}\left(\widehat{\phi}_{ols}-\phi\right)\xrightarrow{d}N\left[\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}1-\phi_2^2 & -\phi_1\left(1+\phi_2\right)\\ -\phi_1\left(1+\phi_2\right) & 1-\phi_2^2\end{pmatrix}\right]$$

    which can be used to test a hypothesis.

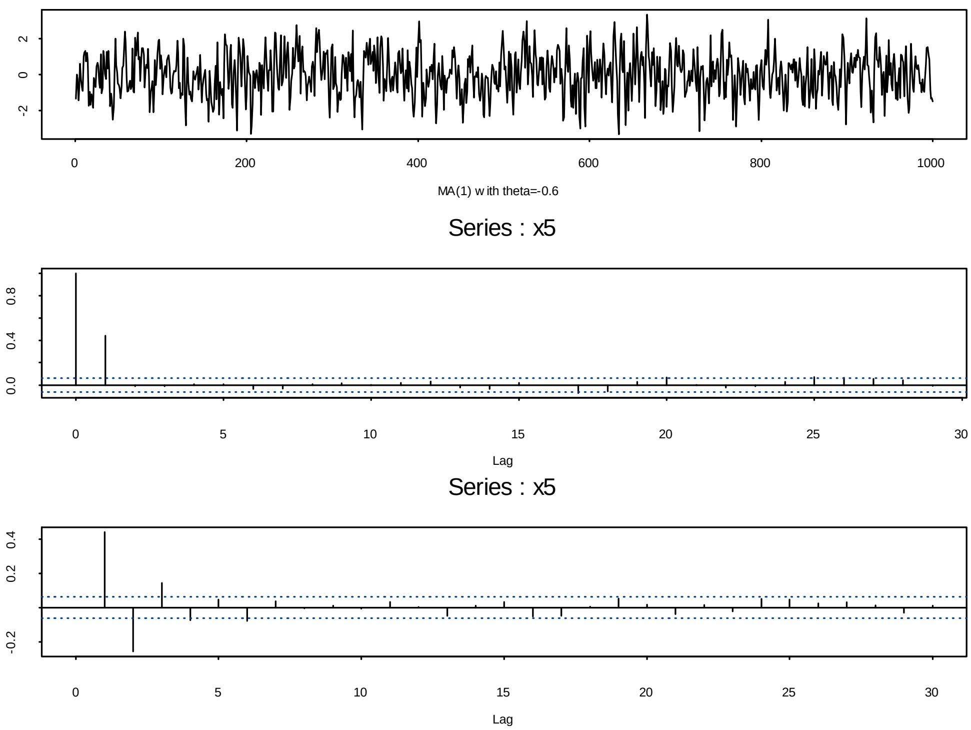
- MA(1): $X_t=\mu+e_t-\theta e_{t-1},\, e_t\sim iid(0,\sigma_e^2)$

    - Always covariance stationary$\forall\theta$
    - $E(X_t)=\mu, Var(X_t)=\sigma_e^2(1+\theta^2)$
    - $\rho_1=\frac{-\theta}{1+\theta^2},\rho_h=0\forall h>1$. ACF cuts off after the first lag.

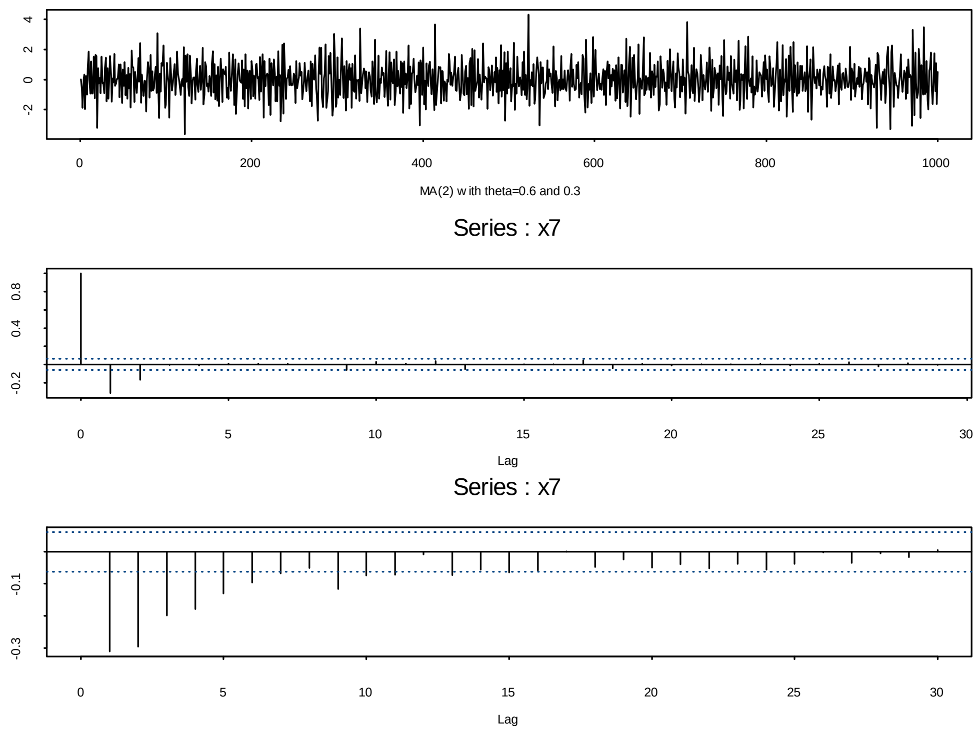
- ACF remains the same when θ becomes $1/\theta$.
- $\alpha_k(k) = \frac{(-1)^k\theta^k(\theta^2-1)}{1-\theta^{2k+2}}$. PACF does not cut off.
- If $|\theta| < 1$, MA(1) process is invertible. In this case PACF geometrically decreases with the number of the lags
- Relationship between MA(1) (with $|\theta| < 1$) and AR(∞)
- Estimation
  1. ML if we know the distribution function of e. $X_t|e_{t-1} \sim N(\mu - \theta e_{t-1}, \sigma_e^2)$.



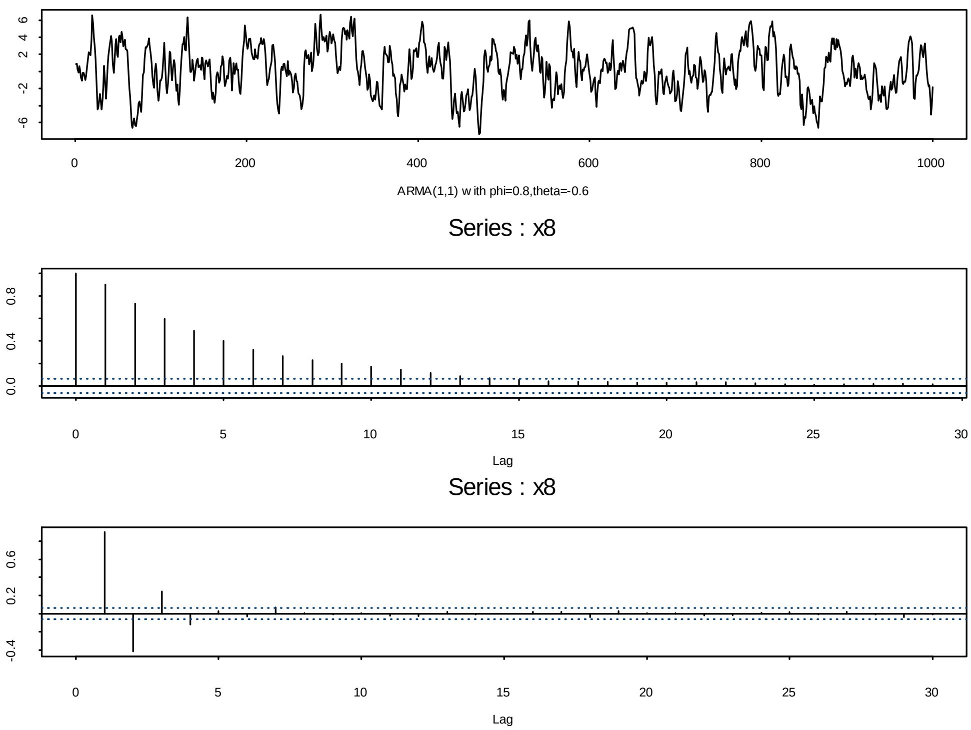
- MA(q): $X_t = \mu + e_t - \theta_1 e_{t-1} - \cdots - \theta_q e_{t-q}, e_t \sim iidN(0, \sigma_e^2)$
  - covariance stationary
  - The PACFs for the first q lags are more complicated than those in the MA(1) process. After the first q lags, however, the PACF geometrically decreases to 0 as the number of the lags increases
  - $\rho_j = 0 \ \ \forall j > q$. ACF will cut off after the q^{th} lag
  - Estimation
    1. ML if we know the distribution function of e.
    2. $\widehat{\theta}_{mle}$ is consistent. $\sqrt{n}\left(\widehat{\theta}_{mle} - \theta\right) \xrightarrow{d} N(0, 1-\theta^2)$ which can be used to test a hypothesis.



- MA(∞): $X_t = \mu + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots, e_t \sim iidN(0, \sigma_e^2)$
  - covariance stationary if $\sum_{j=0}^{\infty} |\theta_j| < \infty$
  - ergodic for the mean if $\sum_{j=0}^{\infty} |\theta_j| < \infty$
- Relationship between AR(1) and MA(∞)



- ARMA(1,1): $X_t = \mu + \phi X_{t-1} + e_t - \theta e_{t-1} \quad e_t \sim iidN(0, \sigma_e^2)$
  - $\rho_1 = \frac{(\phi-\theta)(1-\phi\theta)}{1-2\phi\theta+\theta^2}, \rho_h = \phi\rho(h-1), \forall h > 1$
  - The first ACF depends on the parameters of both the AR part and the MA part. After the first period, the subsequent ACF geometrically decreases to 0 with the rate of decline given by the AR parameter as the lag increases. This is similar to AR(1).
  - In contrast to AR(1), however, the PACF does not cut off. This is similar to MA(1).
  - Estimation
    1. ML if we know the distribution function of e.
- ARMA(0,0) – Gaussian white noise : $X_t = \mu + e_t \quad e_t \sim iidN(0, \sigma_e^2)$



- ARMA(p,q): $X_t = \sum_{i=1}^{p} \phi_i X_{t-i} + e_t - \sum_{i=1}^{q} \theta_i e_{t-i}$ $e_t \sim iidN(0, \sigma_e^2)$

  $\Longleftrightarrow \Phi(L)X_t = \Theta(L)e_t$

  $\Longleftrightarrow X_t = \frac{\Theta(L)}{\Phi(L)} e_t$ (MA representation) under some conditions

  $\Longleftrightarrow e_t = \frac{\Phi(L)}{\Theta(L)} X_t$ (AR representation) under some conditions

  - ACF behaves similar to that in AR(p) process
  - PACF behaves similar to that in MA(q) process
  - Estimation

    1. ML if we know the distribution function of e.

- Autocovariance generating function. When γ_j is absolutely summable, we define the **autocovariance generating function** as $g(z) = \sum_{j=-\infty}^{\infty} \gamma_j z^j$. The **population spectrum** is defined as $s(\varpi) = \frac{1}{2\pi} g(e^{-i\varpi})$.

- Wold Representation (Decomposition): For any zero mean covariance stationary process, X_t, we can always find $\{e_t\}$ and v_t such that

  $X_t = \sum_{j=0}^{\infty} c_j e_{t-j} + v$

  where $\{c_j\}$ is a real sequence with $\sum_{i=0}^{\infty} c_j^2 < \infty$ with $c_0 = 1$, $e_t \backsim WN(0, \sigma^2)$, v_t is purely deterministic, WN stands for "white noise".

4 Forecasting Based on ARMA Processes

- How to forecast a series with an ARMA(p,q) process with known coefficients?
- $\{X_1, ..., X_T\}$−historical observations
- X_{T+h}−(unknown) value of X in future period $T + h$
- $\widehat{X}_{T+h}$−forecast of X_{T+h} based on $\{X_1, ..., X_T\}$(h-period-ahead forecast)
- $\widehat{e}_{T+h} = X_{T+h} - \widehat{X}_{T+h}$−forecast error
- Mean squared error: $E(X_{T+h} - \widehat{X}_{T+h})^2$

- Minimum MSE forecast of X_{T+h} based on $\{X_1, ..., X_T\}$: $\widehat{X}_{T+h} = E(X_{T+h}|X_1, ..., X_T)$. This is known as the optimal forecasts

  Proof: Without loss of generality, assume $h = 1$. Let $g(X_1, ..., X_T)$ be the forecast other than the conditional mean. The MSE is

  $E[X_{T+1} - g(X_1, ..., X_T)]^2 = E[X_{T+1} - E(X_{T+1}|X_1, ..., X_T) + E(X_{T+1}|X_1, ..., X_T) - g(X_1, ..., X_T)]^2$

  $= E[X_{T+1} - E(X_{T+1}|X_1, ..., X_T)]^2$

  $+2E\{[X_{T+1} - E(X_{T+1}|X_1, ..., X_T)][E(X_{T+1}|X_1, ..., X_T) - g(X_1, ..., X_T)]\}$

  $+E[E(X_{T+1}|X_1, ..., X_T) - g(X_1, ..., X_T)]^2$

  Define $\eta_{T+1} = [X_{T+1} - E(X_{T+1}|X_1, ..., X_T)][E(X_{T+1}|X_1, ..., X_T) - g(X_1, ..., X_T)]$.

  Conditional on $X_1, ..., X_T$, both $E(X_{T+1}|X_1, ..., X_T)$ and $g(X_1, ..., X_T)$ are known constants. Hence $E(\eta_{T+1}|X_1, ..., X_T) = 0$.

  Therefore, $E(\eta_{T+1}) = E[E(\eta_{T+1}|X_1, ..., X_T)] = 0$ and

  $E[X_{T+1} - g(X_1, ..., X_T)]^2 = E[X_{T+1} - E(X_{T+1}|X_1, ..., X_T)]^2$

  $+E[E(X_{T+1}|X_1, ..., X_T) - g(X_1, ..., X_T)]^2$

  The second term cannot be smaller than zero. So $g(X_1, ..., X_T) = E(X_{T+1}|X_1, ..., X_T)$ will make the MSE the smallest.

- 1-step-ahead forecast: $h = 1$

- multi-step-ahead forecast: $h > 1$

4.0.1 Forecasting based on AR(1)

- Minimum Mean Square Error (MSE) forecast of X_{T+1} : $E(X_{T+1}|X_1 \ ,..., X_T) = \mu + \phi X_T + E(e_{T+1}|X_1, ..., X_T) = \mu + \phi X_T$
- Forecast error variance: σ_e^2
- Minimum MSE forecast of X_{T+2} :

  $E(X_{T+2}|X_1, ..., X_T) = \mu + \phi\mu + \phi^2 X_T$
- Minimum MSE forecast of X_{T+h} :

  $E(X_{T+h}|X_1, ..., X_T) = \mu + \phi\mu + ... + \phi^{h-1}\mu + \phi^h X_T$
- Forecast error variance: $\sigma_e^2(1 + \phi^2 + \phi^4 + ... + \phi^{2(h-1)})$
- The minimum MSE is known as the optimal forecasts

4.0.2 Forecasting based on MA(1)

- Minimum MSE (optimal) forecast of X_{T+1}:

$$E(X_{T+1}|X_1,\cdots,X_T) = \mu + E(e_{T+1}|X_1,\cdots,X_T) - \theta E(e_T|X_1,\cdots,X_T)$$
$$= \mu - \theta E(e_T|X_1,\cdots,X_T),$$

  where ε_T can be derived from $\{X_1, ..., X_T\}$ with some assumption about e_0. For example, we can assume $e_0 \approx E(e_0) = 0$. Such an approximation will have negligible effect on $\widehat{X}_{T+h}$ when the sample size is approaching infinite. However, when the sample size is small, the effect could be large and this leads to a non-optimal forecast in finite sample. The optimal forecast in finite sample can be obtained by Kalman filter.

- Minimum MSE (optimal) forecast of X_{T+2}:

$$E(X_{T+2}|X_1,\cdots,X_T) = \mu + E(e_{T+2}|X_1,\cdots,X_T) - \theta E(e_{T+1}|X_1,\cdots,X_T) = \mu$$

- Minimum MSE (optimal) forecast of X_{T+h} : $E(X_{T+h}|X_1\ , ..., X_T) = \mu, h \geq 2$

4.0.3 Forecasting based on ARMA(1,1)

- Minimum MSE (optimal) forecast of X_{T+1} : $E(X_{T+1}|X_1 \ , ..., X_T) = \mu + \phi X_T + E(e_{T+1}|X_1, ..., X_T) - \theta E(e_T|X_1, \ ..., X_T) = \mu + \phi X_T - \theta E(e_T|X_1, ..., X_T)$

- Minimum MSE (optimal) forecast of X_{T+h} :

  $E(X_{T+h}|X_1, \ ..., X_T) = \mu + \phi\mu + ... + \phi^{h-1}\mu + \phi^h X_T, h \geq 2$

5 Box-Jenkins Method

- This method applies to stationary ARMA time series.
- Procedures:

  1. When the data becomes stationary, identify plausible values for (p, q), say $(p_1, q_1), (p_2, q_2), ..., (p_k, q_k)$, by examining the sample ACF and sample PACF.
  2. Estimate these k ARMA processes
  3. Choose one model using model selection criteria, such as AIC, or SC (BIC)
  4. Validation: testing hypothesis, checking residuals
  5. Forecast with the selected model