统计分析与建模

高珍

gaozhen@tongji.edu.cn

平稳序列建模

- Regression and Autoregression
- Autoregressive (AR) models
- Moving average (MA) models
- Autoregressive moving average (ARMA) models;

回归模型

- Regressive model:

$$y_i = \alpha + \delta x_i + e_i \tag{1}$$

- Autoregressive (AR) model

$$y_t = \alpha + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + e_t \tag{2}$$

- Distributed Lag (DL) model

$$y_t = \alpha + \delta_0 x_t + \delta_1 x_{t-1} + \cdots + \delta_q x_{t-q} + e_t \tag{3}$$

- Autoregressive Distributed Lag (ADL) model:

$$y_t = \alpha + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \delta_0 x_t + \delta_1 x_{t-1} + \cdots + \delta_q x_{t-q} + e_t \tag{4}$$

Outline

- $AR(p)$ model
- $MA(q)$ model
- $ARMA(p,q)$ model

Autoregressive AR(1)模型

- $y_t = \phi_0 + \phi_1 y_{t-1} + e_t$, where $e_t \sim i.i.d.(0, \sigma^2)$.
- A compact form: $(1 - \phi_1 B) y_t = \phi_0 + e_t$.
- B: back-shift operator:

$$
\begin{aligned}
By_t &= y_{t-1}, \\
B^k y_t &= y_{t-k}, \\
B^0 y_t &= y_t, \\
B^k c &= c.
\end{aligned}
$$

- Characteristic equation: $\phi(x) = 1 - \phi_1 x = 0$.
- Stationary condition: the root of $1 - \phi_1 x = 0$ is outside of the unit circle, then AR(1) is stationary iff $|\phi_1| < 1$.

AR(1)统计特征

Given the stationarity of y_t : $\quad y_t = \phi_0 + \phi_1 y_{t-1} + e_t$ (1)

- Mean: $\mu = Ey_t = \dfrac{\phi_0}{1-\phi_1}$, (2)
- Variance:

$$
\begin{aligned}
E[(y_t-\mu)^2] &= \phi_1^2 E[(y_{t-1}-\mu)^2] + \sigma^2, &(3)\\
\gamma_0 &= \phi_1^2\gamma_0 + \sigma^2 \Rightarrow \gamma_0 = \frac{\sigma^2}{1-\phi_1^2}. &(4)
\end{aligned}
$$

- The autocovariance function (ACVF) at lag j:

$$
\begin{aligned}
\gamma_j &= Cov(y_{t-j}, y_t) &(5)\\
&= \phi_1\gamma_{j-1} = \frac{\phi_1^j\sigma^2}{1-\phi_1^2}, \text{ for all } j \geq 1.
\end{aligned}
$$

- The autocorrelation function (ACF) at lag j:

$$\rho_j = \frac{\gamma_j}{\gamma_0} = \phi_1^j. \tag{6}$$

ACF of AR(1)

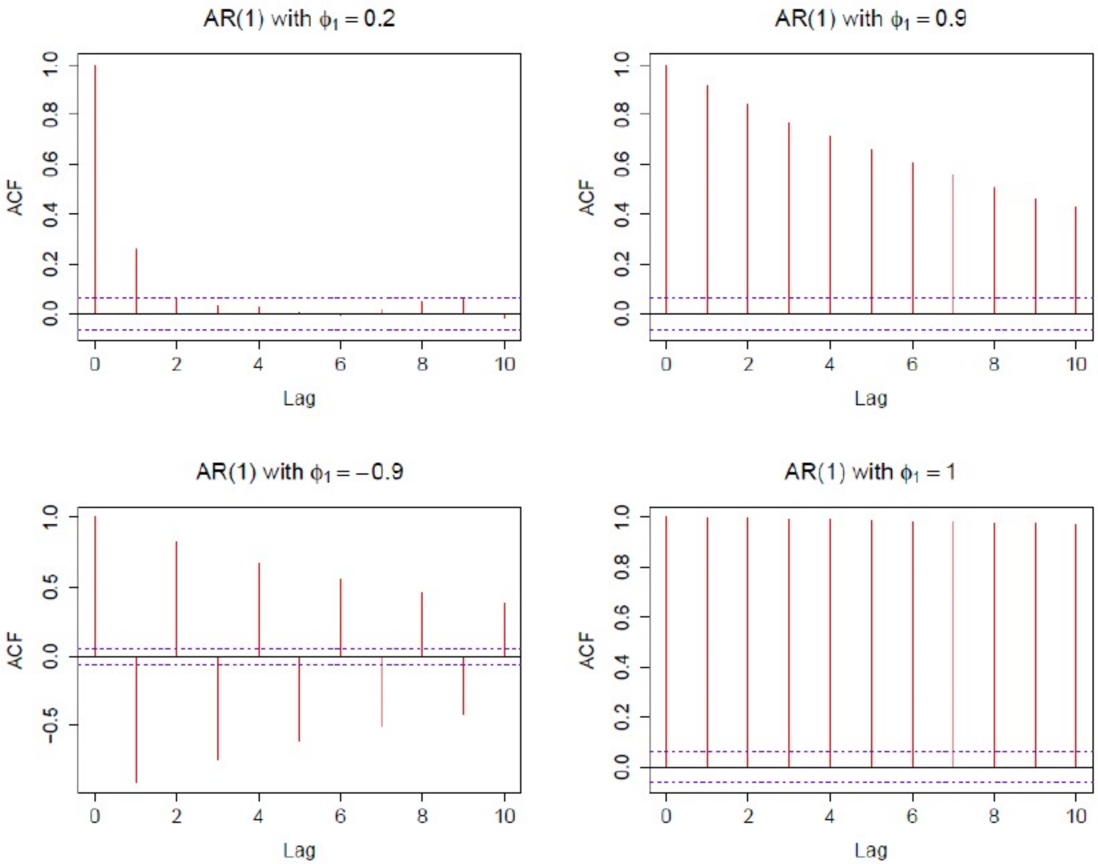


Figure 1: ACF of AR(1) processes

举例: 平稳

$$y_t = 0.5y_{t-1} + e_t$$

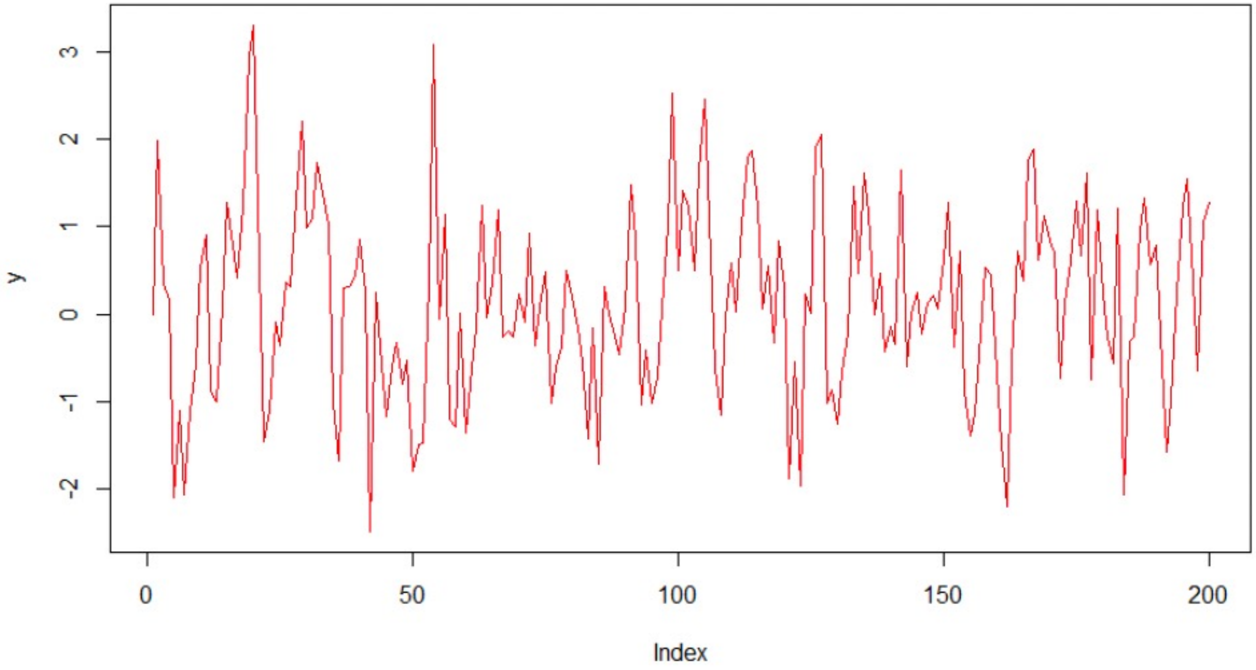
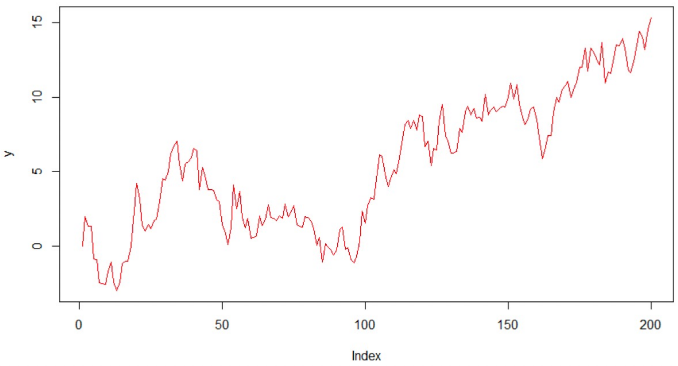


Figure 2: Stationary AR(1)

举例: 非平稳

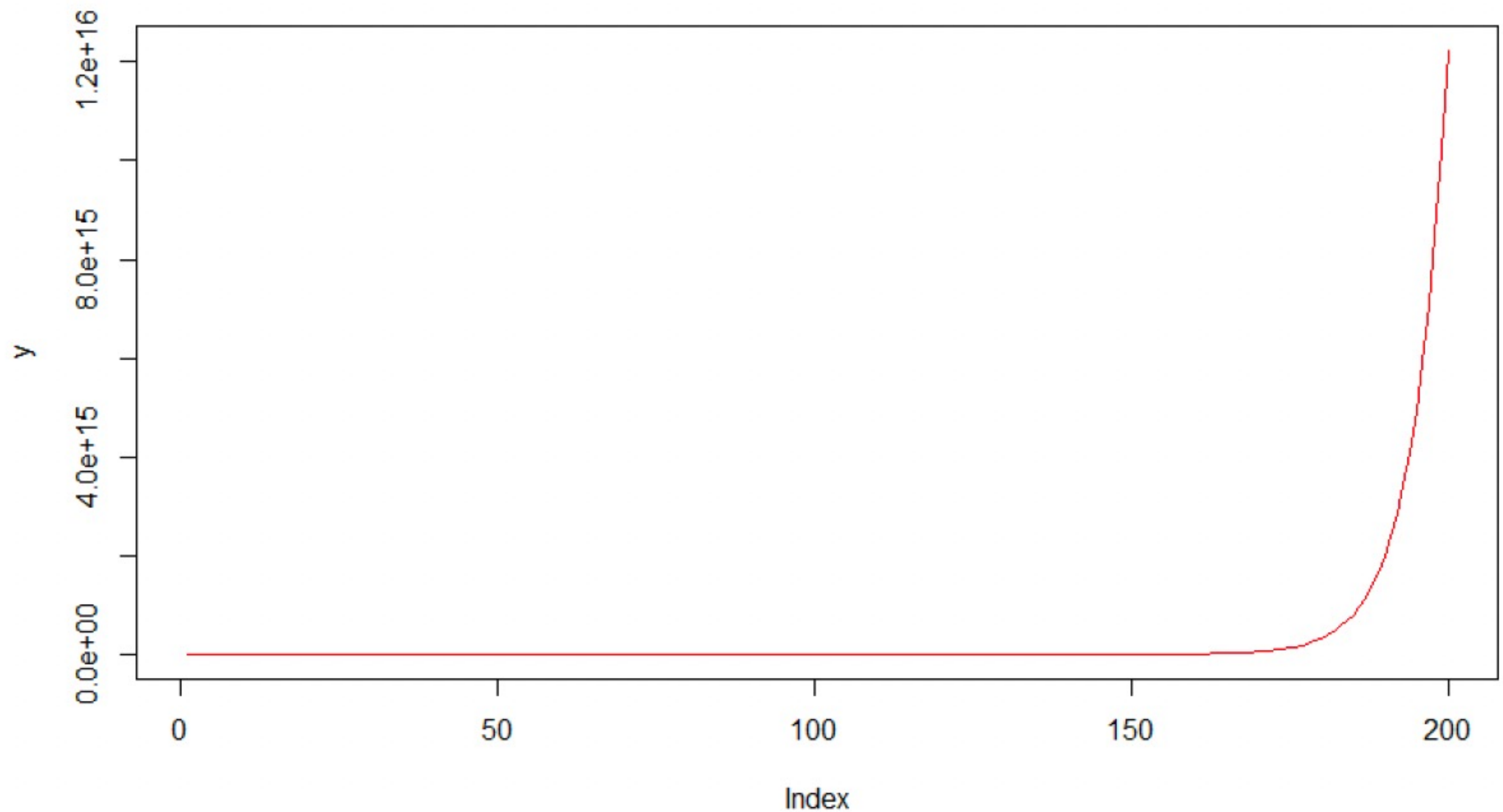


Random Walk

- Case 2: Non-stationary.
  $\phi_1 = 1$, it is not covariance stationary.
  - Assuming $\phi_0 = 0$, AR(1) model can be rewritten as the random walk process:
  $$y_t = e_t + e_{t-1} + \cdots + e_1 + y_0.$$
  - $\gamma_{0,t} \equiv E(y_t - y_0)^2 = t\sigma^2$
    $\gamma_{j,t} \equiv E(y_t - y_0)(y_{t-j} - r_0) = (t-j)\sigma^2$ and
    $\rho_{j,t} = (t-j)/t$ for all $k \geq 0$.
- Similarly, when $\phi_1 = -1$, it is not covariance stationary.

举例: 非平稳

Explosive AR(1) $y_t = 1.2 * y_{t-1} + e_t$



AR(2)

- $y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + e_t$, where $e_t \sim i.i.d.(0, \sigma^2)$.
- $(1 - \phi_1 B - \phi_2 B^2) y_t = \phi_0 + e_t$.
- Characteristic equation: $\phi(x) = 1 - \phi_1 x - \phi_2 x^2 = 0$.
- Stationary condition: the roots of $1 - \phi_1 x - \phi_2 x^2 = 0$ are outside of the unit circle.
- Stationarity condition for AR(2) model is equivalent to

$$\begin{cases} \phi_2 + \phi_1 < 1, \\ \phi_2 - \phi_1 < 1, \\ -1 < \phi_2 < 1. \end{cases}$$

AR(2)统计特征

Given y_t is a stationary AR(2) process

- Mean: $E(y_t) = \phi_0 + \phi_1 E(y_{t-1}) + \phi_2 E(y_{t-2}) + E(e_t),$ (1)

$$\mu = \frac{\phi_0}{1 - \phi_1 - \phi_2} \quad (2)$$

- $$y_t - \mu = \phi_1(y_{t-1} - \mu) + \phi_2(y_{t-2} - \mu) + e_t. \quad (3)$$

- ACVF: Multiplying $(Y_{t-j} - \mu)$ on both sides and taking expectation,

$$\gamma_j = \phi_1 \gamma_{j-1} + \phi_2 \gamma_{j-2}, \text{ for all } j = 1, 2, 3, ... \quad (4)$$

  - $j = 1,$ (5)

$$\gamma_1 = \phi_1 \gamma_0 + \phi_2 \gamma_{-1} = \phi_1 \gamma_0 + \phi_2 \gamma_1. \quad \boxed{\rho_1 = \frac{\gamma_1}{\gamma_0} = \frac{\phi_1}{1 - \phi_2};}$$

  - $j = 2,$ (6)

$$\gamma_2 = \phi_1 \gamma_1 + \phi_2 \gamma_0; \quad \boxed{\rho_j = \phi_1 \rho_{j-1} + \phi_2 \rho_{j-2}}$$

AR(2)统计特征

Variance:

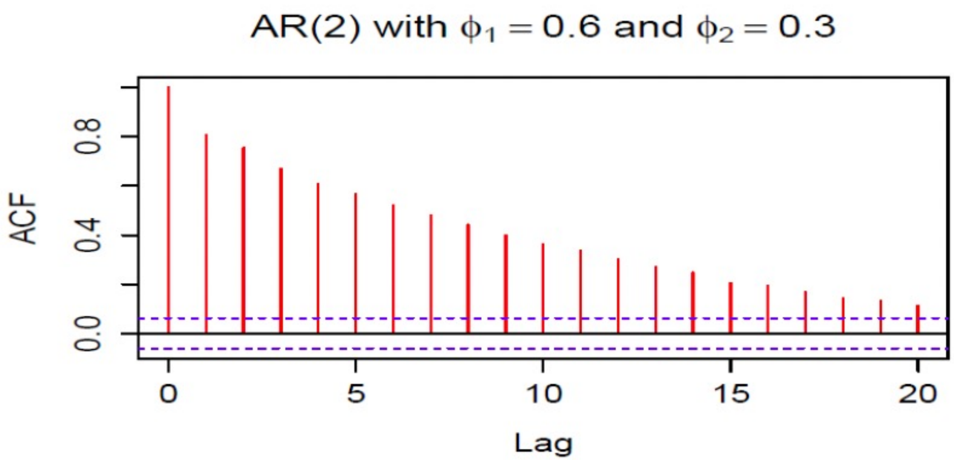
$$\begin{aligned}\gamma_0 &= E(y_t-\mu)^2 \\ &= \phi_1\gamma_1+\phi_2\gamma_2+\sigma^2 \\ &= \phi_1\rho_1\gamma_0+\phi_2\rho_2\gamma_0+\sigma^2; \\ &= \frac{(1-\phi_2)\sigma^2}{(1+\phi_2)((1-\phi_2)^2-\phi_1^2)}.\end{aligned}$$

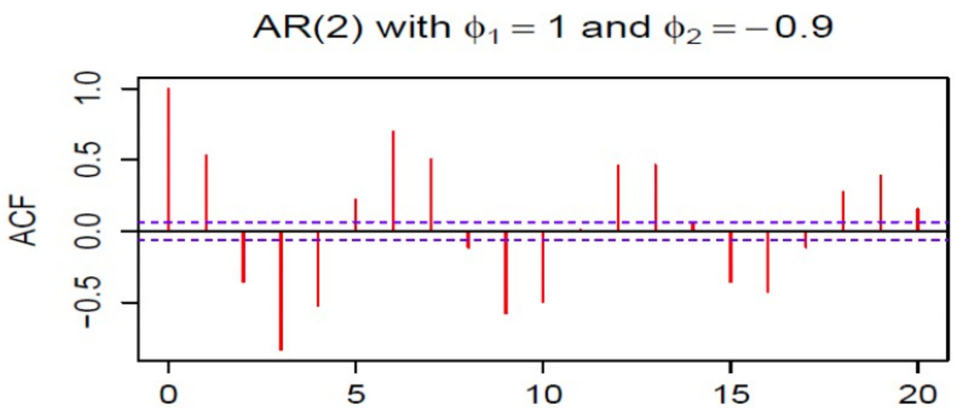
ACF:

$$\begin{aligned}\rho_1 &= \frac{\gamma_1}{\gamma_0}=\frac{\phi_1}{1-\phi_2}; \\ \rho_2 &= \phi_1\rho_1+\phi_2=\frac{\phi_1^2}{1-\phi_2}+\phi_2; \\ \rho_j &= \phi_1\rho_{j-1}+\phi_2\rho_{j-2}, \quad \text{for} \quad j\geq 3.\end{aligned}$$

举例：平稳

$$
\begin{aligned}
\rho_1 &= \frac{\gamma_1}{\gamma_0} = \frac{\phi_1}{1-\phi_2}; \\
\rho_2 &= \phi_1\rho_1 + \phi_2 = \frac{\phi_1^2}{1-\phi_2} + \phi_2; \\
\rho_j &= \phi_1\rho_{j-1} + \phi_2\rho_{j-2}, \quad \text{for} \quad j \geq 3.
\end{aligned}
$$





AR(p)

- $y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + e_t$
- $\phi(B) y_t = \phi_0 + e_t$, where $\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$.
- Condition for Stationarity: the roots of $\phi_p(z) = 0$ lie outside the unit circle.
- ACF of AR(p) model:

$$
\begin{aligned}
\gamma_k &= \phi_1 \gamma_{k-1} + \cdots + \phi_p \gamma_{k-p}, \quad k > 0. \\
\rho_k &= \phi_1 \rho_{k-1} + \cdots + \phi_p \rho_{k-p}, \quad k > 0.
\end{aligned}
$$

(the so-called Yule-Walker equations of ρ_k.)

AR(p)

- Solve the following sets of equations to obtain $\rho_1, \cdots, \rho_p$.

$$\begin{cases} \rho_1 - \phi_1\rho_0 - \cdots - \phi_p\rho_{p-1} = 0, \\ \cdots\cdots \\ \rho_p - \phi_1\rho_{p-1} - \cdots - \phi_p\rho_0 = 0. \end{cases}$$

- Using the following recursive equations to get calculate ρ_k when $k \geq p+1$.

$$\rho_k = \phi_1\rho_{k-1} + \cdots + \phi_p\rho_{k-p}.$$

PACF 偏自相关函数

- The PACF at lag k (denoted by π_k) measures the correlation between y_t and y_{t-k} regardless of their linear relationship with the intermediate variables $\{y_{t-1}, \cdots, y_{t-k+1}\}$.
- If y_t is a normally distributed times series, then

$$\pi_k = Corr(y_t, y_{t-k} | y_{t-1}, \cdots, y_{t-k+1})$$

- This definition is equivalent to say that

$$\begin{aligned}\pi_k &= Corr(y_t - E(y_t | y_{t+1}, \ldots, y_{t+k-1}), y_{t+k} - E(y_{t+k} | y_{t+1}, \ldots, y_{t+k-1}))\\ &= Corr(y_t - \hat{y}_t, y_{t+k} - \hat{y}_{t+k}).\end{aligned}$$

- On linear regression theory, $\hat{y}_t$ and $\hat{y}_{t+k}$ are the best linear estimates of y_t and y_{t+k} (respectively) based on the values of $y_{t+1}, \ldots, y_{t+k-1}$.

PACF

- According to the above definitions, the partial autocorrelation coefficient of order k is computed as the least squares estimator of the coefficient ϕ_{kk} in

$$y_t = \phi_{k1} y_{t-1} + \cdots + \underbrace{\phi_{kk}}_{\pi_k} y_{t-k} + e_t \tag{5}$$

  where y_t is assumed to be zero mean.

- We will only include a further lagged variable y_{t-k} in the model for y_t if y_{t-k} makes a genuine and additional contribution to y_t in addition to those from $y_{t-1}, \cdots, y_{t-k+1}$.

ACF & PACF of AR models

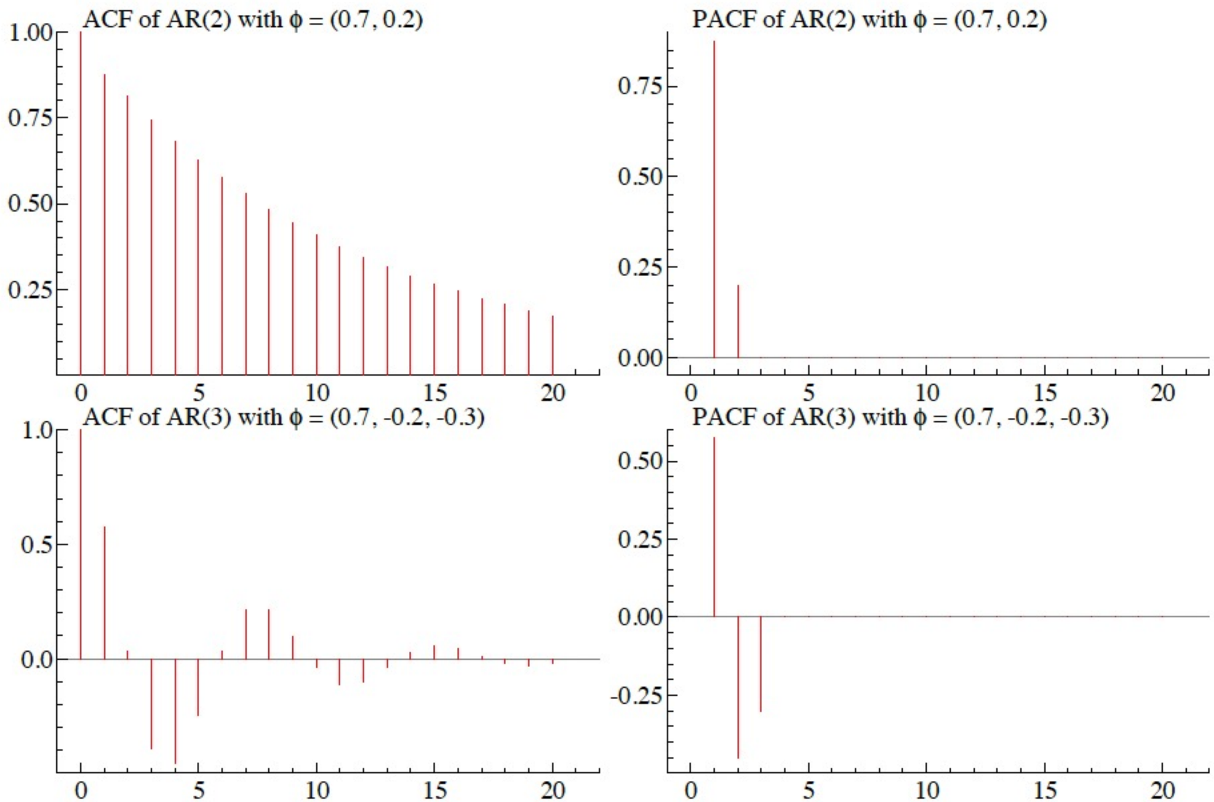


Figure 7: The PACF of AR(p) Models

Outline

- $AR(p)$ model
- $MA(q)$ model
- $ARMA(p, q)$ model

Moving average(MA) model

- MA(q):

$$y_t = \theta_0 + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q},$$

  where $q \geq 0$ is a finite integer and $\{e_t\} \sim i.i.d.(0, \sigma_e^2)$.
- MA(0) is actually a i.i.d. sequence if $\theta_0 = 0$.
- First proposed by E. Slutsky in 1927 to explain some cycle phenomena in economic data etc.
- In some textbooks, they use the following definition equation:

$$y_t = \theta_0 + e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_q e_{t-q},$$

$MA(1)$统计特征

- $y_t = \theta_0 + e_t - \theta_1 e_{t-1}$.
- $\mu = E(y_t) = \theta_0$
- Variance $\gamma_0 = Var(y_t) = \sigma_e^2(1 + \theta_1^2)$.
- ACVF at lag 1 is

$$\begin{aligned}\gamma_1 &= Cov(y_t, y_{t-1}) \\ &= Cov(e_t - \theta_1 e_{t-1}, e_{t-1} - \theta_2 e_{t-2}) \\ &= Cov(-\theta_1 e_{t-1}, e_{t-1}) = -\theta_1 \sigma_e^2,\end{aligned}$$

- ACF at lag 1 is

$$\rho_1 = \frac{-\theta_1}{1 + \theta_1^2}.$$

$MA(1)$统计特征

$$\begin{cases} E(y_t) = \theta_0, \\ \gamma_0 = \sigma_e^2(1+\theta_1^2), \\ \rho_1 = \dfrac{-\theta_1}{1+\theta_1^2}, \\ \rho_k = 0, \quad \forall\ k \geq 2. \end{cases}$$

- The ACVF at lag 2 is $\boxed{y_t = \theta_0 + e_t - \theta_1 e_{t-1}}$

$$\begin{aligned} \gamma_2 &= Cov(y_t, y_{t-2}) \\ &= Cov(e_t - \theta_1 e_{t-1}, e_{t-2} - \theta_1 e_{t-3}) \\ &= 0, \end{aligned}$$

- Similarly, $\gamma_k = Cov(y_t, y_{t-k}) = 0$, and $\rho_k = 0$, whenever $k \geq 2$;
- That is, the process has no correlation beyond lag 1.

ACF of $MA(1)$

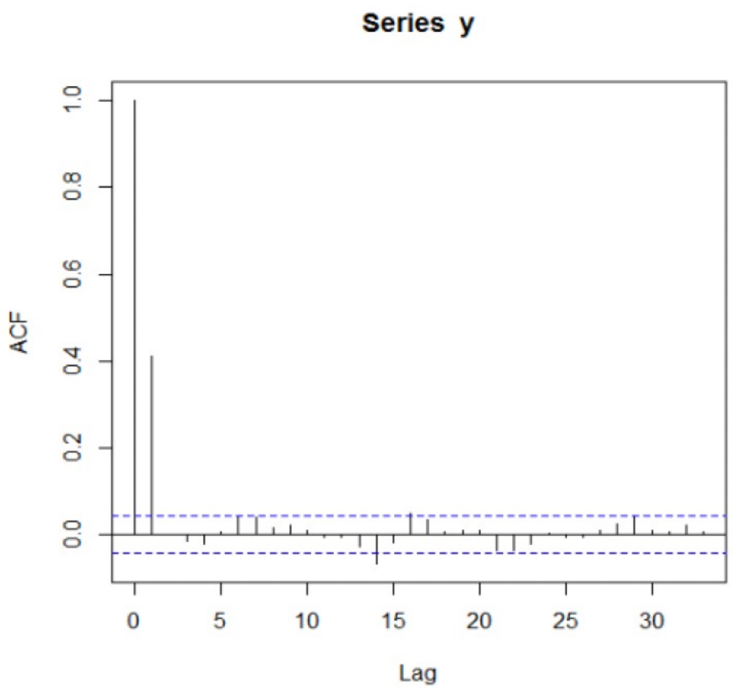


Figure 8: The ACFs of MA(1) model cuts off from lag 2.

$$\begin{cases} E(y_t) = \theta_0, \\ \gamma_0 = \sigma_e^2(1+\theta_1^2), \\ \rho_1 = \dfrac{-\theta_1}{1+\theta_1^2}, \\ \rho_k = 0, \quad \forall\ k \geq 2. \end{cases}$$

$MA(2)$

Consider a MA(2) process,

$$y_t = \theta_0 + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2}.$$

- Mean: $\mu = E(y_t) = \theta_0$.
- Variance:

$$\begin{aligned} \gamma_0 &= Var(y_t) = Var(e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2}) \\ &= (1 + \theta_1^2 + \theta_2^2)\sigma_e^2, \end{aligned}$$

$MA(2)$统计特征

$$y_t = \theta_0 + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2}$$

- ACVF at lag 1:

$$\begin{aligned}
\gamma_1 &= Cov(y_t, y_{t-1}) \\
&= Cov(e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2}, e_{t-1} - \theta_1 e_{t-2} - \theta_2 e_{t-3}) \\
&= (-\theta_1 + \theta_1\theta_2)\sigma_e^2,
\end{aligned}$$

- ACF at lag 2:

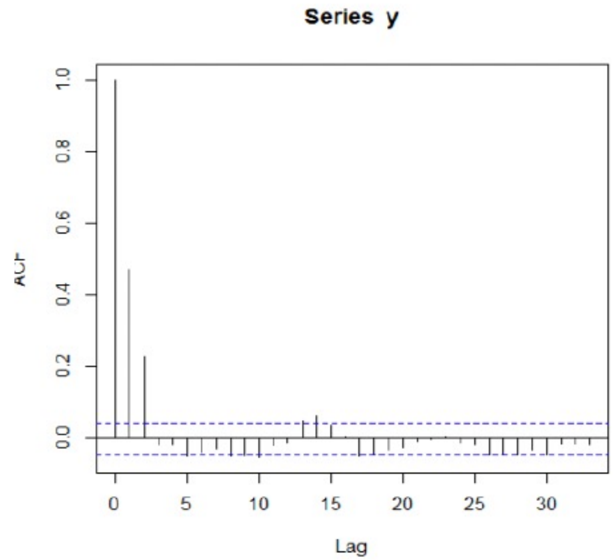
$$\begin{aligned}
\gamma_2 &= Cov(y_t, y_{t-2}) \\
&= Cov(e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2}, e_{t-2} - \theta_1 e_{t-3} - \theta_2 e_{t-4}) \\
&= Cov(-\theta_2 e_{t-2}, e_{t-2}) = -\theta_2\sigma_e^2.
\end{aligned}$$

- $\gamma_k = 0$, for all $k \geq 3$.

ACF $of\ MA(2)$

The ACF of the MA(2) model is

$$\begin{cases} \rho_1 = \dfrac{-\theta_1 + \theta_1\theta_2}{1 + \theta_1^2 + \theta_2^2}, \\ \rho_2 = \dfrac{-\theta_2}{1 + \theta_1^2 + \theta_2^2}, \\ \rho_k = 0, \quad \forall k \geq 3. \end{cases}$$



$MA(q)$

- $y_t = \theta_0 + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \ldots - \theta_q e_{t-q}, e_t \sim i.i.d.(0, \sigma^2)$.
- Lag form: $y_t = \theta_0 + (1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q)e_t$.
- MA(q) model is always weakly stationary. (Why?)
- Mean: $\mu = E(y_t) = \theta_0$;
- Variance:$\gamma_0 = (1 + \theta_1^2 + \cdots + \theta_q^2)\sigma^2$;
- ACF at lag j:

$$\rho_j = \frac{\theta_j + \theta_{j+1}\theta_1 + \theta_{j+2}\theta_2 + \ldots + \theta_q\theta_{q-j}}{1 + \theta_1^2 + \ldots + \theta_q^2}, \text{for } j = 1, 2, \ldots, q;$$

  and $\rho_j = 0$ for $j > q$.
- Invertibility: All roots of $1 - \theta_1 x \theta_2 x^2 - \cdots \theta_q x^q = 0$ lie out of unit circle.

Outline

- $AR(p)$ model
- $MA(q)$ model
- $ARMA(p,q)$ model

Autoregressive moving average(ARMA) model

ARMA(1,1):

- $y_t = \phi_1 y_{t-1} + \phi_0 + e_t - \theta_1 e_{t-1}$.
  or Lag form: $(1 - \phi_1 B) y_t = \phi_0 + (1 - \theta_1 B) e_t$, where $e_t \sim i.i.d.(0, \sigma_e^2)$.
- Stationary condition: same as AR(1)
- Invertible condition: same as MA(1)
- Mean: $\mu = E(y_t) = \frac{\phi_0}{1-\phi_1}$ (same as AR(1))
- Variance: $\gamma_0 = Var(y_t) = \dfrac{(1 - 2\phi_1\theta_1 + \theta_1^2)\sigma_e^2}{1 - \phi_1^2}$
- ACF: $\rho_k = \phi_1 \rho_{k-1}$ for $k > 1$ and $\rho_1 = \phi_1 - \dfrac{\theta_1 \sigma_e^2}{\gamma_0}$

$ARMA(p,q)$

- A general ARMA(p,q) model is in the form:

$$\begin{aligned} y_t &= \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} \\ &\quad + e_t - \theta_1 e_{t-1} - \cdots - \theta_q e_{t-q}. \end{aligned}$$

- Lag form:

$$\phi(B) y_t = \phi_0 + \theta(B) e_t,$$

  where $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$ and $\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$.
- It is assumed that the polynomials $\phi(B)$ and $\theta(B)$ can not have common factors.
- Again, for a stationary process, we can rewrite the model as

$$\phi(B)(y_t - \mu) = \theta(B) e_t.$$

$ARMA(p,q)$统计特征

Given stationarity:

- Mean: $\mu = \dfrac{\phi_0}{1 - \phi_1 - \cdots - \phi_p}$.
- ACF: the correlation coefficient ρ_j satisfies that

$$\rho_j - \phi_1 \rho_{j-1} - \cdots - \phi_p \rho_{j-p} = 0, \quad \text{for } j > q,$$

then the ACF satisfy the difference equation $\phi(B)\rho_j = 0$ for $j > q$ with $\rho_1, \ldots, \rho_q$ as initial conditions.

谢谢！

高珍
gaozhen@tongji.edu.cn