时间序列分析（TIME SERIES ANALYSIS）

主讲：吴尚
复旦大学管理学院统计与数据科学系

模型识别

- 样本自相关函数和MA模型的识别
- 偏自相关函数（PACF）和AR模型的识别
- EACF和ARMA模型的识别
- 非平稳模型识别

样本自相关函数

$$r_k = \frac{\sum_{t=k+1}^{n}(Y_t - \bar{Y})(Y_{t-k} - \bar{Y})}{\sum_{t=1}^{n}(Y_t - \bar{Y})^2}, k = 1,2,\ldots$$

- 对于任何固定的m，当n趋于无穷时，$\sqrt{n}(r_1 - \rho_1), \sqrt{n}(r_2 - \rho_2), \ldots, \sqrt{n}(r_m - \rho_m)$的联合分布逼近均值为0，协方差矩阵为$C = (c_{ij})$的多元正态分布（6.1.2）
- 对于白噪声，$Var(r_k) \approx \frac{1}{n}, Corr(r_k, r_j) \approx 0, k \neq j$
- 对于AR(1)，$Var(r_1) \approx \frac{1-\phi^2}{n}$，$Var(r_k) \approx \frac{1+\phi^2}{n(1-\phi^2)}$对于较大的$k$
- 对于MA(1)，$Var(r_1) \approx \frac{1-3\rho_1^2+4\rho_1^4}{n}$，$Var(r_k) \approx \frac{1+2\rho_1^2}{n}$

MA(q)的样本自相关函数

- 对于MA(q)，当$k > q$时，

$$c_{kk} = 1 + 2\sum_{j=1}^{q}\rho_j^2$$

$$\sqrt{n}(r_k - \rho_k(= 0)) \to N(0, c_{kk})$$

$$Var(r_k) \approx \frac{c_{kk}}{n} = \frac{1}{n}\left[1 + 2\sum_{j=1}^{q}\rho_j^2\right]$$

- 因此

$$P\left(-2\sqrt{\frac{c_{kk}}{n}} \le r_k \le 2\sqrt{\frac{c_{kk}}{n}}\right) \approx 0.95$$

MA模型识别

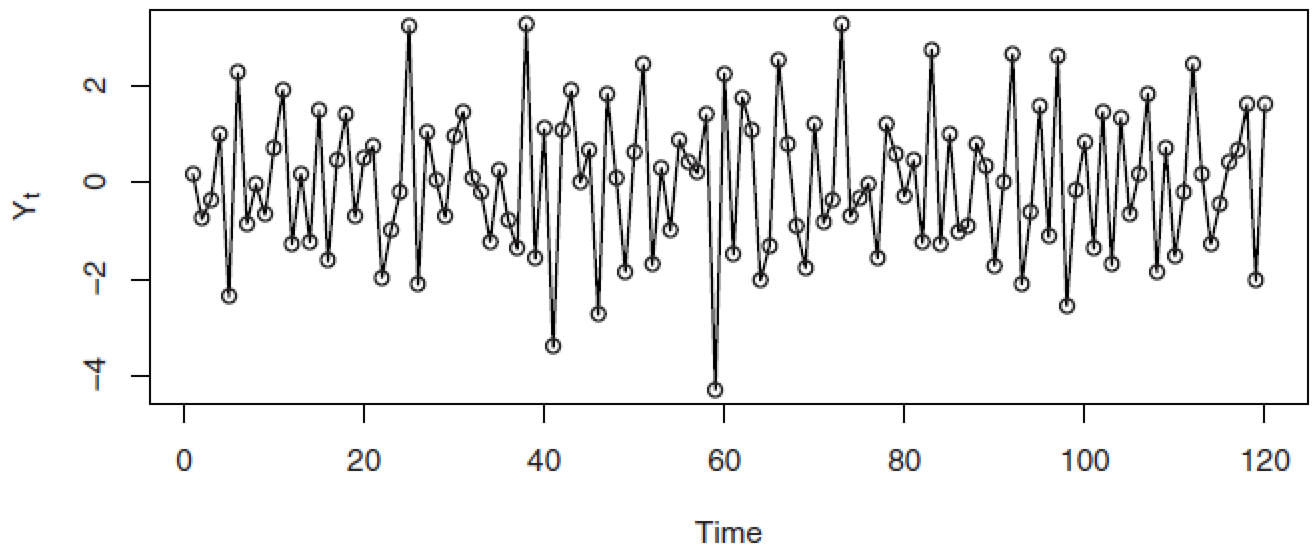
- 当真实模型为MA(q)时，对于每个$k > q$，r_k在正负2倍的标准差$\pm 2\hat{s}_q$之间的概率约为0.95，其中标准差可以用下式估计

$$\hat{s}_q = \sqrt{\frac{1}{n}\left[1 + 2\sum_{j=1}^{q} r_j^2\right]}$$

- 如果样本自相关系数在最初的q阶明显大于2倍标准差范围，而后几乎95%的自相关系数都落在2倍标准差的范围以内，而且由非零自相关系数衰减为小值波动的过程非常突然，这时通常视为自相关系数截尾，阶数为q.

MA(1)模型识别范例

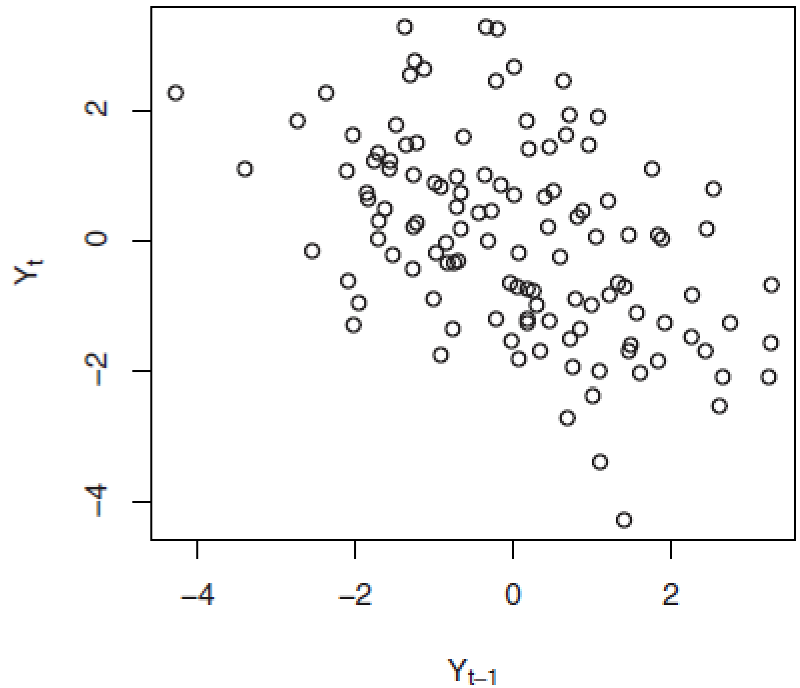
Exhibit 4.5 **Time Plot of an MA(1) Process with $\theta = +0.9$**




```
> win.graph(width=4.875,height=3,pointsize=8)
> data(ma1.1.s)
> plot(ma1.1.s,ylab=expression(Y[t]),type='o')
```

MA(1)模型识别范例

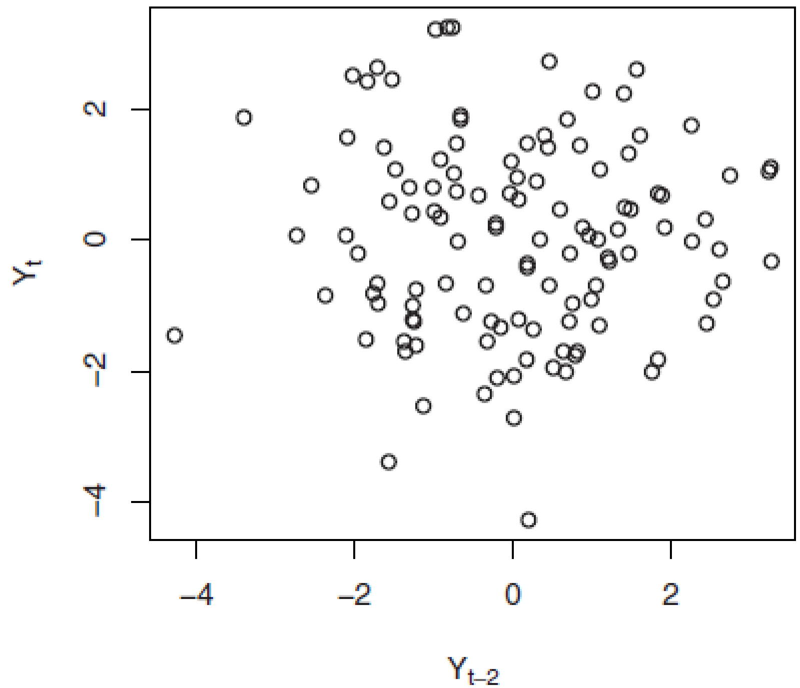
Exhibit 4.6 Plot of Y_t versus Y_{t-1} for MA(1) Series in Exhibit 4.5




```
> win.graph(width=3, height=3,pointsize=8)
> plot(y=ma1.1.s,x=zlag(ma1.1.s),ylab=expression(Y[t]),
   xlab=expression(Y[t-1]),type='p')
```

MA(1)模型识别范例

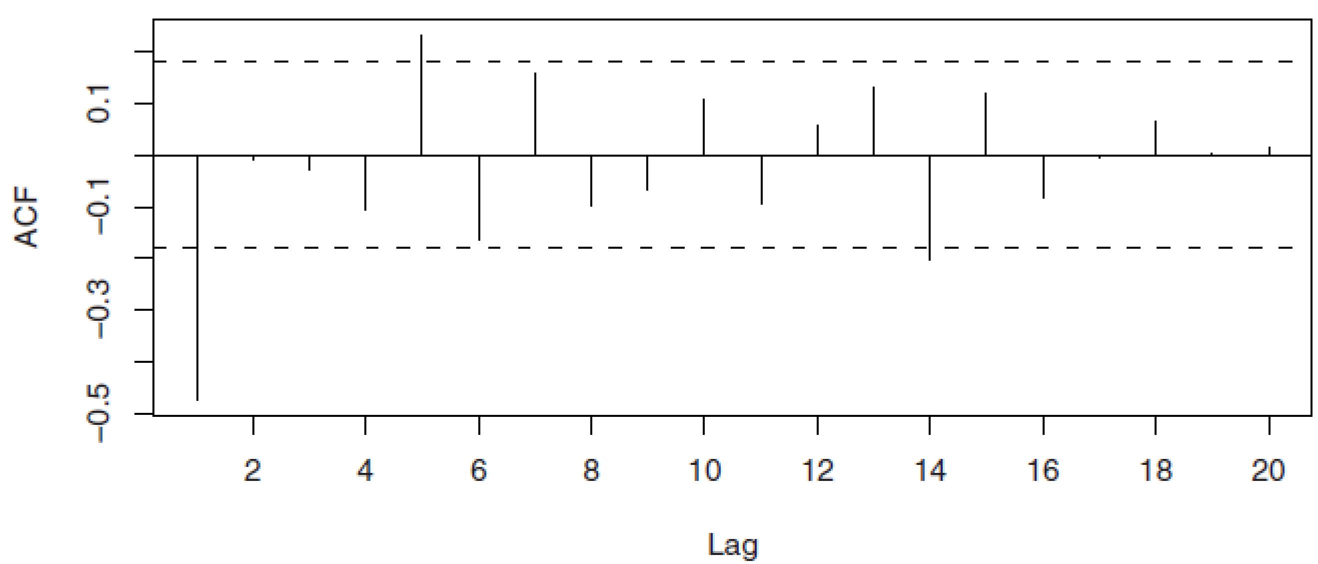
Exhibit 4.7 Plot of Y_t versus Y_{t-2} for MA(1) Series in Exhibit 4.5




```
> plot(y=ma1.1.s,x=zlag(ma1.1.s,2),ylab=expression(Y[t]),
   xlab=expression(Y[t-2]),type='p')
```

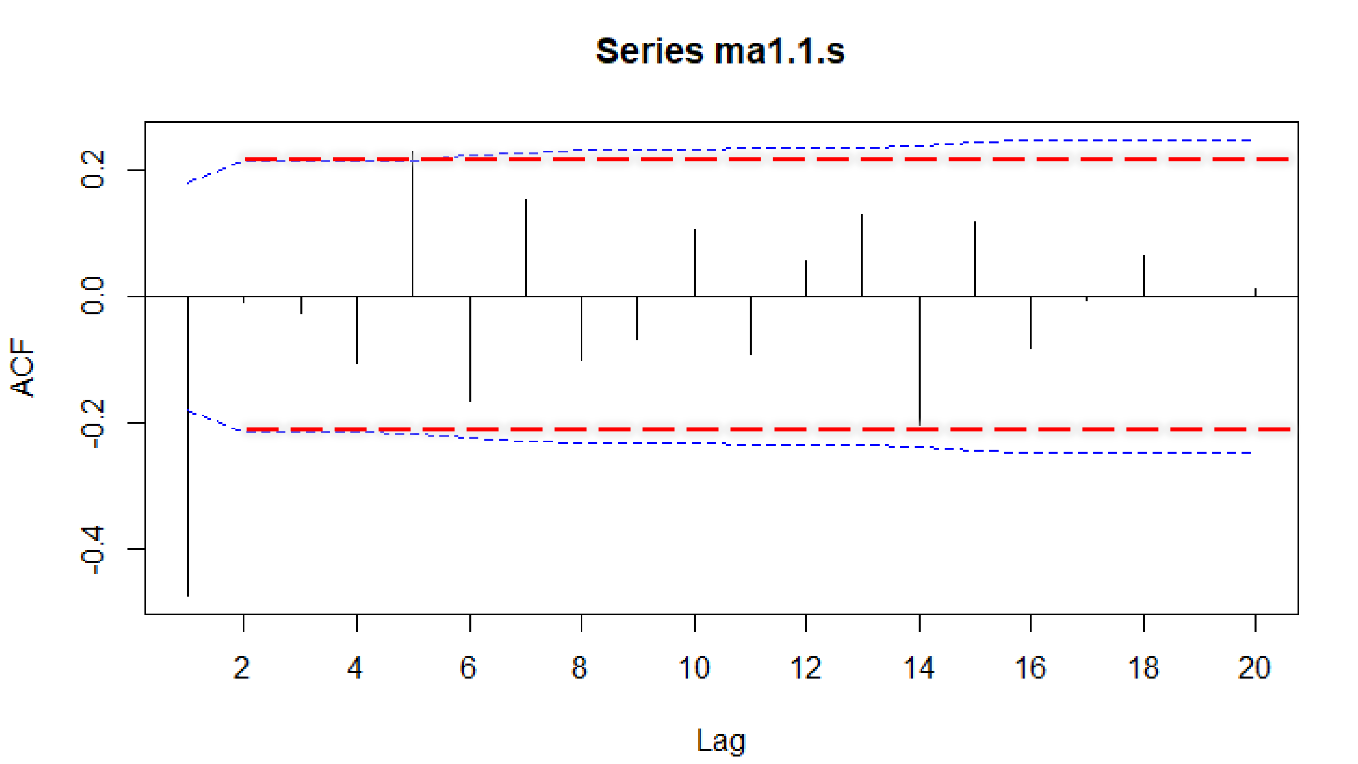
MA(1)模型识别范例

Exhibit 6.5 Sample Autocorrelation of an MA(1) Process with $\theta = 0.9$




```
> data(ma1.1.s)
> win.graph(width=4.875,height=3,pointsize=8)
> acf(ma1.1.s,xaxp=c(0,20,10))
```

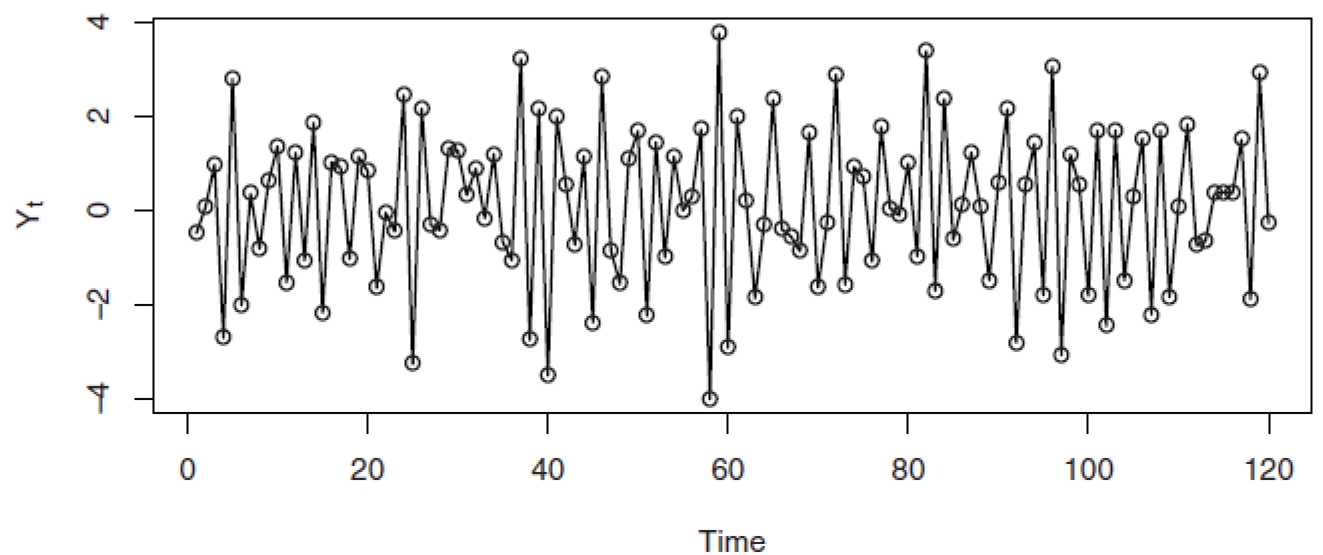
MA(1)模型识别范例




```
> acf(ma1.1.s,ci.type='ma',xaxp=c(0,20,10))
```

MA(2)模型识别范例

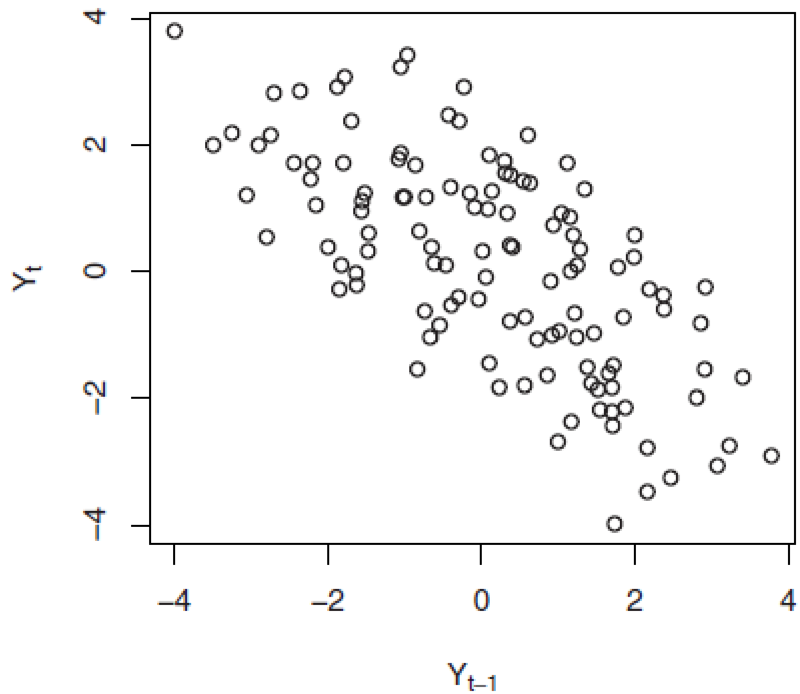
Exhibit 4.8 Time Plot of an MA(2) Process with $\theta_1 = 1$ and $\theta_2 = -0.6$




```
> win.graph(width=4.875, height=3,pointsize=8)
> data(ma2.s); plot(ma2.s,ylab=expression(Y[t]),type='o')
```

MA(2)模型识别范例

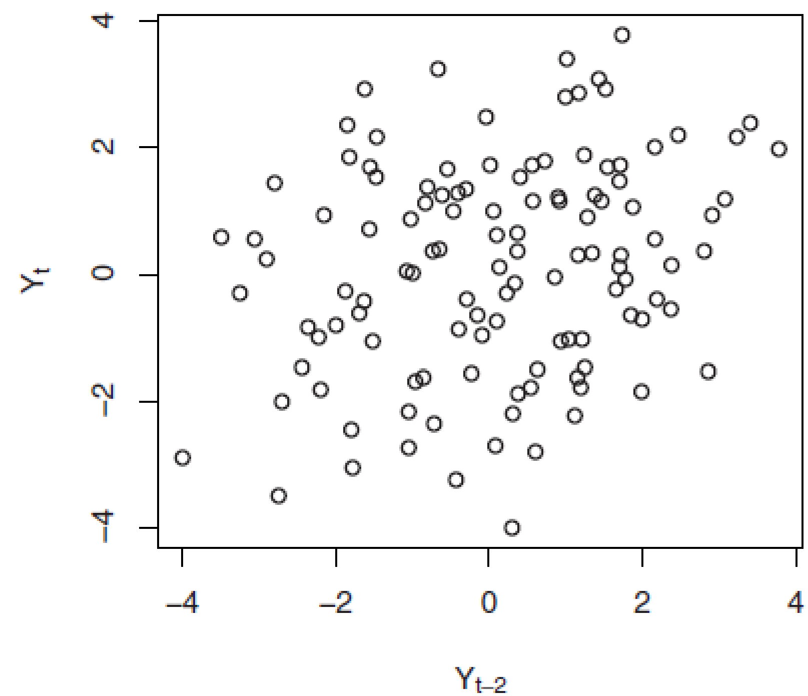
Exhibit 4.9 **Plot of Y_t versus Y_{t-1} for MA(2) Series in Exhibit 4.8**




```
> win.graph(width=3,height=3,pointsize=8)
> plot(y=ma2.s,x=zlag(ma2.s),ylab=expression(Y[t]),
   xlab=expression(Y[t-1]),type='p')
```

MA(2)模型识别范例

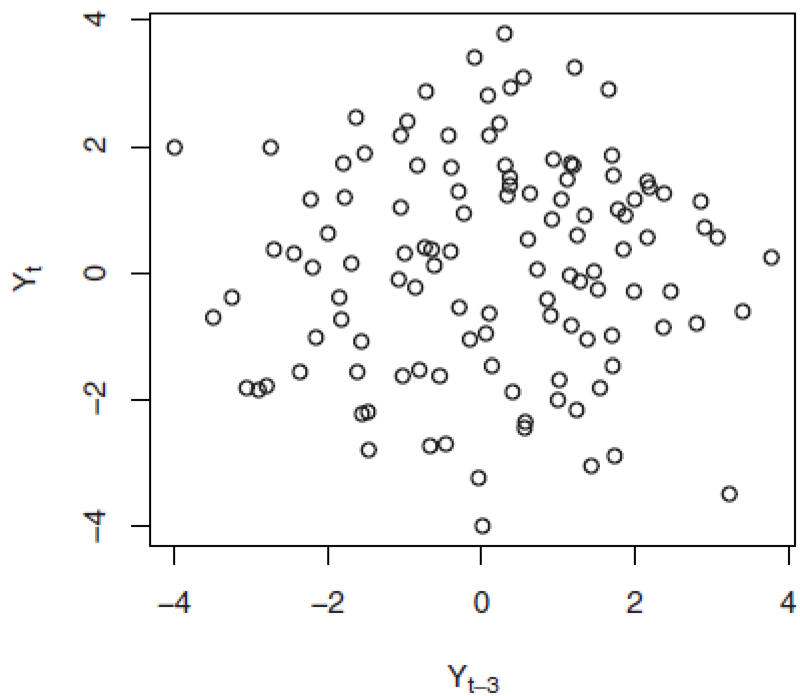
Exhibit 4.10 Plot of Y_t versus Y_{t-2} for MA(2) Series in Exhibit 4.8




```
> plot(y=ma2.s,x=zlag(ma2.s,2),ylab=expression(Y[t]),
   xlab=expression(Y[t-2]),type='p')
```

MA(2)模型识别范例

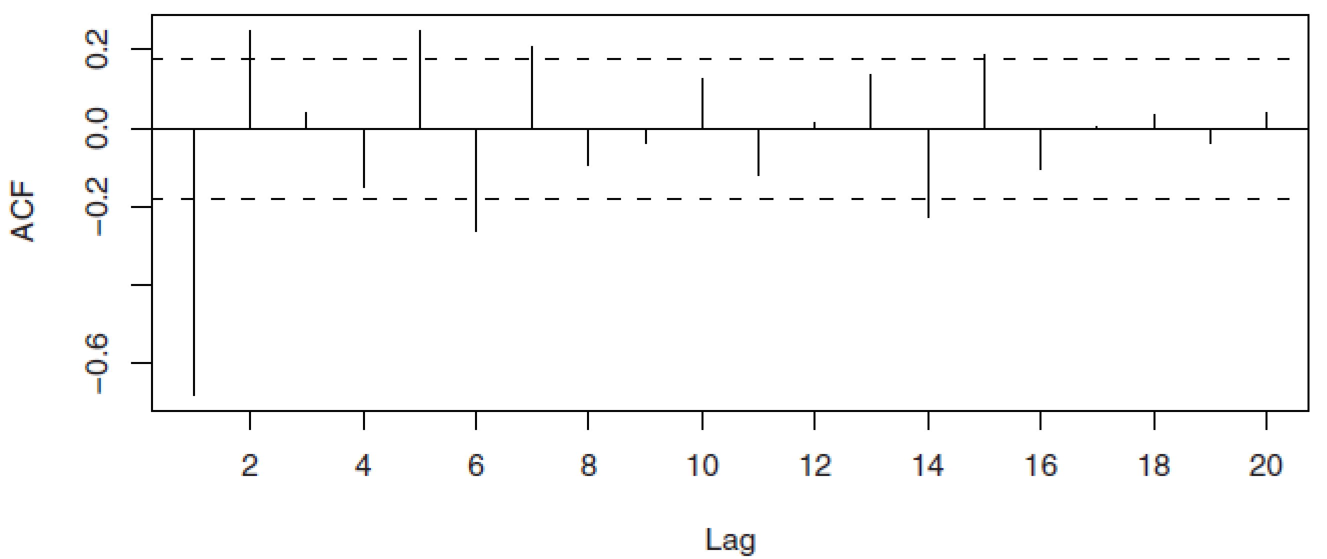
Exhibit 4.11 Plot of Y_t versus Y_{t-3} for MA(2) Series in Exhibit 4.8




```
> plot(y=ma2.s,x=zlag(ma2.s,3),ylab=expression(Y[t]),
  xlab=expression(Y[t-3]),type='p')
```

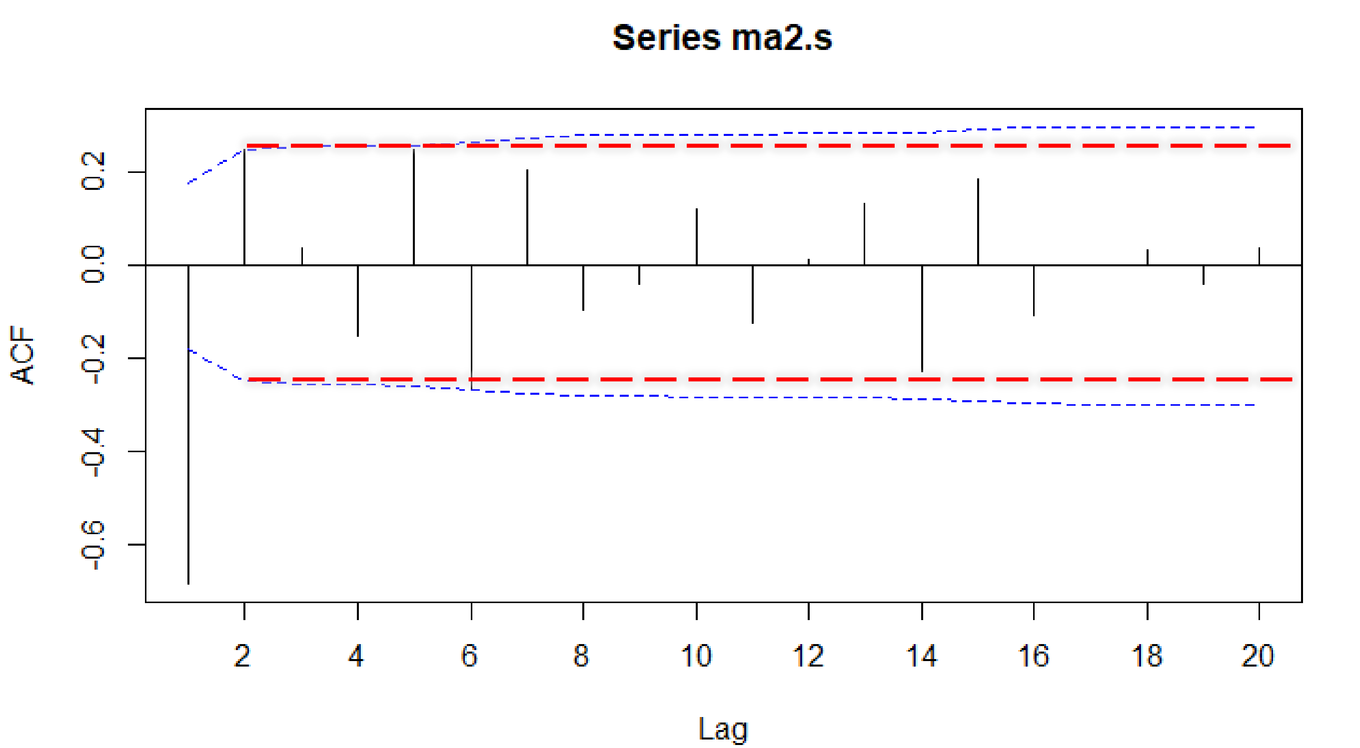
MA(2)模型识别范例

Exhibit 6.8 Sample ACF for an MA(2) Process with $\theta_1 = 1$ and $\theta_2 = -0.6$




```
> data(ma2.s); acf(ma2.s,xaxp=c(0,20,10))
```

MA(2)模型识别范例




```
> acf(ma2.s,ci.type='ma',xaxp=c(0,20,10))
```

MA模型识别

$$r_k = \frac{\sum_{t=k+1}^{n}(Y_t - \bar{Y})(Y_{t-k} - \bar{Y})}{\sum_{t=1}^{n}(Y_t - \bar{Y})^2}, k = 1,2, \ldots$$

- 对于白噪声(MA(0))，当$k > 0$时，$Var(r_k) \approx \frac{1}{n}$
- 对于MA(q)，当$k > q$时，$Var(r_k) \approx \frac{1}{n}\left[1 + 2\sum_{j=1}^{q}\rho_j^2\right]$
- r_k在正负2倍的标准差$\pm 2\hat{s}_q$之间的概率约为0.95，其中 $\hat{s}_q = \sqrt{\frac{1}{n}\left[1 + 2\sum_{j=1}^{q} r_j^2\right]}$
- 先判断序列是否为白噪声，再依次判断是否为MA(1), MA(2), ……

偏自相关函数（PACF）

- 回顾AR(k) 模型的Yule-Walker方程组

$$
\begin{aligned}
\rho_1 &= \phi_1 + \phi_2\rho_1 + \cdots + \phi_k\rho_{k-1} \\
\rho_2 &= \phi_1\rho_1 + \phi_2 + \cdots + \phi_k\rho_{k-2} \\
&\vdots \\
\rho_k &= \phi_1\rho_{k-1} + \phi_2\rho_{k-2} + \cdots + \phi_k
\end{aligned}
$$

- 对于任意$p < k$, 令$\phi_{p+1} = 0, \phi_{p+2} = 0, \dots, \phi_k = 0$，则得到的方程组对于AR(p)模型也满足，这很自然，因为AR(p)相当于后面$k - p$个系数全为零的AR(k)模型。
- 我们知道AR(p)的自相关函数$\{\rho_k\}$不截尾，上述思想提供了一个可能截尾的序列。

偏自相关函数（PACF）

- 对于任意平稳序列，已知$\rho_1, \rho_2, \dots$，对于给定的k，考虑方程组

$$
\begin{aligned}
\rho_1 &= \phi_{k1} + \phi_{k2}\rho_1 + \cdots + \phi_{kk}\rho_{k-1} \\
\rho_2 &= \phi_{k1}\rho_1 + \phi_{k2} + \cdots + \phi_{kk}\rho_{k-2} \\
&\vdots \\
\rho_k &= \phi_{k1}\rho_{k-1} + \phi_{k2}\rho_{k-2} + \cdots + \phi_{kk}
\end{aligned}
$$

- 我们把 $\{\phi_{kk}\}$称为 偏自相关函数（PACF）序列，有如下重要结论:
- 对于AR(p)过程，当$k \geq p$时，

$$
\phi_{kj} = \begin{cases} \phi_j & j = 1, \dots, p \\ 0 & j = p+1, \dots, k \end{cases}
$$

- 特别地，$\phi_{pp} = \phi_p \neq 0, \phi_{kk} = 0, k > p$, 即AR(p)过程的偏自相关函数序列$p$阶截尾。

偏自相关函数（PACF）的递推式

- Levinson (1947) 和 Durbin (1960) 给出了PACF的递推式，

$$\phi_{kk} = \frac{\rho_k - \sum_{j=1}^{k-1} \phi_{k-1,j}\rho_{k-j}}{1 - \sum_{j=1}^{k-1} \phi_{k-1,j}\rho_j}$$

$$\phi_{kj} = \phi_{k-1,j} - \phi_{kk}\phi_{k-1,k-j}, \qquad j = 1,2,\dots,k-1$$

偏自相关函数（PACF）的其它定义

- (1) $\phi_{kk} = Corr(Y_t - \beta_1 Y_{t-1} - \beta_2 Y_{t-2} - \cdots - \beta_{k-1} Y_{t-k+1},$
  $$Y_{t-k} - \beta_1 Y_{t-k+1} - \beta_2 Y_{t-k+2} - \cdots - \beta_{k-1} Y_{t-1})$$
- 对于零均值序列，
  $\beta_1, \beta_2, \dots, \beta_{k-1}$ 最小化均方误差 $E(Y_t - \beta_1 Y_{t-1} - \beta_2 Y_{t-2} - \cdots - \beta_{k-1} Y_{t-k+1})^2$
- 对于AR(p)过程，当$k > p$时，$\beta_j = \phi_j, j \le p$; $\beta_j = 0, j > p.$
- $\phi_{kk} = Corr\left(e_t, Y_{t-k} - \phi_1 Y_{t-k+1} - \phi_2 Y_{t-k+2} - \cdots - \phi_p Y_{t-k+p}\right) = 0$

- (2) 对于正态分布序列，$\phi_{kk} = Corr(Y_t, Y_{t-k} | Y_{t-1}, Y_{t-2}, \dots, Y_{t-k+1})$
- 对于AR(p)过程，当$k > p$时，
  $$\phi_{kk} = Corr\left(\phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + e_t, Y_{t-k} \middle| Y_{t-1}, Y_{t-2}, \dots, Y_{t-k+1}\right) = 0$$

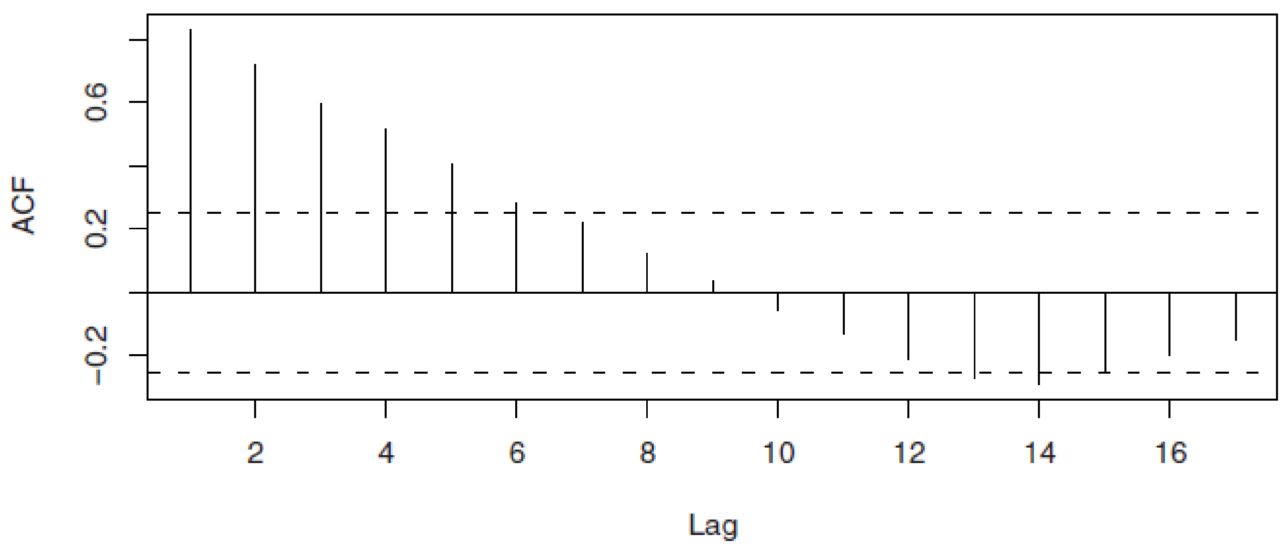
样本偏自相关函数

- 前面所定义的偏自相关函数都是基于序列的性质，由其内在的自相关函数计算得出（与自相关函数类似，是模型本身的参数）；当我们有具体样本数值时，可以用样本自相关函数r_k代替ρ_k来计算，例如求解方程组、递推式等，记为$\hat{\phi}_{kk}$.
- 由于AR(p)过程的PACF序列p阶截尾，当$k > p$时，$\phi_{kk} = 0$, 所以$\hat{\phi}_{kk}$应该也接近0，实际上，Quenoulle (1949) 证明了对于AR(p)过程，当$k > p$时，$\hat{\phi}_{kk}$近似服从均值为0，方差为$1/n$的正态分布，所以其在正负两倍标准差$\pm 2/\sqrt{n}$之间的概率约为0.95.

$$P\left(-\frac{2}{\sqrt{n}} \leq \hat{\phi}_{kk} \leq \frac{2}{\sqrt{n}}\right) \approx 0.95$$

AR(1)模型识别范例

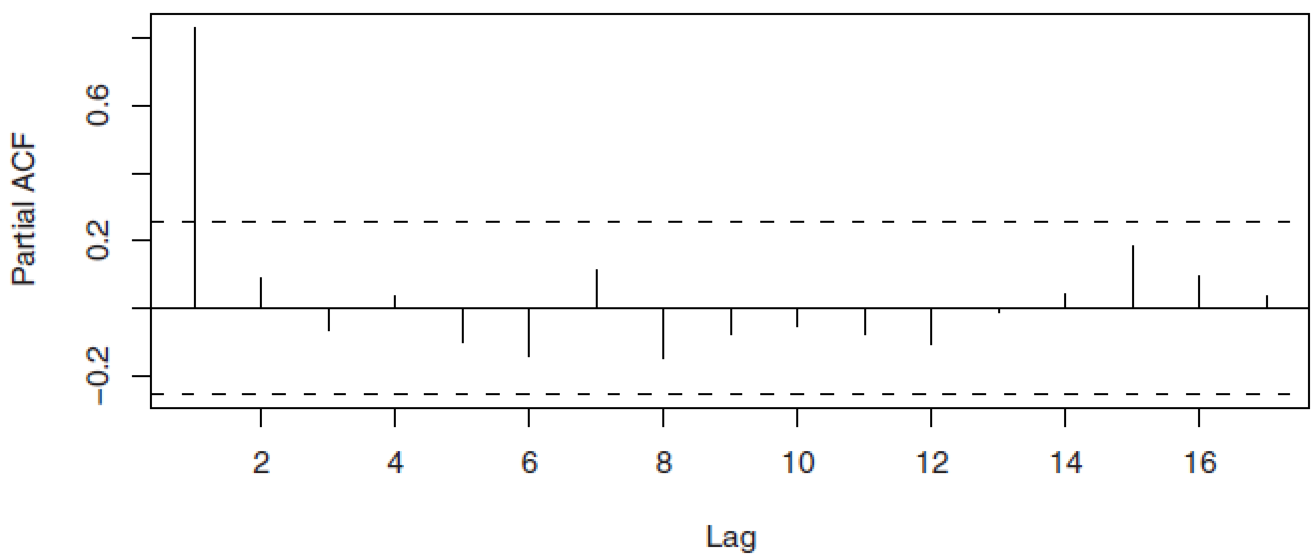
Exhibit 6.10 Sample ACF for an AR(1) Process with $\phi = 0.9$




```
> data(ar1.s); acf(ar1.s,xaxp=c(0,20,10))
```

AR(1)模型识别范例

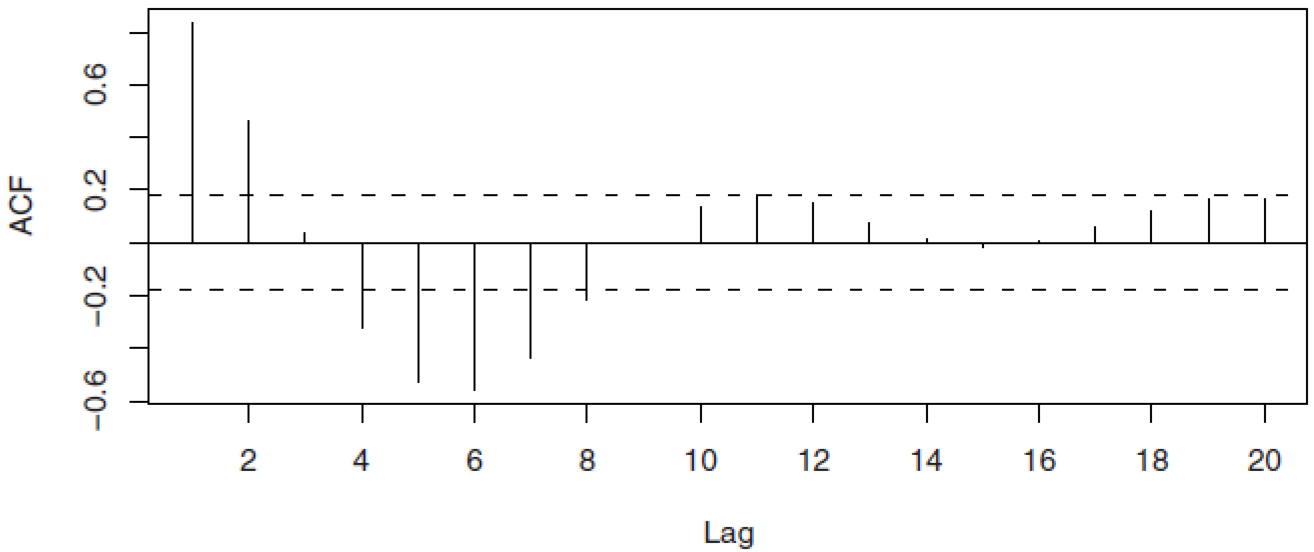
Exhibit 6.11 Sample Partial ACF for an AR(1) Process with ϕ = 0.9




```
> pacf(ar1.s,xaxp=c(0,20,10))
```

AR(2)模型识别范例

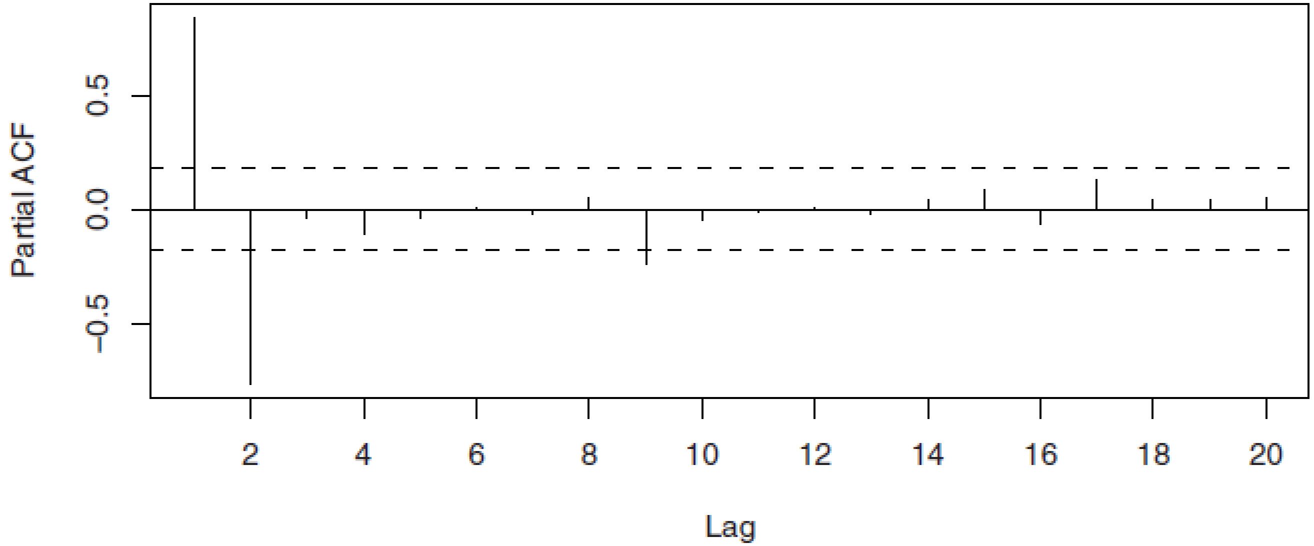
Exhibit 6.12 Sample ACF for an AR(2) Process with $\phi_1 = 1.5$ and $\phi_2 = -0.75$




```
> acf(ar2.s,xaxp=c(0,20,10))
```

AR(2)模型识别范例

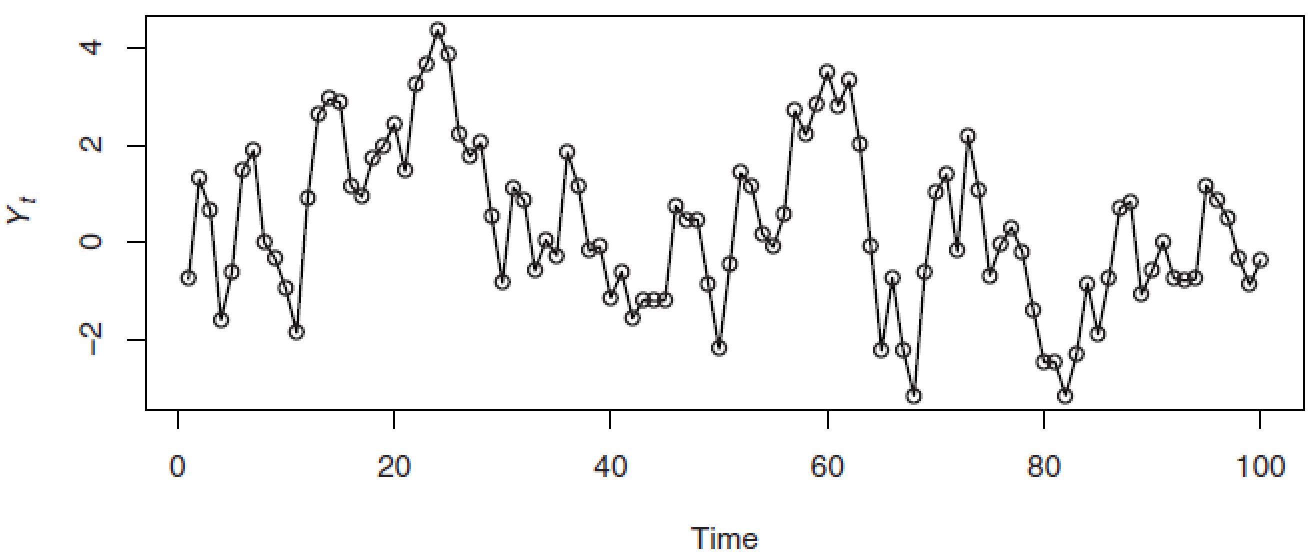
Exhibit 6.13 Sample PACF for an AR(2) Process with ϕ_1 = 1.5 and ϕ_2 = −0.75




```
> pacf(ar2.s,xaxp=c(0,20,10))
```

ARMA(1, 1)模型识别

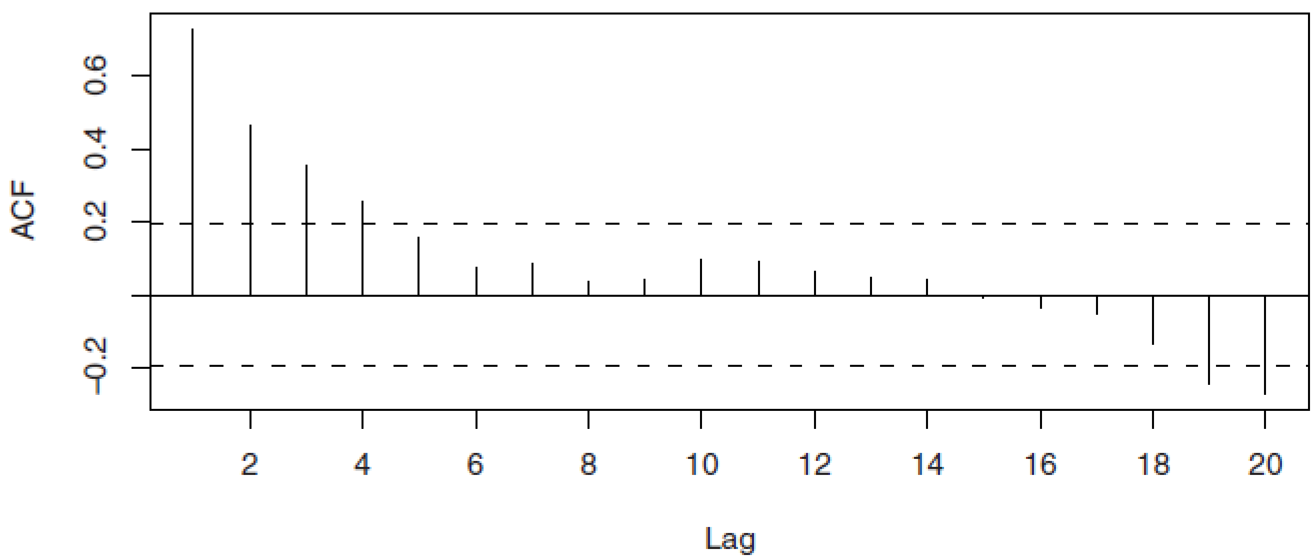
Exhibit 6.14 Simulated ARMA(1,1) Series with $\phi = 0.6$ and $\theta = -0.3$.




```
> data(arma11.s)
> plot(arma11.s, type='o',ylab=expression(Y[t]))
```

ARMA(1, 1)模型识别

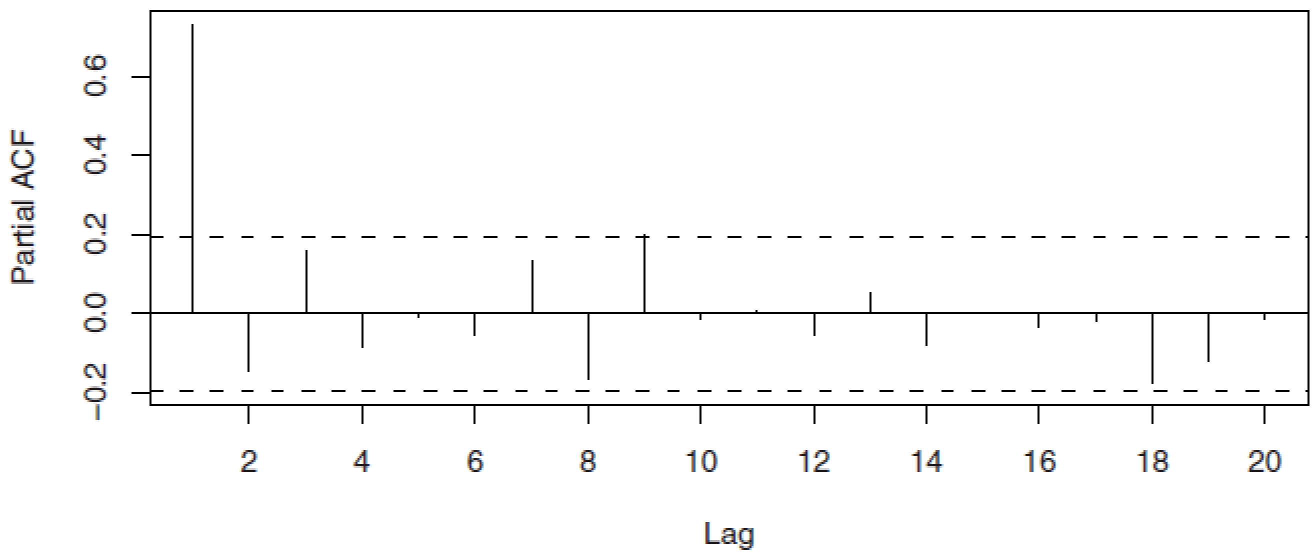
Exhibit 6.15 Sample ACF for Simulated ARMA(1,1) Series




```
> acf(arma11.s,xaxp=c(0,20,10))
```

ARMA(1, 1)模型识别

Exhibit 6.16　Sample PACF for Simulated ARMA(1,1) Series




```
> pacf(arma11.s,xaxp=c(0,20,10))
```

ARMA模型识别和EACF

- 由于ARMA模型的ACF和PACF均不截尾，所以只用ACF和PACF是无法给ARMA模型定阶的。
- 扩展的自相关函数（EACF）基本思路：通过线性回归滤出AR部分，只剩下MA部分，剩下部分的ACF截尾。
- 对于ARMA(1, 1)模型，$Y_t = \phi Y_{t-1} + e_t - \theta e_{t-1}$，以$Y_t$为因变量，$Y_{t-1}$为自变量做简单线性回归，得到参数$\hat{\phi} \approx r_1$，$\hat{\phi}$并不是$\phi$的一致估计量，因为它收敛到$\rho_1 \neq \phi$. 记$\varepsilon_t^1 = Y_t - \hat{\phi} Y_{t-1}$为第一次回归的残差；
- 第二次回归以Y_t为因变量，Y_{t-1}和ε_{t-1}^1为自变量做多元线性回归，得到的Y_{t-1}的系数$\tilde{\phi}$是ϕ的一致估计量，令$W_{t,1,1} = Y_t - \tilde{\phi} Y_{t-1}$, 则$\{W_{t,1,1}\}$近似是MA(1)模型。

EACF的解释

- 对于ARMA(1, 2)模型$Y_t = \phi Y_{t-1} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2}$， 前两次回归和之前一样，第一次以 Y_t 为因变量，Y_{t-1}为自变量做回归，记 ε_t^1为第一次回归的残差；第二次以Y_t为因变量， Y_{t-1}和ε_{t-1}^1为自变量做回归，记ε_t^2为第二次回归的残差；
- 第三次以Y_t为因变量， $Y_{t-1}, \varepsilon_{t-1}^2$和$\varepsilon_{t-2}^1$为自变量做多元线性回归，得到的$Y_{t-1}$的系数$\tilde{\phi}$是$\phi$的一致估计量，令$W_{t,1,2} = Y_t - \tilde{\phi} Y_{t-1}$, 则$\{W_{t,1,2}\}$近似是MA(2)模型。
- 注意，如果真实模型为ARMA(1, 1), 那么做了三次回归后$\tilde{\phi}$仍是ϕ的一致估计量，这时$\{W_{t,1,2}\}$近似是MA(1)模型。

EACF的解释

- 对于序列$\{Y_t\}$,假设它的AR部分为k阶，第一次以Y_t为因变量，$Y_{t-1},\dots,Y_{t-k}$为自变量做回归，记ε_t^1为第一次回归的残差；第二次以Y_t为因变量，$Y_{t-1},\dots,Y_{t-k}$和ε_{t-1}^1为自变量做多元线性回归，记ε_t^2为第二次回归的残差；……
- 第$j+1$次以Y_t为因变量，$Y_{t-1},\dots,Y_{t-k},\varepsilon_{t-1}^j,\varepsilon_{t-2}^{j-1},\dots,\varepsilon_{t-j}^1$为自变量做回归，得到了$Y_{t-1},\dots,Y_{t-k}$的系数$\tilde{\phi}_1,\dots\tilde{\phi}_k$，令$W_{t,k,j}=Y_t-\tilde{\phi}_1Y_{t-1}-\cdots-\tilde{\phi}_kY_{t-k}$
- 如果真实模型为ARMA(p, q), 对于$k=p,j\geq q$，经过j+1次回归后$\tilde{\phi}_1,\dots\tilde{\phi}_p$是$\phi_1,\dots,\phi_p$的一致估计量，这时$\{W_{t,k,j}\}$近似是MA(q)模型，所以其ACF序列$q$阶截尾。

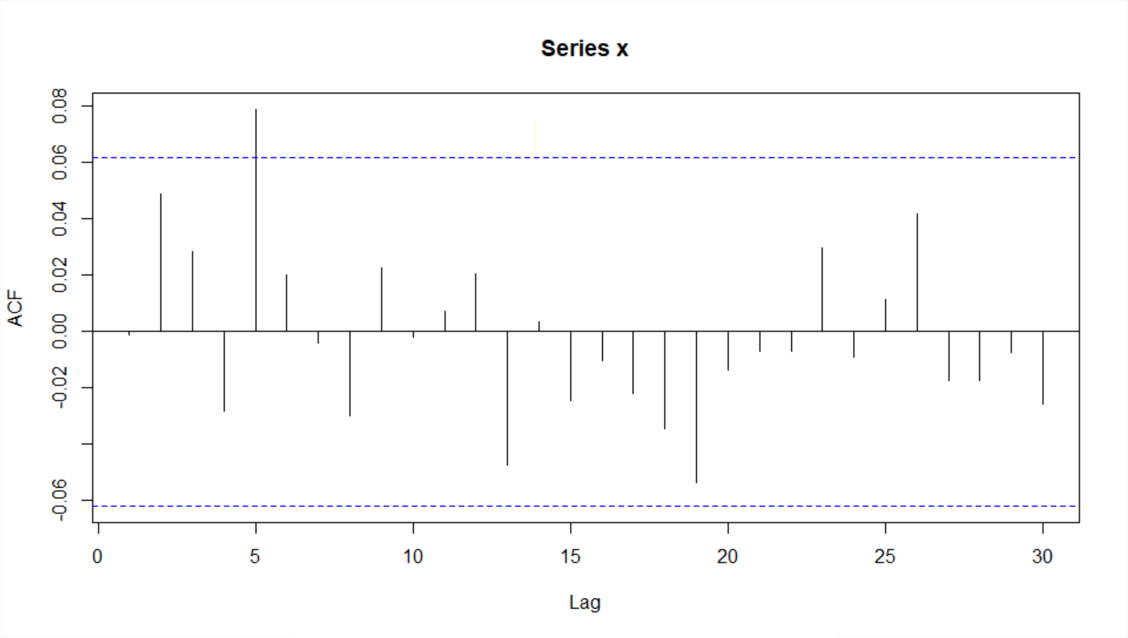
EACF的解释

- EACF是这样一张表，对于给定的数据，假设其AR部分为k阶，做了j+1次回归后得到$\{W_{t,k,j}\}$，如果其j+1阶样本ACF显著不为0（绝对值大于$1.96/\sqrt{n-j-k}$），则EACF表的第k行第j列标为“x”，否则标为“o”
- 如果真实模型为ARMA(p, q), 对于$k=p, j\geq q$，经过j+1次回归后$\{W_{t,k,j}\}$近似是MA(q)模型，其j+1阶样本ACF应当接近0，所以EACF表的第p行从第q列开始大概率为“o”
- 当$k>p$时，会出现过度拟合问题，这将导致$\{W_{t,k,j}\}$的MA阶数增加，例如，EACF表的第p+1行从第q+1列开始才大概率为“o”，所以ARMA(p, q)过程的EACF表理论上有一个由“o”构成的三角模式，顶点为第p行第q列。

例一：白噪声

- 对于白噪声序列$\{e_t\}$，假设它的AR部分为0阶，则无论做几次回归都有$W_{t,0,j} = e_t$， ACF为0阶截尾。
- 如果假设它的AR部分为1阶，则无论做几次回归，$\{W_{t,1,j}\}$都满足MA(1)模型，ACF为1阶截尾。
- 如果假设它的AR部分为k阶，则无论做几次回归， $W_{t,k,j}= e_t - \tilde{\phi}_1 e_{t-1} - \cdots - \tilde{\phi}_k e_{t-k}$都满足MA(k)模型，ACF为k阶截尾。
- 所以白噪声的EACF表的第k行从第k列开始大概率为“o”，但是也有意外情况，如果该白噪声的样本ACF在第q阶显著不为0，则EACF表的第q-1列会出现一定的“x”,这是EACF表中常见的一种现象。

例一： 白噪声




```
> eacf(x)
AR/MA
  0 1 2 3 4 5 6 7 8 9 10 11 12 13
0 o o o o x o o o o o o  o  o  o
1 o o o o x o o o o o o  o  o  o
2 x x o o x o o o o o o  o  o  o
3 x x x o x o o o o o o  o  o  o
4 x x x x x o o o o o o  o  o  o
5 x x x x x o o o o o o  o  o  o
6 x x x x x o o o o o o  o  o  o
7 x x x x o o o o o o o  o  o  o
```

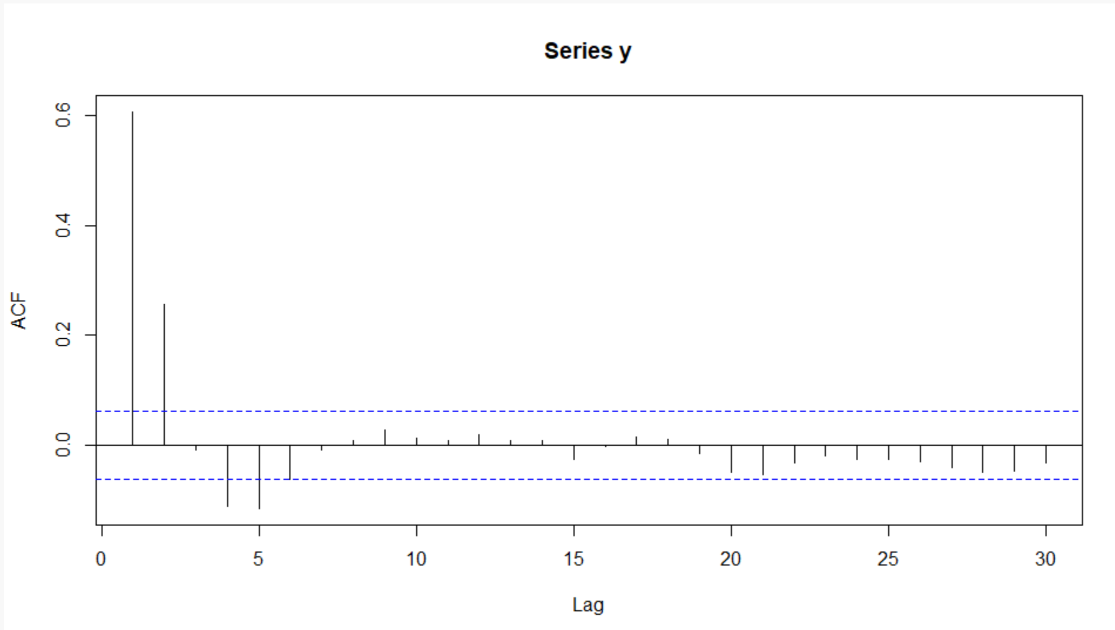
例二：ARMA(2, 2)

- 对于ARMA(2, 2)序列$\{Y_t\}$，当假设它的AR部分为0阶或1阶时，做回归无法有效去除AR部分，所以ACF无明显截尾。
- 如果假设它的AR部分为2阶，则做3次或更多次回归之后，$\{W_{t,2,j}\}$有效去除了AR部分，满足MA(2)模型，ACF为2阶截尾。其3阶样本ACF不显著，所以EACF表的第2行从第2列开始大概率为“o”
- 如果假设它的AR部分为k>2阶，则做k+1次回归之后，$\{W_{t,k,j}\}$大致满足MA(k)模型，ACF为k阶截尾。其k+1阶样本ACF不显著，所以EACF表的第k行从第k列开始大概率为“o”，EACF表中的“o”呈现出以第2行第2列为“顶点”的三角模式。

例二：ARMA(2, 2)

- 对于ARMA(2, 2)序列$\{Y_t\}$，当$k = 0, 1$时，做回归无法有效去除AR部分，所以ACF无明显截尾。
- EACF的第2行有什么特点?
- EACF的第k行有什么特点?
- EACF表中的“o”大致呈现出以第2行第2列为“顶点”的三角模式。

例二： ARMA(2, 2)




```
> eacf(y)
AR/MA
  0 1 2 3 4 5 6 7 8 9 10 11 12 13
0 x x o x x o o o o o o  o  o  o
1 x x o x x o o o o o o  o  o  o
2 x x o o o o o o o o o  o  o  o
3 o o x o o o o o o o o  o  o  o
4 x o x o o o o o o o o  o  o  o
5 x x o o o o o o o o o  o  o  o
6 x x x o o o o o o o o  o  o  o
7 x x x x o o o o o o o  o  o  o
```

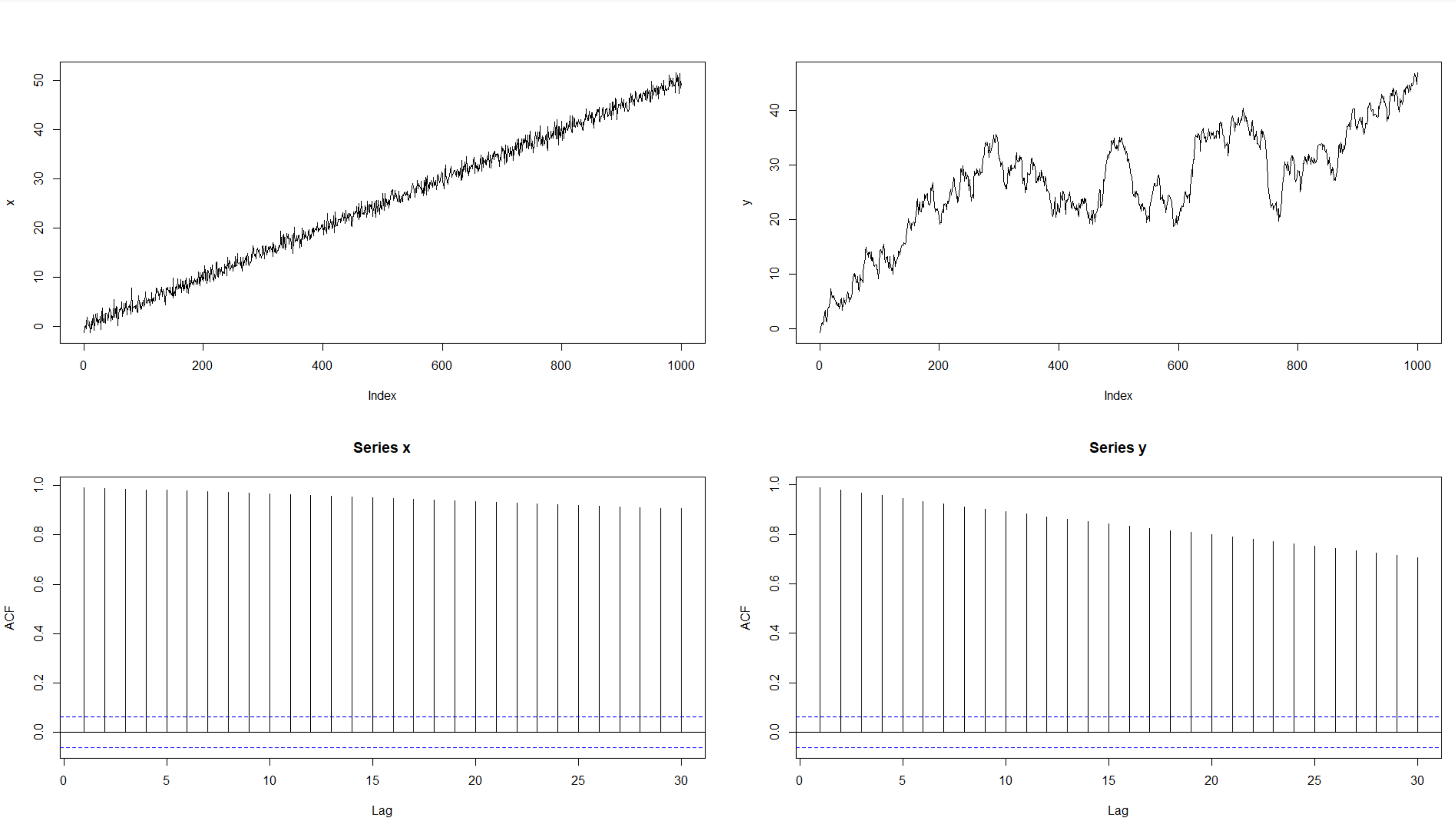
非平稳模型识别

- 当样本ACF呈现出缓慢下降的趋势时，我们可以尝试对原序列进行一次差分，对差分后的序列观察ACF、PACF、EACF等，尝试建立平稳模型。
- 如果一次差分后的样本ACF仍然缓慢下降，可以考虑再次差分，但为了避免过度差分，建议仔细查看每次差分后的序列本身及其自相关特性：模型尽量简洁，但也不能草率。
- 可以结合数据实际背景做变换，如对数差分变换，Box-Cox变换等。

例：确定趋势VS随机趋势

$X_t = 0.05t + e_t$

$Y_t = Y_{t-1} + e_t$



例：确定趋势VS随机趋势

$\nabla X_t = 0.05 + e_t - e_{t-1}$

$\nabla Y_t = e_t$