



廈門大學

XIAMEN UNIVERSITY

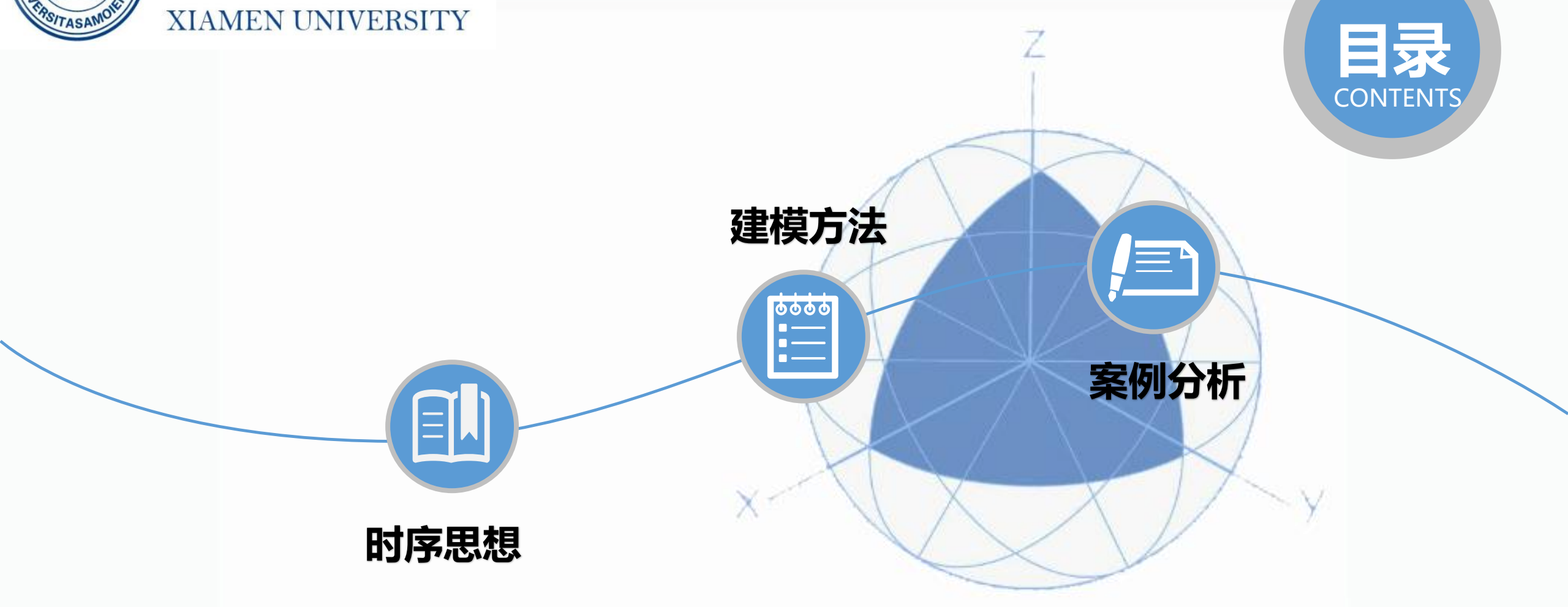
# 时间序列分析方法

谭忠





厦门大学  
XIAMEN UNIVERSITY



时序思想

建模方法

案例分析



廈門大學  
XIAMEN UNIVERSITY

Part 1

# 时序思想 与建模方法





## 时间序列分析方法概述

从统计学的内容来看，统计所研究和处理的是一批有“实际背景”的数据，尽管数据的背景和类型各不相同，但从数据的形成来看，无非是横剖面数据和纵剖面数据两类（也称为静态数据和动态数据）。



横剖面数据是有若干相关现象在某一时点上所处的状态组成的，它反映一定时间、地点等客观条件下诸相关现象之间的存在的内在数值联系。研究这种数据结构测统计方法就是多元统计方法。纵剖面数据是由某一现象或者若干现象在不同时刻上的状态所形成的数据，它反映的是现象以及现象之间关系的发展变化规律性。研究这种数据的统计方法就是时间序列分析。



## 时间序列数据的特点

□**时间性**：序列中的数据或者数据点的位置依赖于时间，即数据的取值依赖于时间的变化，但不一定是时间  $t$  的严格函数.

□**随机性**：每一时刻上的取值或者数据点的位置具有一定的随机性，不可能完全准确地用历史值预测.

□**相关性**：前后时刻 (不一定是相邻时刻)的数值或者数据点有一定的相关性，这种相关性就是系统的动态规律性.

最后，从整体上来看，时间序列往往呈现某种趋势性或者出现周期性变化的现象.



## 10.2.1 基本概念

### 1、随机过程与时间序列

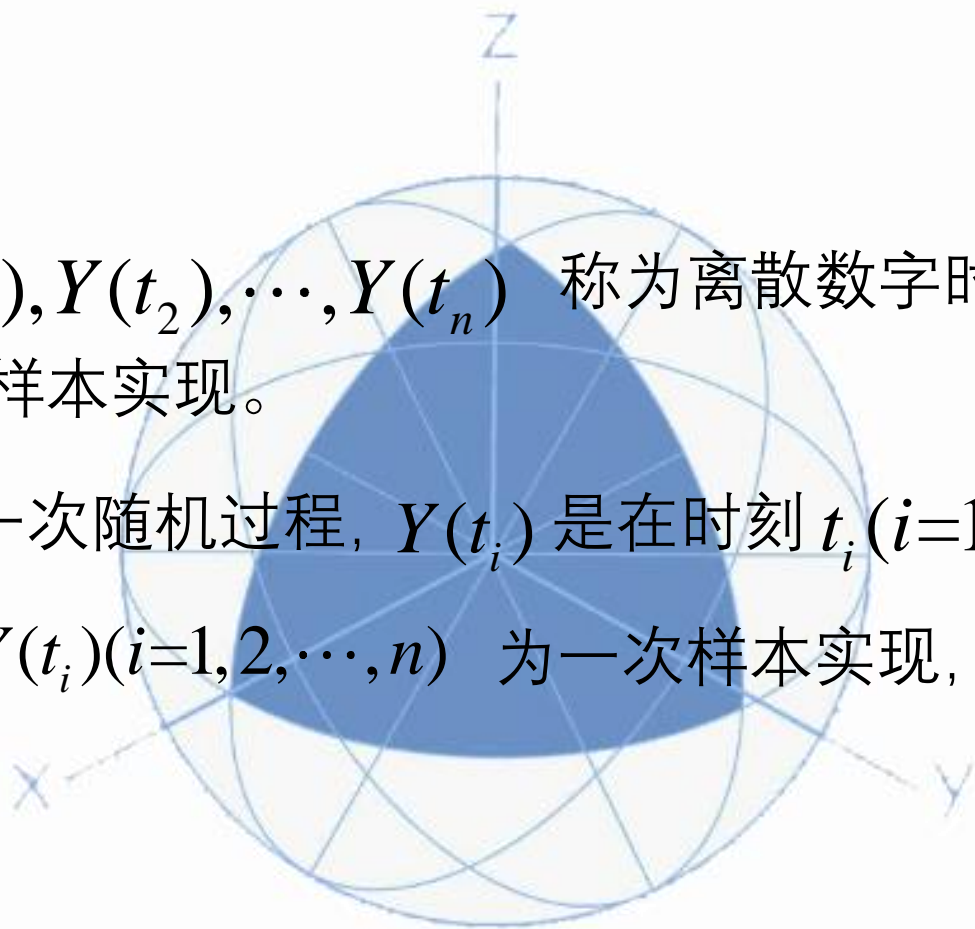
若  $\forall t \in T, Y_t$  是一随机变量，则称  $\{Y_t, t \in T\}$  为一随机过程，由此可见，随机过程  $Y_t$  是依赖于时间  $t$  的一簇随机变量。

从数学意义上而言，若我们对某一过程中某一个变量或一组变量  $Y_t$  进行观察测量，在一系列时刻  $t_1, t_2, \dots, t_n (t_1 < t_2 < \dots < t_n)$



得到的离散有序数集合  $Y(t_1), Y(t_2), \dots, Y(t_n)$  称为离散数字时间序列，即随机过程的一次样本实现。

如果，我们设  $\{Y_t, t \in T\}$  是一次随机过程， $Y(t_i)$  是在时刻  $t_i (i=1, 2, \dots, n)$  对过程  $Y_t$  的观测值，则称  $Y(t_i) (i=1, 2, \dots, n)$  为一次样本实现，也就是一个时间序列。



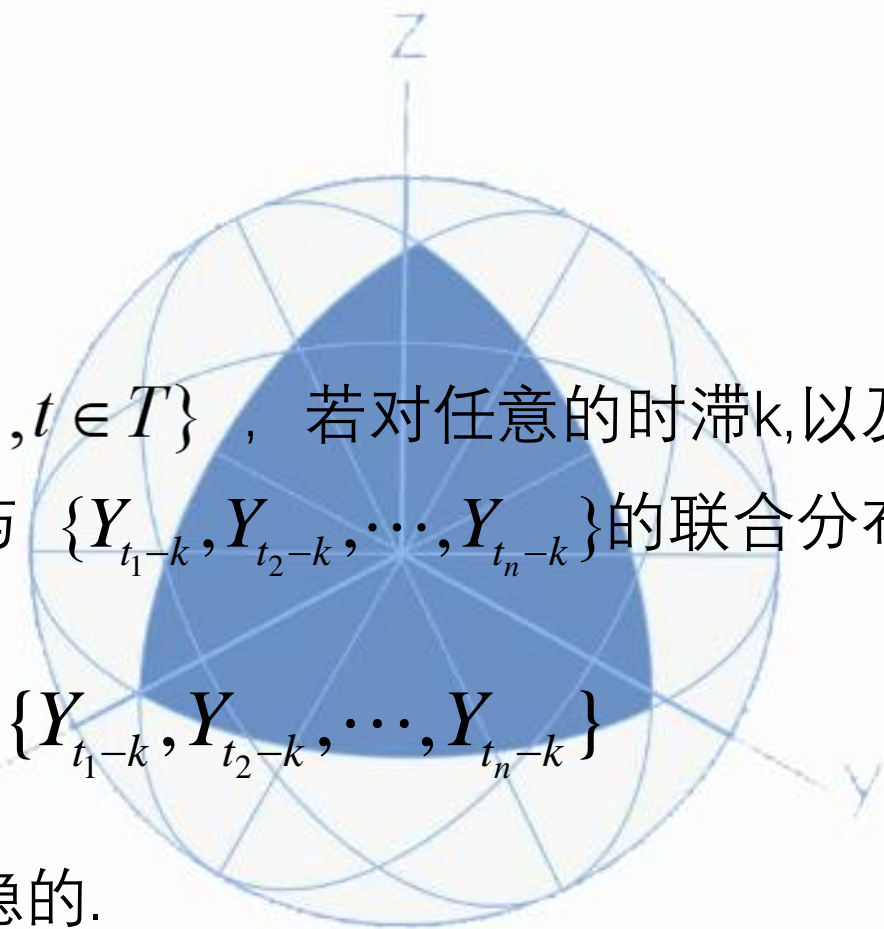


## 2、平稳性

□ 严平稳：对于随机过程  $\{Y_t, t \in T\}$ ，若对任意的时滞  $k$ ，以及时点  $t_1, t_2, \dots, t_n$  有  $\{Y_{t_1}, Y_{t_2}, \dots, Y_{t_n}\}$  与  $\{Y_{t_1-k}, Y_{t_2-k}, \dots, Y_{t_n-k}\}$  的联合分布相同，即

$$F_n\{Y_{t_1}, Y_{t_2}, \dots, Y_{t_n}\} = F_n\{Y_{t_1-k}, Y_{t_2-k}, \dots, Y_{t_n-k}\}$$

这时我们称随机过程  $Y_t$  是严平稳的.



因为分布函数完整地描述了随机变量的统计特性，所以上式表明该序列的统计特性不随时间的变化而变化。然而在实际中，严平稳的概念过于严格且不易进行检验。另外，在相关应用中，与概率分布相比而言，我们更关心变量的均值、方差与协方差等，所以有了下面的弱平稳概念。



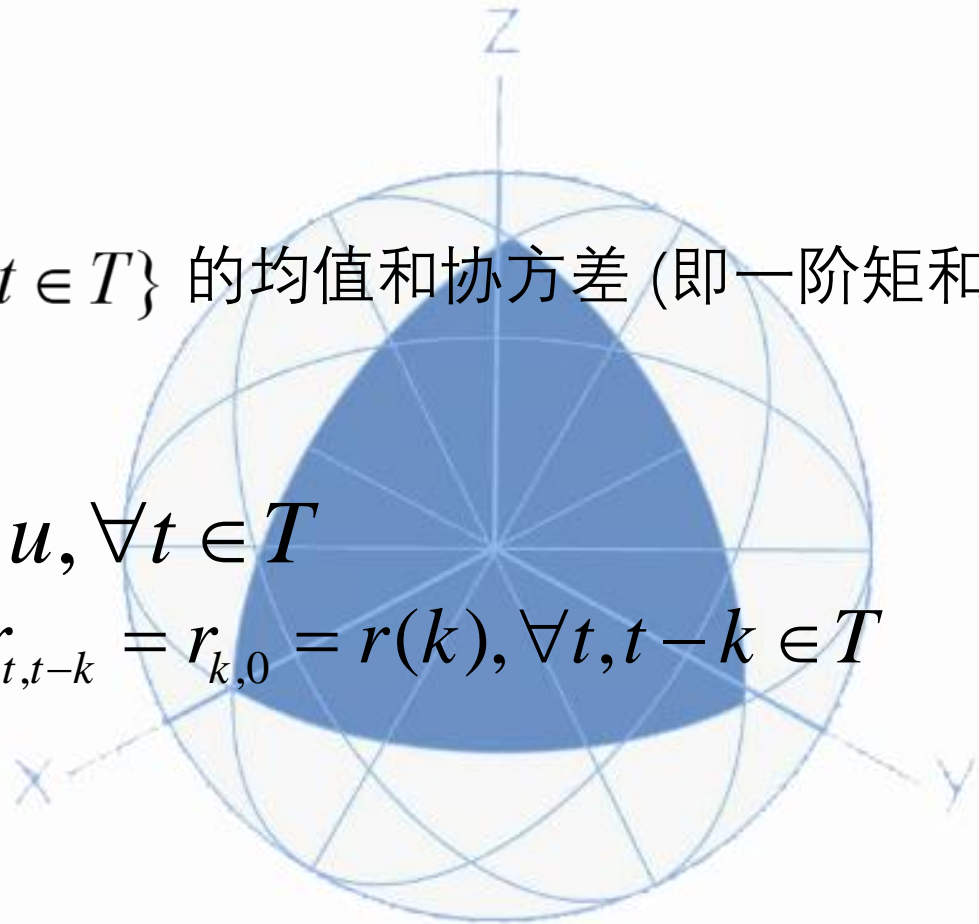


□ **弱平稳**: 若随机过程  $\{Y_t, t \in T\}$  的均值和协方差 (即一阶矩和二阶矩) 存在, 且满足:

$$E(Y_t) = u, \forall t \in T$$

$$E[(Y_t - u)(Y_{t-k} - u)] = r_{t,t-k} = r_{k,0} = r(k), \forall t, t-k \in T$$

则称该过程是弱平稳过程.






严平稳对时间推移的不变性表现在统计平均的概率分布上，而弱平稳对时间推移的不变性 表现在统计平均的一、二阶矩上. 一般说来，严平稳比弱平稳要求要“严”，二者之间的联系表现在：

(1) 严平稳  $\Rightarrow$  弱平稳：因为概率分布存在，但是对应的期望和协方差不一定存在，如柯西 分布的一、二阶矩均不存在.

(2) 弱平稳  $\Rightarrow$  严平稳：这是显然的.





但是对于正态过程而言，严平稳 $\Leftrightarrow$ 弱平稳，这是因为一方面正态过程的一二阶矩是存在的，所以严平稳一定能推出弱平稳. 另一方面因为正态过程的概率密度函数是由均值和自协方差函数完全确定的，当均值和自协方差函数不随时间平移而变化时，对应的概率分布函数也不会随时间的平移而变化，从而得出一个弱平稳的正态过程肯定也是严平稳的.

### 3、自相关函数与样本自相关函数

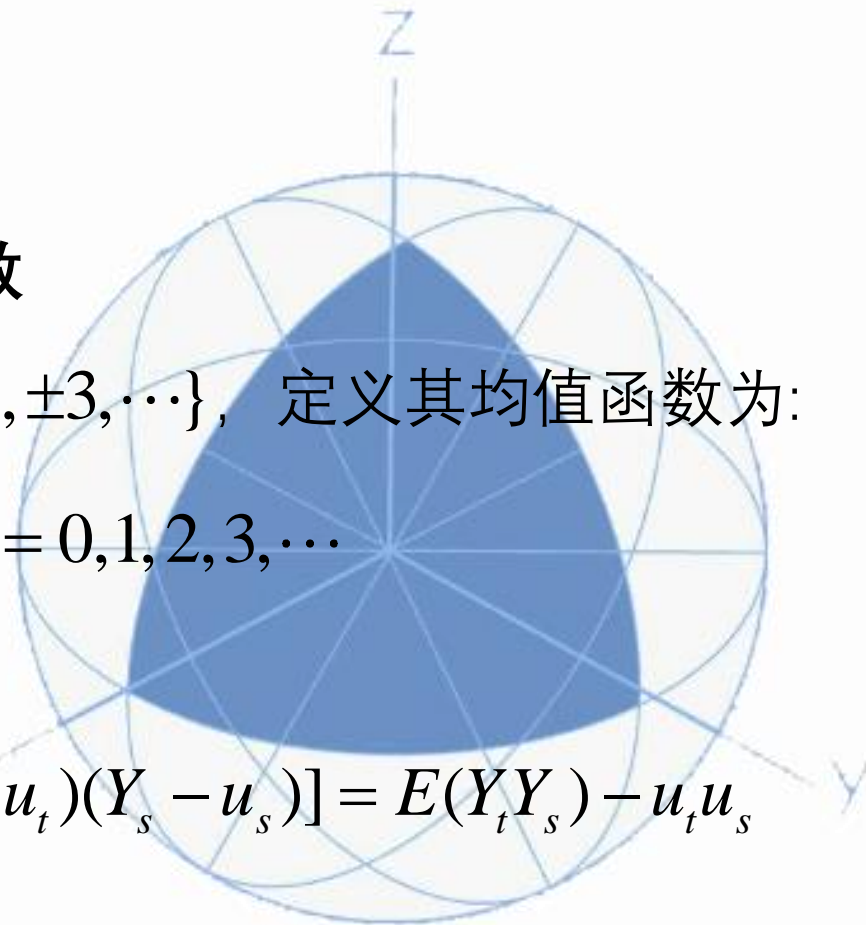
对一个时间序列  $\{Y_t : t = 0, \pm 1, \pm 2, \pm 3, \dots\}$ , 定义其均值函数为:

$$u_t = E(Y_t), t = 0, 1, 2, 3, \dots$$

自协方差函数为

$$r_{t,s} = Cov(Y_t, Y_s) = E[(Y_t - u_t)(Y_s - u_s)] = E(Y_t Y_s) - u_t u_s$$

其中  $t, s = 0, \pm 1, \pm 2, \pm 3, \dots$





自相关函数:

$$\rho_{t,s} = \text{Corr}(Y_t, Y_s) = \frac{\text{Cov}(Y_t, Y_s)}{\sqrt{\text{Var}(Y_t)} \sqrt{\text{Var}(Y_s)}} = \frac{r_{t,s}}{\sqrt{r_{t,t} r_{s,s}}}$$

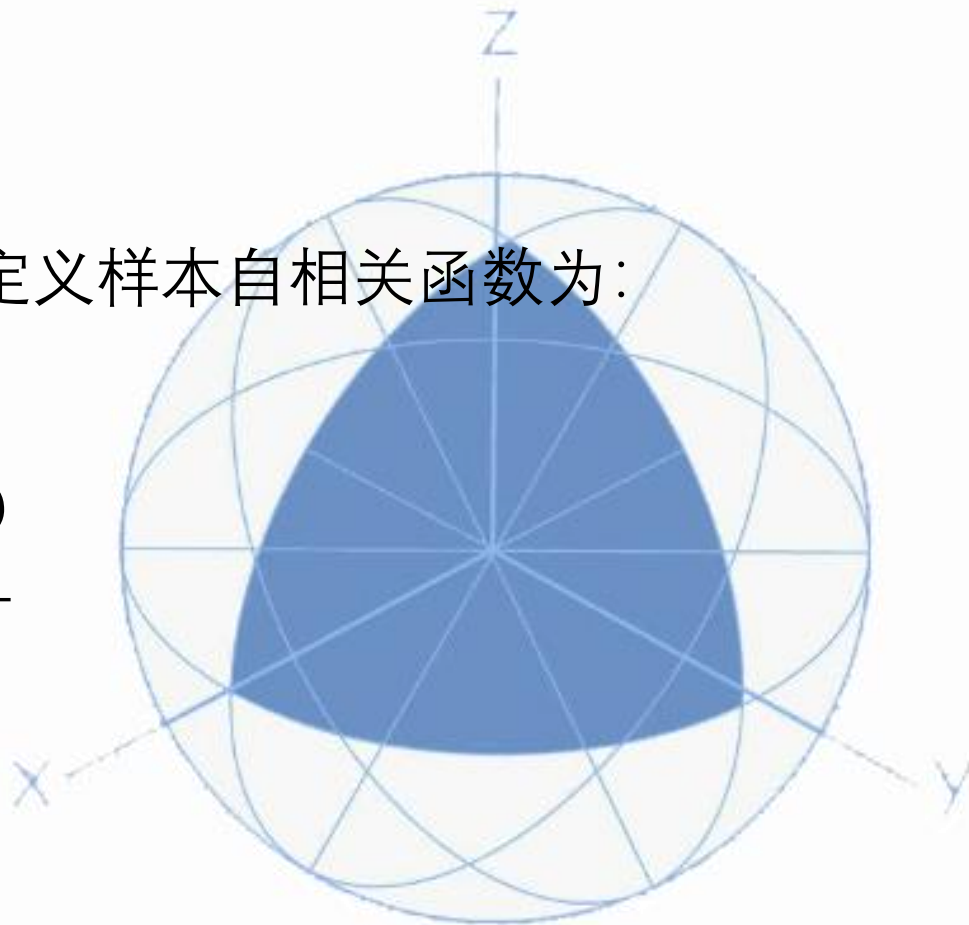
对于平稳时间序列而言,  $Y_t$  和  $Y_s$  的协方差对于时间的依赖只与时间间隔  $|t-s|$  有关, 而与具体的时刻  $t, s$  无关, 从而我们有

$$r_k = \text{Cov}(Y_t, Y_{t-k}), \rho_k = \text{Corr}(Y_t, Y_{t-k}) = \frac{r_k}{r_0}$$



对于观测序列  $Y_1, Y_2, \dots, Y_n$ , 定义样本自相关函数为:

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



#### 4、白噪声

通常我们所说的白噪声过程是指具有 0 均值，且独立同分布的一列随机变量序列 $\{e_t\}$ .显然白噪声过程是严平稳的，因为

$$\begin{aligned} & P(e_{t_1} \leq x_1, e_{t_2} \leq x_2, \cdots, e_{t_n} \leq x_n) \\ &= P(e_{t_1} \leq x_1)P(e_{t_2} \leq x_2) \cdots P(e_{t_n} \leq x_n) \\ &= P(e_{t_1-k} \leq x_1)P(e_{t_2-k} \leq x_2) \cdots P(e_{t_n-k} \leq x_n) \\ &= P(e_{t_1-k} \leq x_1, e_{t_2-k} \leq x_2, \cdots, e_{t_n-k} \leq x_n) \end{aligned}$$

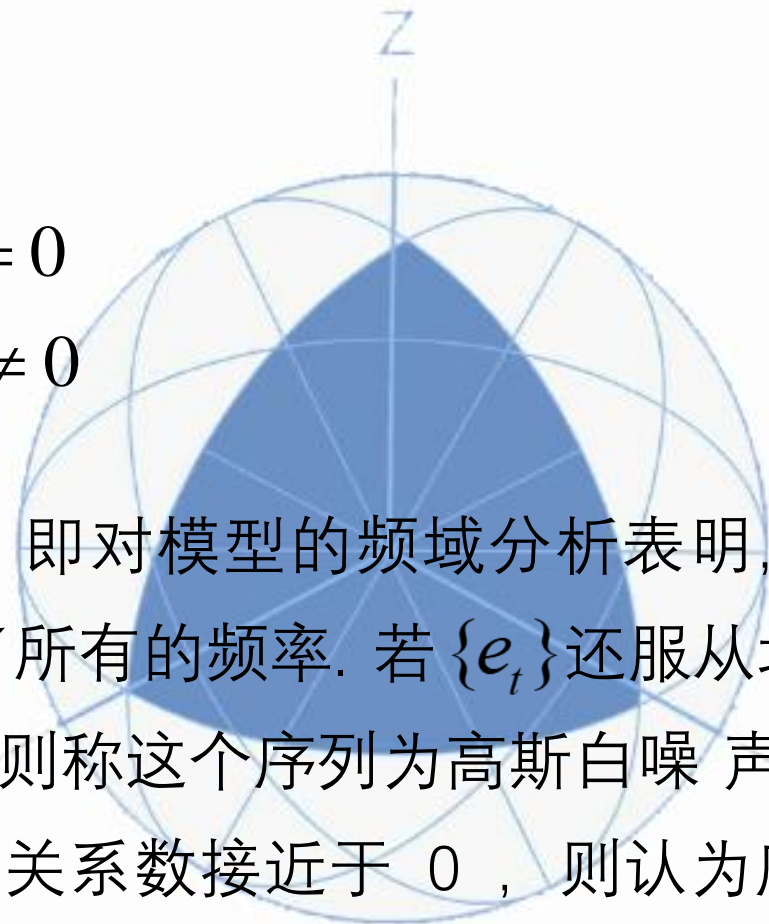




且自相关系数:

$$\rho_k = \begin{cases} 1 & k = 0 \\ 0 & k \neq 0 \end{cases}$$

白噪声这一术语来自如下事实，即对模型的频域分析表明，与白光类似，模型中平等的包含了所有的频率. 若  $\{e_t\}$  还服从均值为 0，方差为  $\sigma^2$  的正态分布，则称这个序列为高斯白噪声. 在实际应用中，如果所有样本自相关系数接近于 0，则认为序列是白噪声序列.





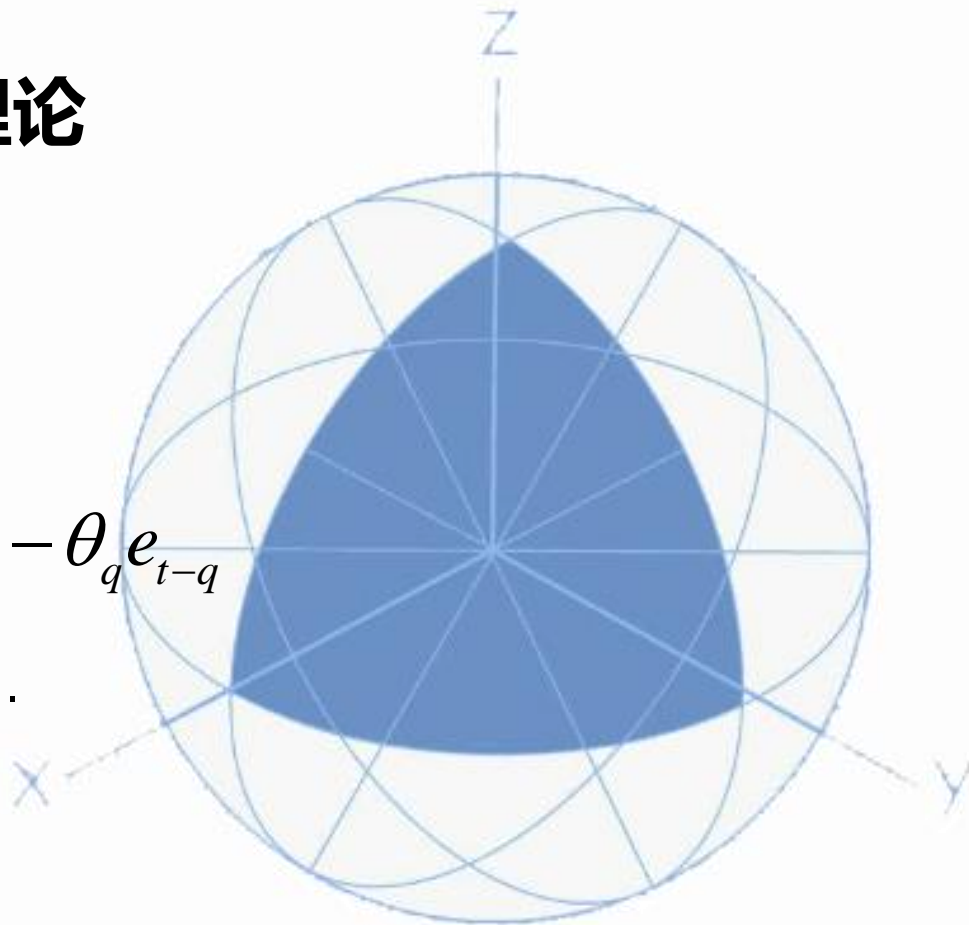
## 10.2.2 常见时间序列模型与理论

### 1、滑动平均模型 MA(q)

滑动平均模型的一般形式为：

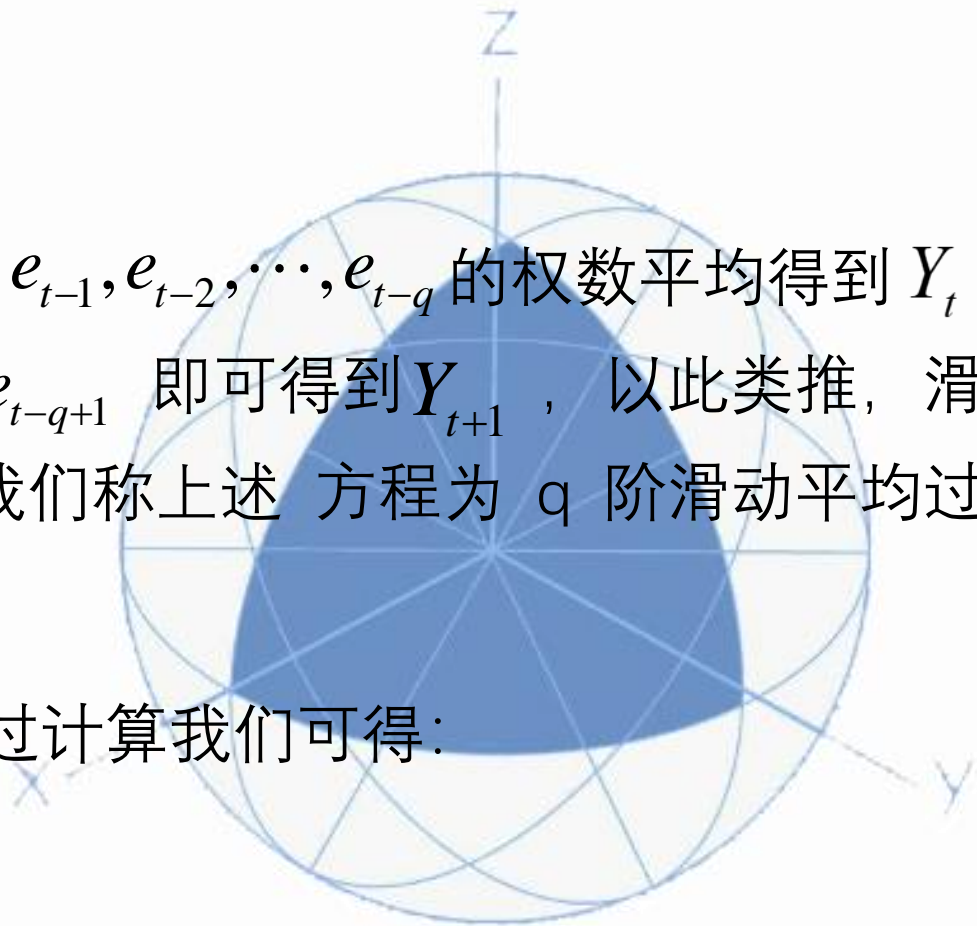
$$Y_t = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q}$$

其中  $e_t$  为独立的白噪声序列.



用  $1, -\theta_1, -\theta_2, \dots, -\theta_q$  作为  $e_t, e_{t-1}, e_{t-2}, \dots, e_{t-q}$  的权数平均得到  $Y_t$  ,  
再滑动权数至  $e_{t+1}, e_t, e_{t-1}, \dots, e_{t-q+1}$  即可得到  $Y_{t+1}$  , 以此类推, 滑  
动平均的术语也由此而来, 我们称上述 方程为  $q$  阶滑动平均过  
程, 简记为  $MA(q)$  .

对于一般的  $MA(q)$  过程, 通过计算我们可得:





$$E(Y_t) = 0, r_0 = \text{Var}(Y_t) = (1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2) \sigma^2$$

$$\rho_k = \begin{cases} \frac{-\theta_k + \theta_1 \theta_{k+1} + \theta_2 \theta_{k+2} + \cdots + \theta_{q-k} \theta_q}{1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2}, & k = 1, 2, \cdots, q \\ 0, & k > q \end{cases}$$

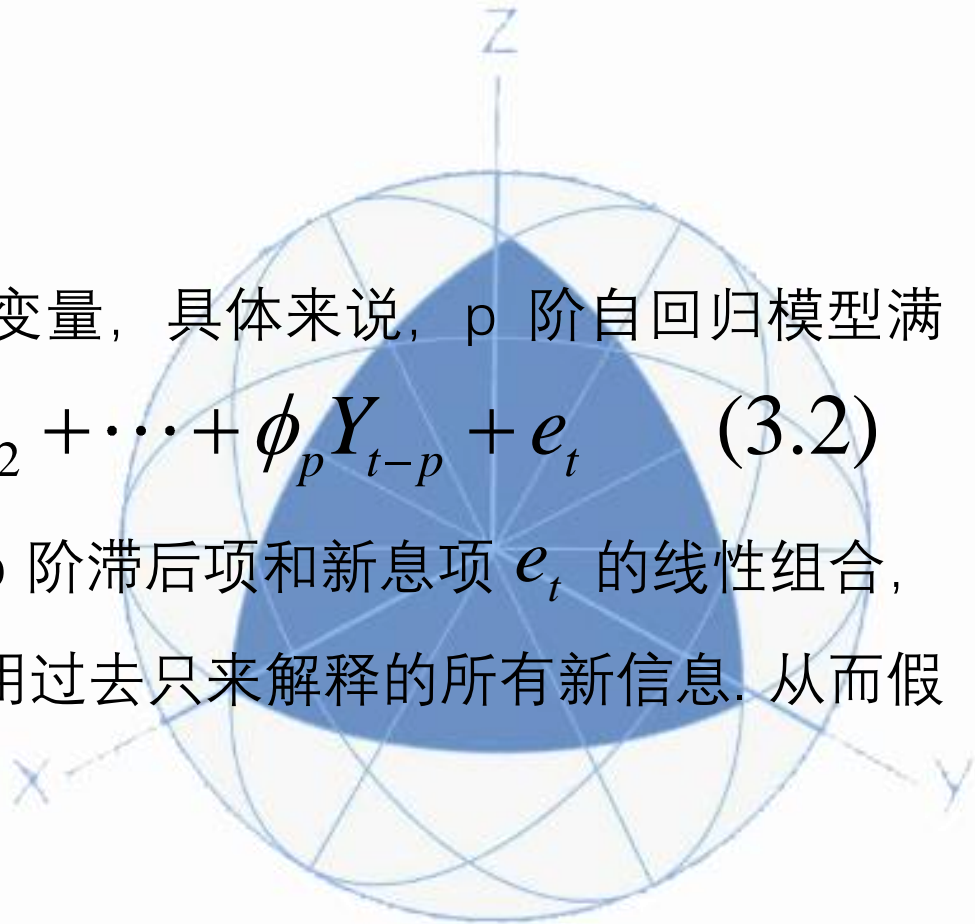
从这我们也可以看出 MA(q) 过程的自相关系数  $\rho_k$  在滞后 q 阶以后取值为零，我们称之为 q 阶截尾。



## 2、一般自回归模型 AR(p)

自回归模型就是用自身作为回归变量，具体来说，p 阶自回归模型满足表达式：
$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + e_t \quad (3.2)$$

即序列  $Y_t$  的当期值是自身最近 p 阶滞后项和新息项  $e_t$  的线性组合，其中  $e_t$  包括了序列在 t 期无法用过去只来解释的所有新信息。从而假设  $e_t$  独立于  $Y_{t-1}, Y_{t-2}, \cdots, Y_{t-p}$





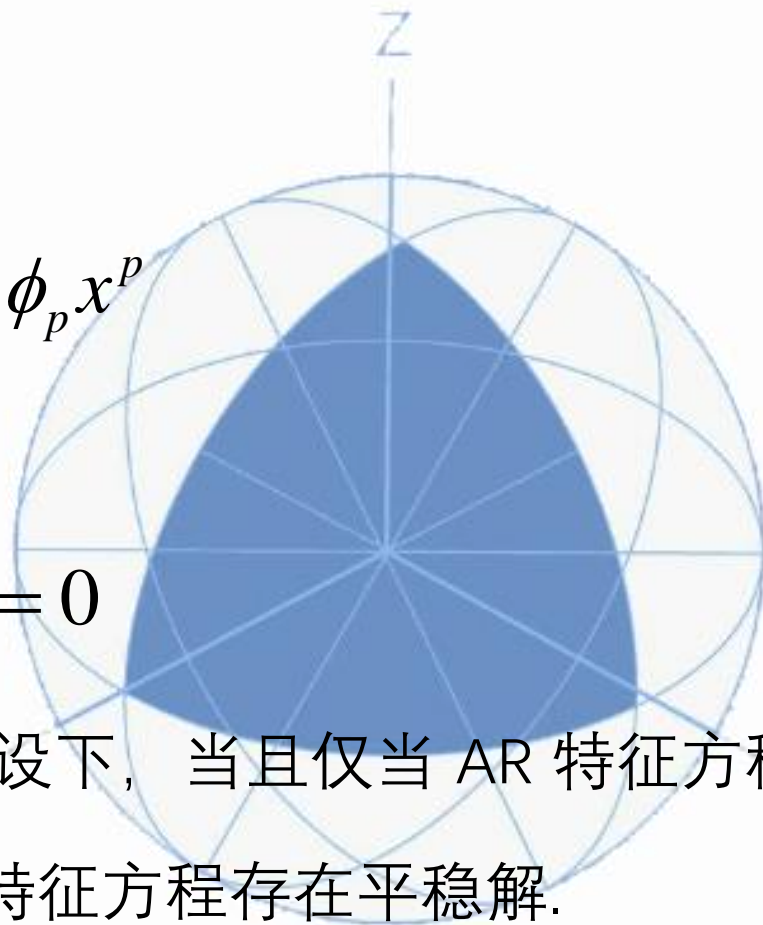
相应的 AR 特征多项式为:

$$\phi(x) = 1 - \phi_1 x - \phi_2 x^2 - \cdots - \phi_p x^p$$

从而相应的 AR 特征方程为:

$$1 - \phi_1 x - \phi_2 x^2 - \cdots - \phi_p x^p = 0$$

在  $e_t$  独立于  $Y_{t-1}, Y_{t-2}, \cdots, Y_{t-p}$  的假设下, 当且仅当 AR 特征方程每一个根的绝对值 (模) 都小于 1, 特征方程存在平稳解.





假设序列平稳并且均值为零，在 (3.2) 式两边同乘以  $Y_{t-k}$  并且期望然后除以  $r_0$ ，可得

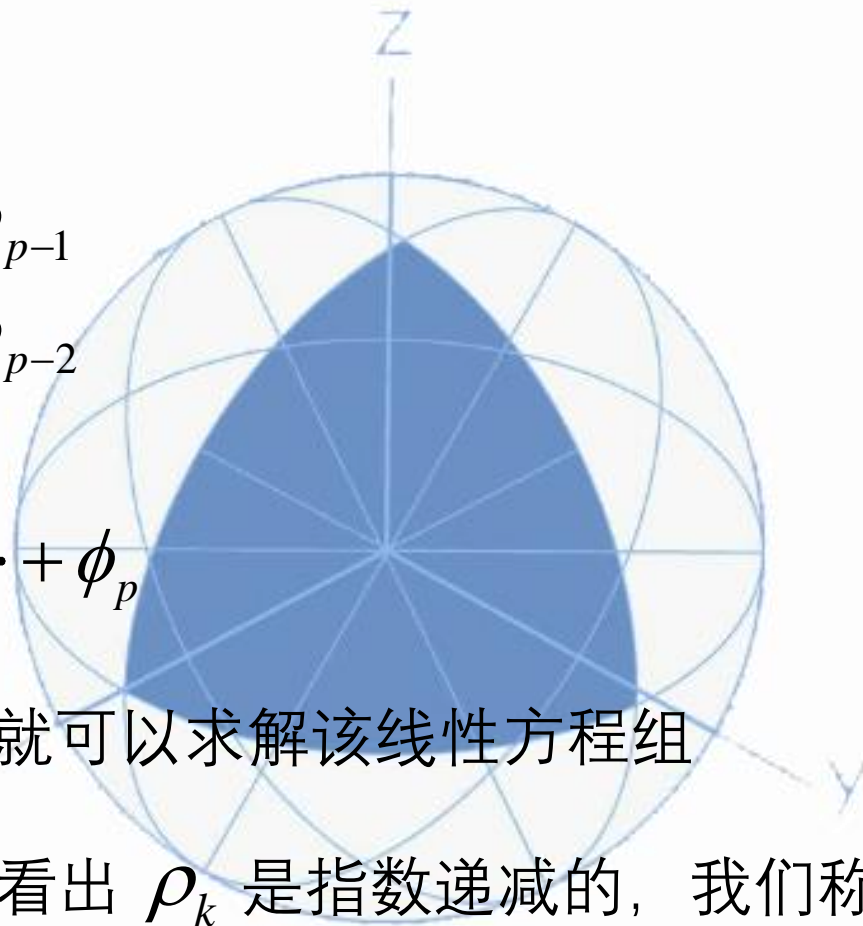
$$\rho_k = \phi_1 \rho_{k-1} + \phi_2 \rho_{k-2} + \cdots + \phi_p \rho_{k-p}, k \geq 1$$

把  $k = 1, 2, \dots, p$  代入，根据  $\rho_0 = 1, \rho_{-k} = \rho_k$  得到 Yule-Walker 方程组：



$$\begin{cases} \rho_1 = \phi_1 + \phi_2 \rho_1 + \phi_3 \rho_2 + \cdots + \phi_p \rho_{p-1} \\ \rho_2 = \phi_1 \rho_1 + \phi_2 + \phi_3 \rho_1 + \cdots + \phi_p \rho_{p-2} \\ \dots\dots\dots \\ \rho_p = \phi_1 \rho_{p-1} + \phi_2 \rho_{p-2} + \phi_3 \rho_{p-3} \cdots + \phi_p \end{cases}$$

当给定  $\phi_1, \phi_2, \dots, \phi_p$  的值, 我们就可以求解该线性方程组  
得到  $\rho_1, \rho_2, \dots, \rho_p$  的值. 我们可以看出  $\rho_k$  是指数递减的, 我们称  
之为 p 阶拖尾.



### 3、自回归滑动平均混合模型 ARMA(p,q)

如果序列中部分是自回归，部分是滑动平均，我们可以得到一个相当普遍的时间序列模型，其表达式如下所示：

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q}$$

记为 ARMA(p,q).



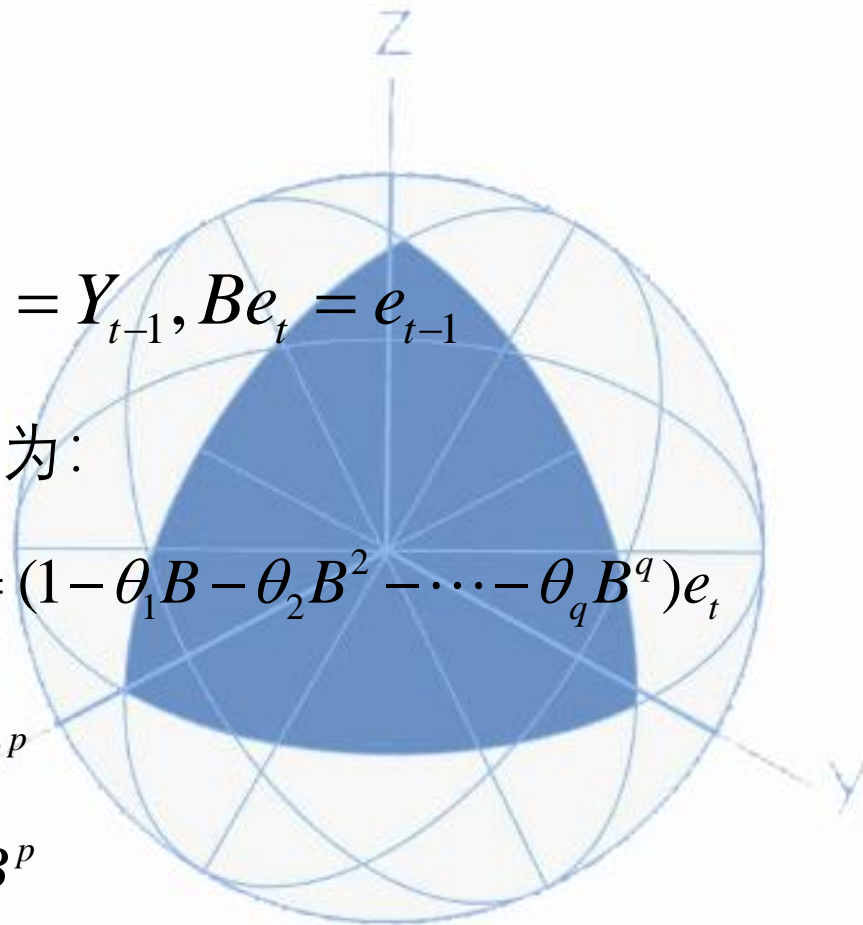
我们记延迟算子为  $B$ ，则有  $BY_t = Y_{t-1}, Be_t = e_{t-1}$

则 ARMA(p,q) 写成滞后算子形式为：

$$(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p) Y_t = (1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q) e_t$$

令  $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$

$$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$$





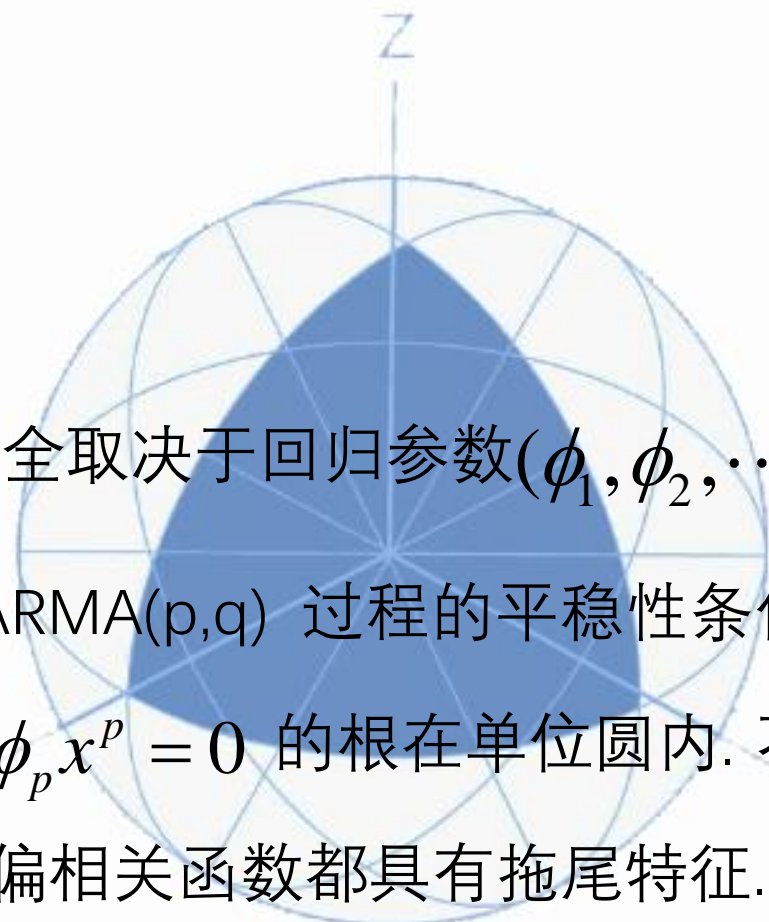
则有  $\phi(B)Y_t = \theta(B)e_t$

因为 ARMA(p,q) 过程的平稳性完全取决于回归参数  $(\phi_1, \phi_2, \dots, \phi_p)$

而与移动平均参数均无关, 即 ARMA(p,q) 过程的平稳性条件为

特征方程  $1 - \phi_1 x - \phi_2 x^2 - \dots - \phi_p x^p = 0$  的根在单位圆内. 不过

ARMA(p,q) 过程的自相关函数和偏相关函数都具有拖尾特征.



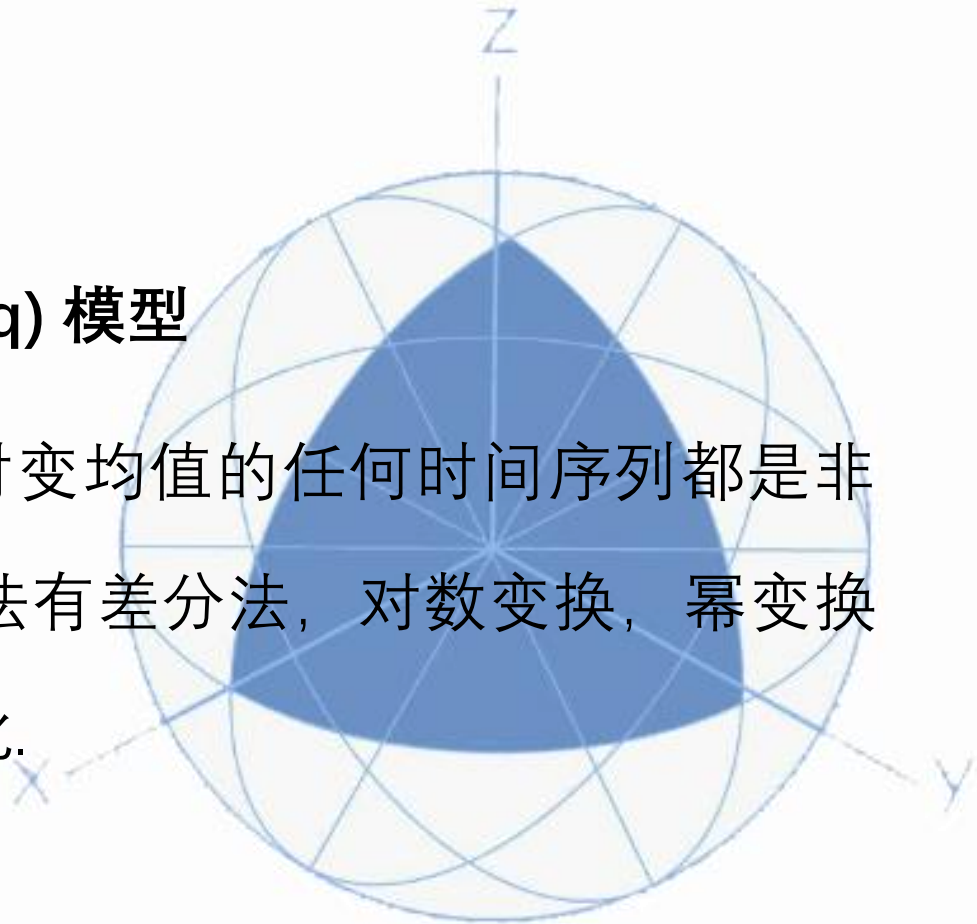
综上，我们将时间序列常见的这三种模型的性质汇总如下表所示：

	$AR(p)$	$MA(q)$	$ARMA(p, q)$
模型方程	$\phi_p(B)Y_t = e_t$	$Y_t = \theta_q(B)e_t$	$\phi_p(B)Y_t = \theta_q(B)e_t$
自相关 acf	拖尾	q 阶截尾	拖尾
偏自相关 pacf	p 阶截尾	拖尾	拖尾



#### 4、非平稳时间序列 ARIMA(p,d,q) 模型

由平稳性的条件我们可知具有时变均值的任何时间序列都是非平稳的，常见的平稳化处理 方法有差分法，对数变换，幂变换等. 我们这里主要介绍差分平稳化.





若一时间序列  $Y_t$  的  $d$  次差分  $W_t = \nabla^d Y_t$  是一个平稳的 ARMA 过程, 且  $W_t$  服从 ARMA(p,q) 模型, 我们就称  $Y_t$  是 ARIMA(p,d,q) 过程. 通常情况下  $d = 1$  或者最多取到 2, 太高次的差分可能会引起过度差分.

根据延迟算子  $B$ , 我们有  $\nabla Y_t = Y_t - Y_{t-1} = (1 - B)Y_t$

同样的二次差分可以表示为  $\nabla^2 Y_t = (1 - B)^2 Y_t$



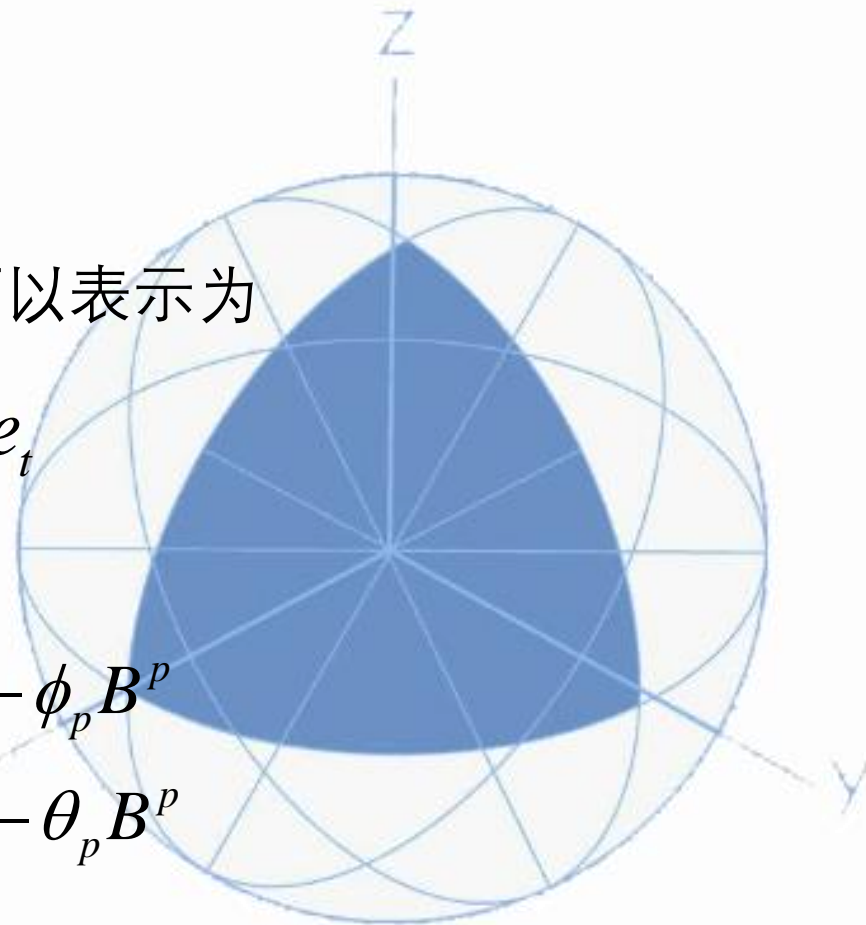
从而一般的 ARIMA(p,d,q) 模型可以表示为

$$\phi(B)(1-B)^d Y_t = \theta(B)e_t$$

其中

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p$$

$$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \dots - \theta_q B^q$$





## 5、时间序列一般建模过程

- (1) 数据预处理：包括无量纲化处理，缺失值处理以及异常值处理等.
- (2) 平稳性检验：时序图直观观察，自相关系数呈衰减趋势，ADF 检验等. 常用的平稳化方法有对数处理，差分处理以及幂变换等.



(3) 模型定阶：主要通过模型的样本自相关图线以及偏自相关图线的截尾特征进行模型阶数的确定。

(4) 参数估计：常用的参数估计方法有矩估计，最小二乘估计和极大似然估计等。

(5) 模型诊断：主要包括检验残差的正态性以及自相关性

(6) 模型预测：通过检验的之后的模型就可以用来做预测了，主要的预测方法是最小均方误差预测。

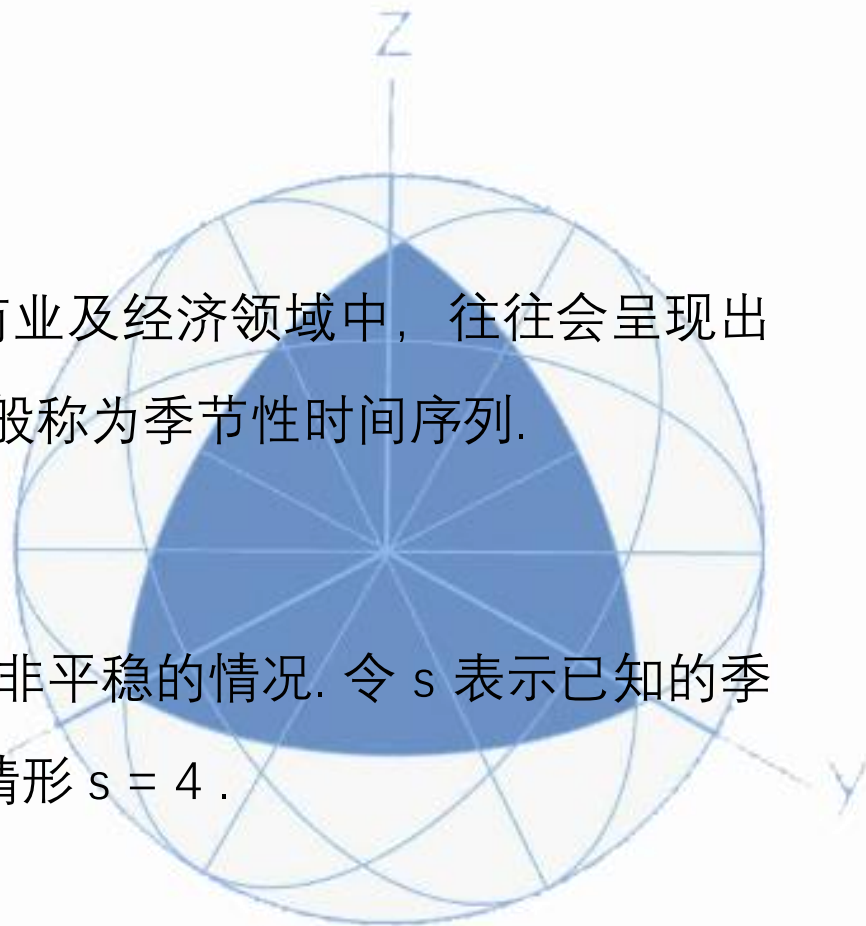


## 10.2.3 季节模型

在应用时间序列的很多领域中，尤其是在商业及经济领域中，往往会呈现出一定的循环和 周期趋势. 这样的序列我们一般称为季节性时间序列.

### 1、平稳季节 ARMA 模型

我们先来研究平稳模型，后面将进一步考虑非平稳的情况. 令  $s$  表示已知的季节周期；月 度序列情况  $s = 12$ ，季度序列情形  $s = 4$  .





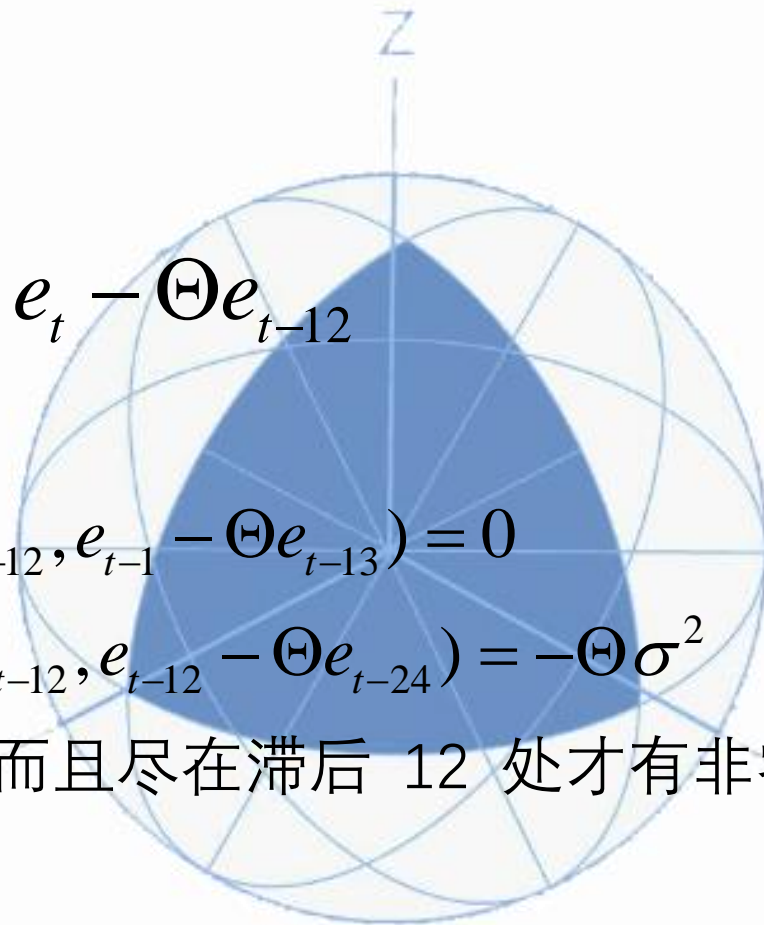
考虑如下生成的时间序列:  $Y_t = e_t - \Theta e_{t-12}$

从而

$$\text{Cov}(Y_t, Y_{t-1}) = \text{Cov}(e_t - \Theta e_{t-12}, e_{t-1} - \Theta e_{t-13}) = 0$$

$$\text{Cov}(Y_t, Y_{t-12}) = \text{Cov}(e_t - \Theta e_{t-12}, e_{t-12} - \Theta e_{t-24}) = -\Theta \sigma^2$$

很容易看出, 该序列是平稳的, 而且仅在滞后 12 处才有非零的自相关性.



根据这些想法，我们可以定义季节周期为  $s$  的  $Q$  阶季节 MA( $Q$ ) 模型如下：
$$Y_t = e_t - \Theta_1 e_{t-s} - \Theta_2 e_{t-2s} - \cdots - \Theta_Q e_{t-Qs}$$

其季节 MA 特征多项式为 
$$\Theta(x) = 1 - \Theta_1 x - \Theta_2 x^{2s} - \cdots - \Theta_Q x^{Qs}$$

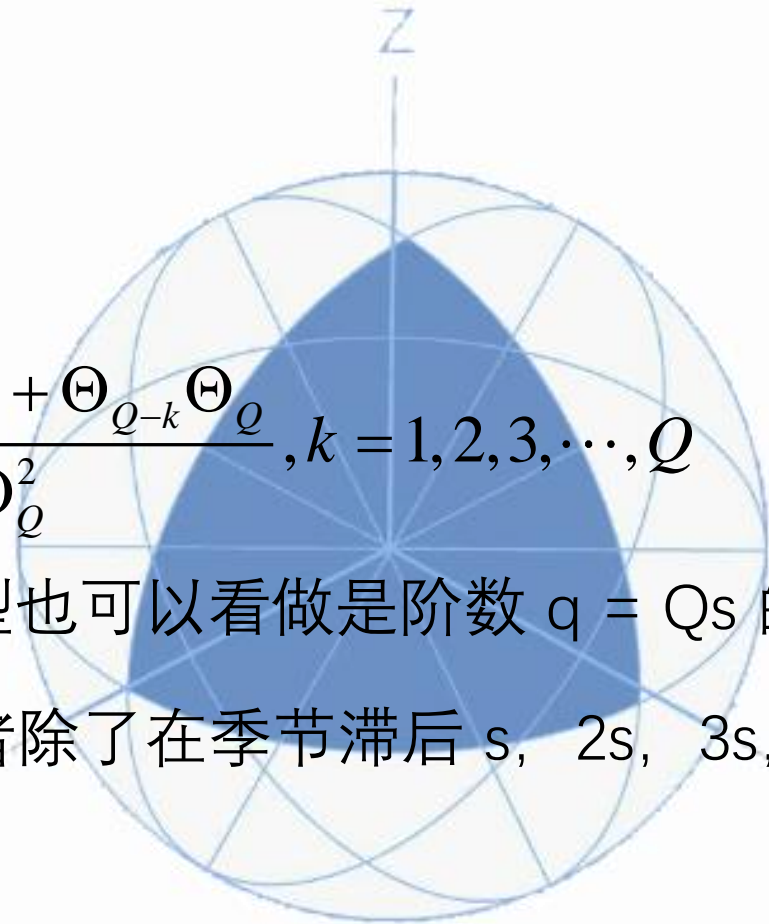
显然，该序列总是平稳的，且其自相关函数只在  $s, 2s, 3s, \dots, Qs$  等季节滞后上非零.



特別的,

$$\rho_{ks} = \frac{-\Theta_k + \Theta_1 \Theta_{k+1} + \Theta_2 \Theta_{k+2} + \cdots + \Theta_{Q-k} \Theta_Q}{1 + \Theta_1^2 + \Theta_2^2 + \cdots + \Theta_Q^2}, k = 1, 2, 3, \dots, Q$$

需要注意的是, 季节 MA(Q) 模型也可以看做是阶数  $q = Qs$  的非季节性 MA 模型的特例, 但后者除了在季节滞后  $s, 2s, 3s, \dots, Qs$  处, 所有的  $\theta$  值都取零.

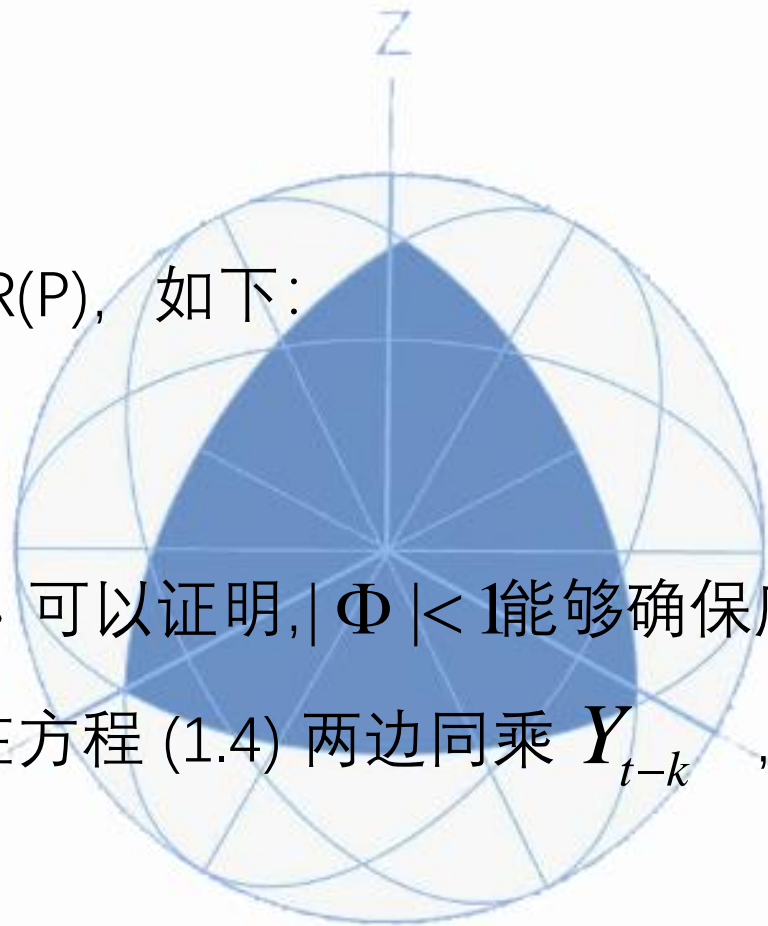




同样可以定义季节自回归模型 AR(P), 如下:

$$Y_t = \Phi Y_{t-12} + e_t \quad (1.4)$$

其中  $|\Phi| < 1$ ,  $e_t$  独立于  $Y_{t-1}, Y_{t-2}, \dots$  可以证明,  $|\Phi| < 1$  能够确保序列是平稳的, 故易证  $E(Y_t) = 0$ ; 在方程 (1.4) 两边同乘  $Y_{t-k}$ , 取期望值, 并除以  $r_0$ , 得到



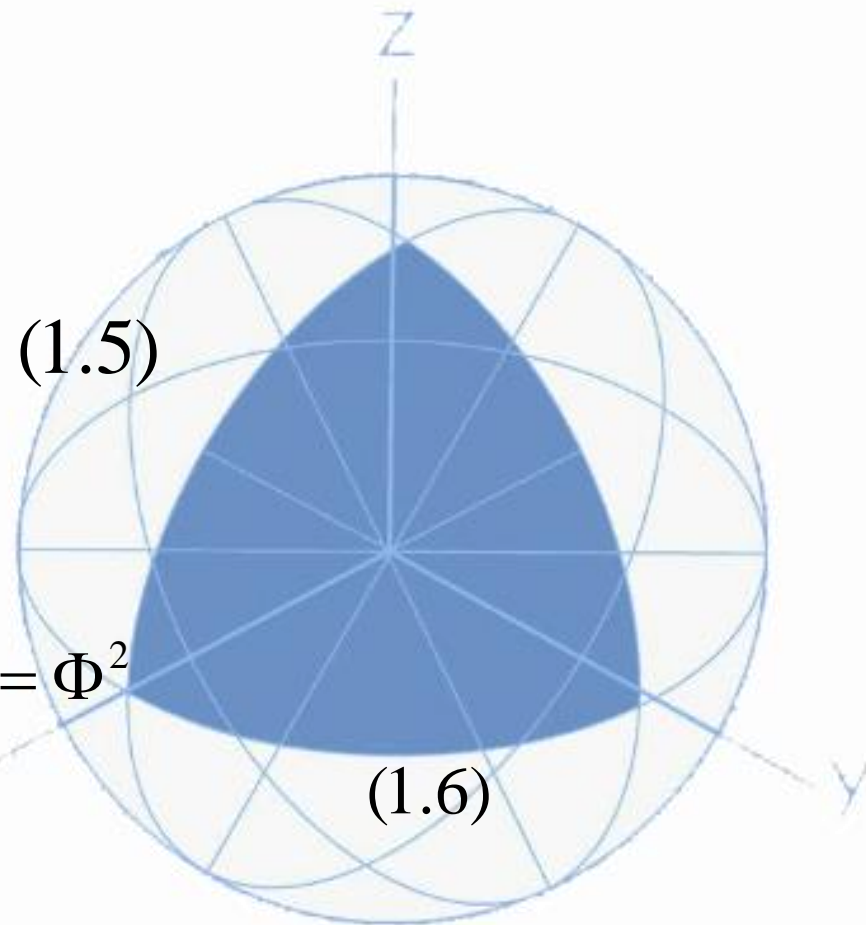


$$\rho_k = \Phi \rho_{k-12}, k \geq 1 \quad (1.5)$$

易得,

$$\rho_{12} = \Phi \rho_0 = \Phi, \rho_{24} = \Phi \rho_{12} = \Phi^2$$

$$\rho_{12k} = \Phi^k, k = 1, 2, \dots$$





进一步，先后去方程 (1.5) 中的  $k$  分别为 1 和 11，并  
利用  $\rho_k = \rho_{-k}$ ，得到  $\rho_1 = \Phi \rho_{11}, \rho_{11} = \Phi \rho_1$   
则我们可证，除了在季节之后 12, 24, 36, ... 处以  
外， $\rho_k = 0$  总成立. 在之后处取以上值时，自相关函数如  
同 AR(1) 模型一样以指数方式衰减.

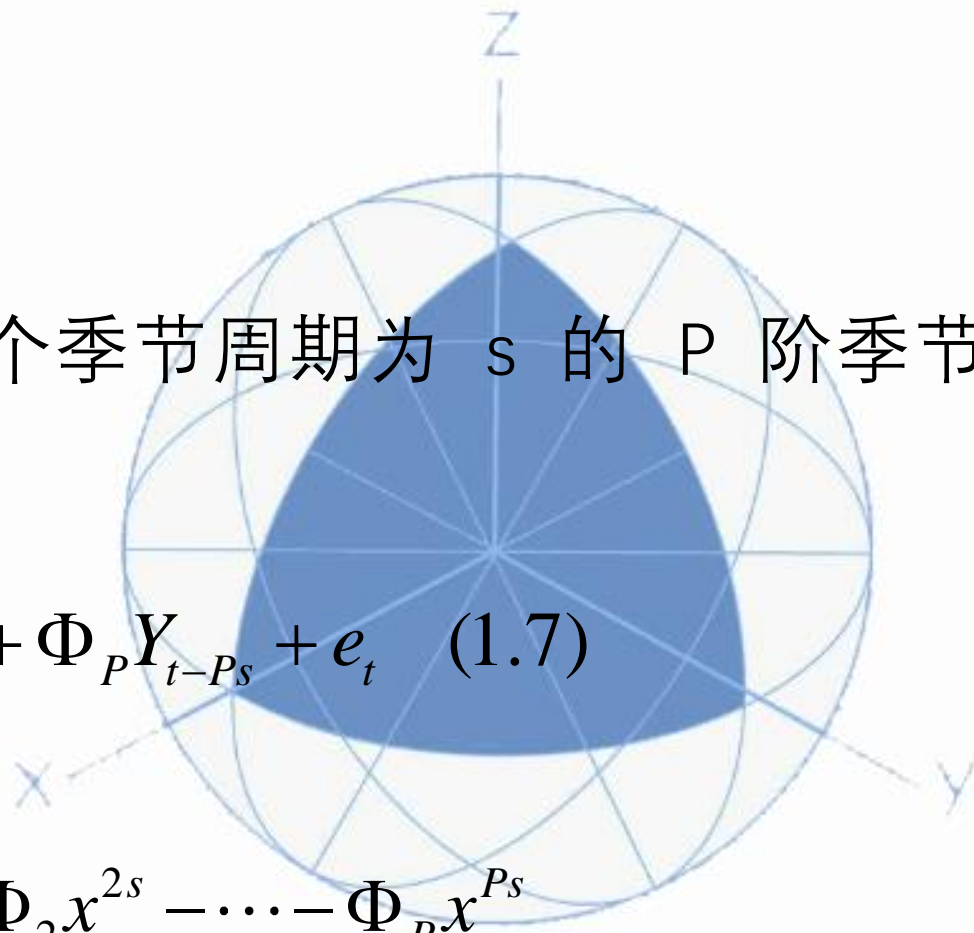


记住这个例子，定义一个季节周期为  $s$  的  $P$  阶季节  
AR( $P$ ) 模型如下：

$$Y_t = \Phi_1 Y_{t-s} + \Phi_2 Y_{t-2s} + \cdots + \Phi_P Y_{t-Ps} + e_t \quad (1.7)$$

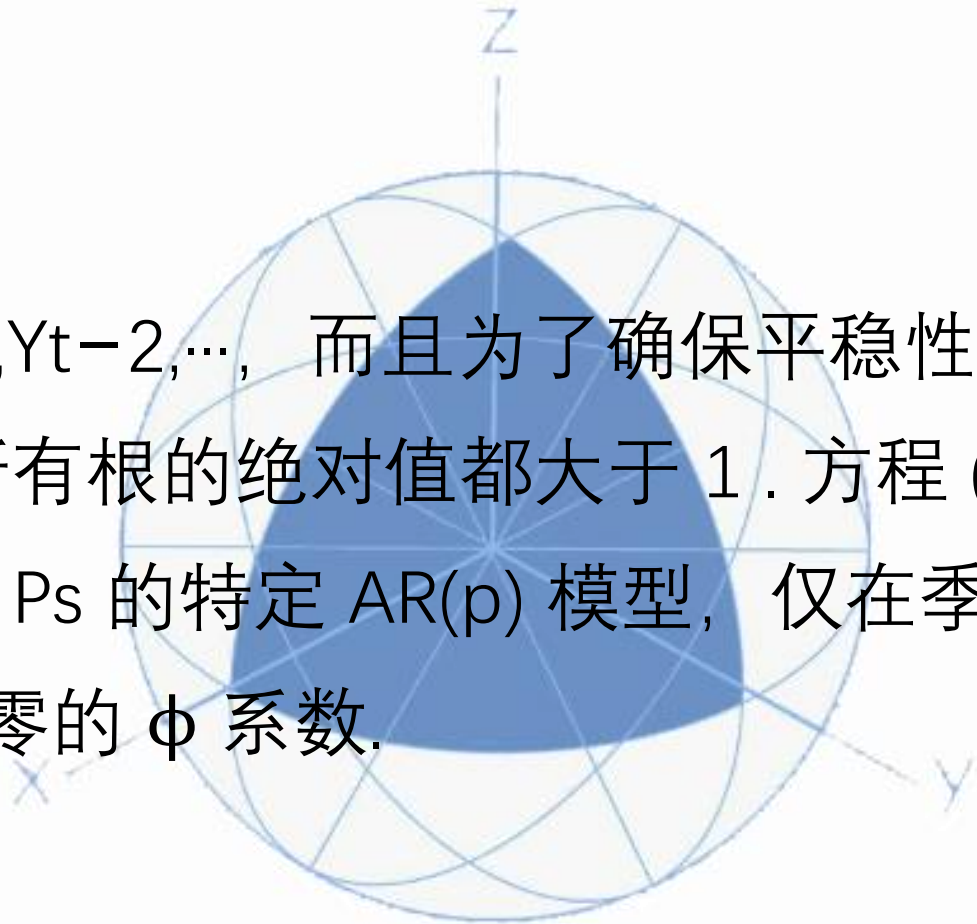
其特征多项式为

$$\Phi(x) = 1 - \Phi_1 x^s - \Phi_2 x^{2s} - \cdots - \Phi_P x^{Ps}$$





通常要求  $e_t$  独立于  $Y_{t-1}, Y_{t-2}, \dots$ , 而且为了确保平稳性, 要求特征方程  $\Phi(x) = 0$  的所有根的绝对值都大于 1. 方程 (1.7) 可以看做是一个阶数  $p = P_s$  的特定  $AR(p)$  模型, 仅在季节滞后  $s, 2s, 3s, \dots, P_s$  处才有非零的  $\phi$  系数.



可以证明，自相关函数仅在滞后  $s, 2s, 3s, \dots$  处非零，其行为很像是指数衰减函数和阻尼正弦函数的组合。特别地，方程 (1.4)、方程 (1.5) 和方程 (1.6) 易推广到广义季节 AR(1) 模型，此时  $\rho_{ks} = \Phi^k, k=1, 2, \dots$  在滞后为其他值时是零相关的。

## 2、平稳乘法季节 ARIMA 模型

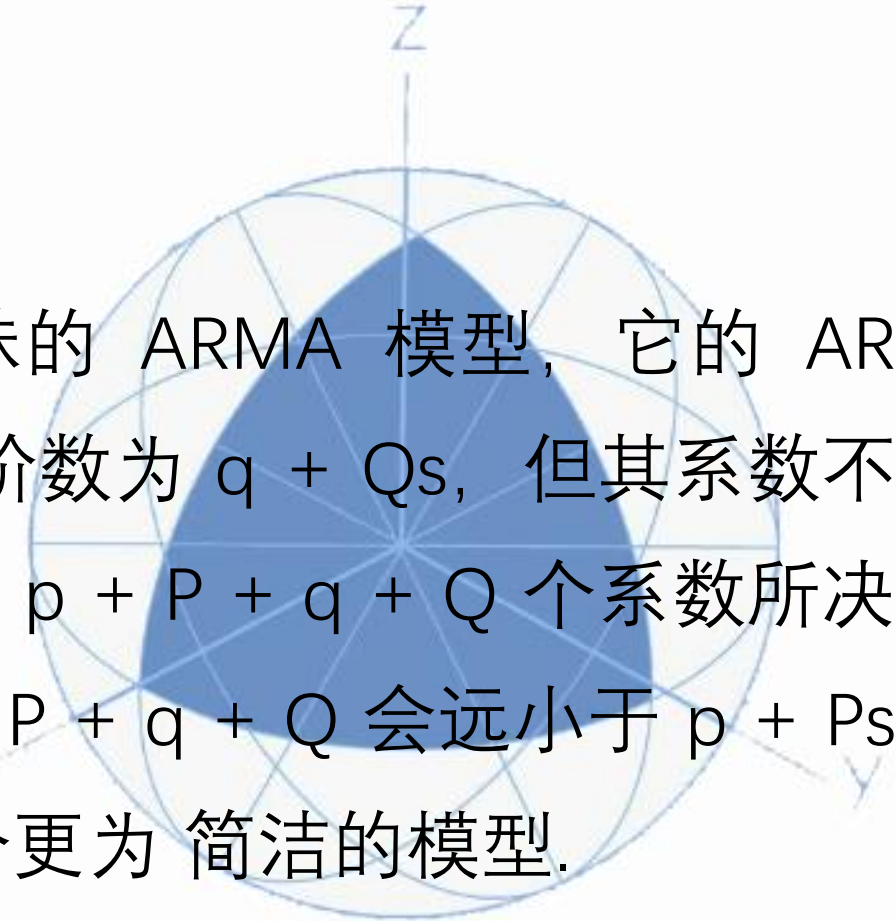
我们几乎不需要那些仅在季节性滞后上包含自相关性的模型. 结合在季节与非季节 ARMA模型上的思考, 可以构造出一类简洁的模型, 它们不仅在季节滞后上包含相关性, 而且在邻近序列值的更小滞后上亦然.



下面我们定义周期为  $s$  的乘法季节 ARMA( $p, q$ )  $\times$  ( $P, Q$ ) ,  
模型是 AR 特征多项式为  $\phi(x)\Phi(x)$ , MA 特征多项式为

$\Theta(x)\theta(x)$  的模型, 其中

$$\begin{cases} \phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p \\ \Phi(B) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \dots - \Phi_P B^{Ps} \end{cases} \quad \text{与} \quad \begin{cases} \theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \dots - \theta_q B^q \\ \Theta(B) = 1 - \Theta_1 B^s - \Theta_2 B^{2s} - \dots - \Theta_Q B^{Qs} \end{cases}$$



再次注意，这只是一个特殊的 ARMA 模型，它的 AR 项阶数为  $p + P_s$ ，MA 项阶数为  $q + Q_s$ ，但其系数不具有完全的一般性，而仅由  $p + P + q + Q$  个系数所决定。若  $s = 12$ ，则由于  $p + P + q + Q$  会远小于  $p + P_s + q + Q_s$ ，因此将得到一个更为简洁的模型。



### 3、非平稳季节 ARIMA 模型

季节差分是非平稳季节过程建模的一个重要工具，时间序列  $Y_t$  的周期为  $s$  的季节差分用  $\nabla_s Y_t$  表示，定义如下：
$$\nabla_s Y_t = Y_t - Y_{t-s} \quad (3.1)$$

例如，考虑月度序列 1 月至 1 月，2 月至 2 月，... 相继年份中的数据变化情况。注意，长度为  $n$  的序列其季节差分序列的长度为  $n-s$ ，也就是说，季节差分后丢失了  $s$  个数据值。

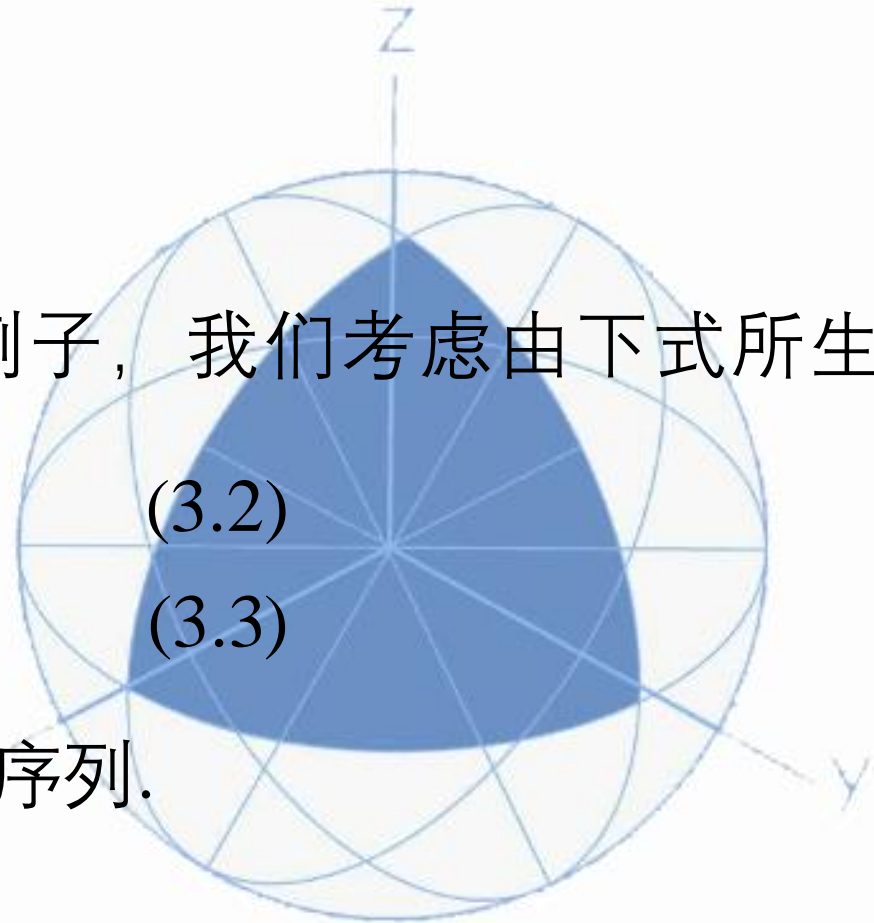


作为使用季节差分的一个例子，我们考虑由下式所生产的一个过程

$$Y_t = S_t + e_t \quad (3.2)$$

$$S_t = S_{t-s} + \varepsilon_t \quad (3.3)$$

其中  $\varepsilon_t, e_t$  是独立的白噪声序列.



这里的  $S_t$  是“季节随机游动序列”. 由于  $S_t$  是非平稳的, 显然  $Y_t$  也是非平稳的. 然后, 若按照 (3.1) 那样对  $Y_t$  作季节差分变换, 则将得到

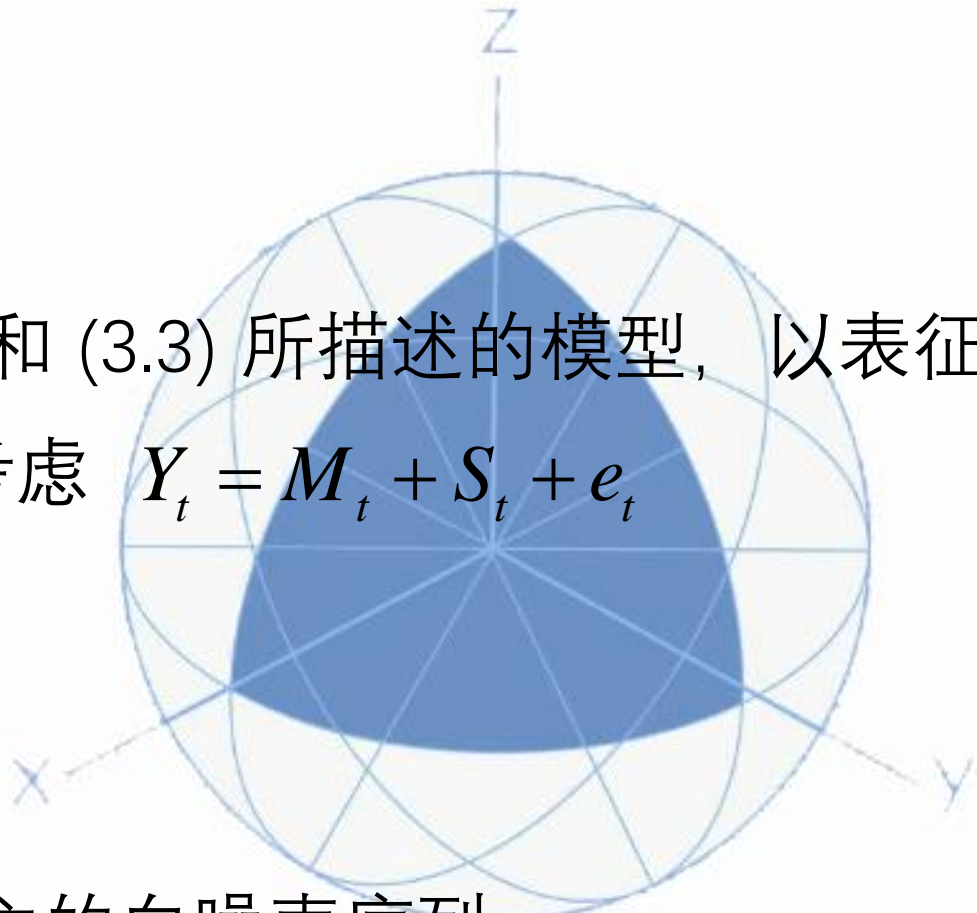
$$\nabla_s Y_t = S_t - S_{t-s} + e_t - e_{t-s} = \varepsilon_t + e_t - e_{t-s}$$

经过简单计算可知  $\nabla_s Y_t$  是平稳的, 且其自相关函数与 MA(1) 模型相同.

也可以推广方程组 (3.2) 和 (3.3) 所描述的模型, 以表征非季节慢变随即趋势, 考虑  $Y_t = M_t + S_t + e_t$

其中  $S_t = S_{t-s} + \varepsilon_t$   
 $M_t = M_{t-1} + \xi_t$

其中  $e_t, \varepsilon_t, \xi_t$  是相互独立的白噪声序列.



这里将同时应用季节差分和普通的非季节差分来得到

$$\begin{aligned}\nabla\nabla_s Y_t &= \nabla(M_t - M_{t-s} + \varepsilon_t + e_t - e_{t-s}) \\ &= (\xi_t + \varepsilon_t + e_t) - (\varepsilon_{t-1} + e_{t-1}) - (\xi_{t-s} + e_{t-s}) + e_{t-s-1}\end{aligned}$$

这里所定义的过程是平稳的，且只在  $1, s-1, s, s+1$  等滞后上有非零自相关性，这同季节周期为  $s$  的乘法季节模型  $\text{ARMA}(p,q) \times (P,Q)$  的自相关结构是一致的。





由此我们可以引出非平稳季节模型的定义，过程  $Y_t$  称为季节周期为  $s$ 、非季节阶数为  $p$ ， $d$  和  $q$ ，季节阶数为  $P$ ， $D$  和  $Q$  的乘法季节 ARIMA 模型，前提是差分序列

$$W_t = \nabla^d \nabla_s^D Y_t$$

满足某季节周期为  $s$  的  $ARMA(p, q) \times (P, Q)_s$  模型， $Y_t$  称为季节周期为  $s$  的  $ARIMA(p, d, q) \times (P, D, Q)_s$  模型.



#### 4、季节模型的预测

计算季节  $ARIMA$  模型预测最简单的方法是对模型递归的应用差分方程形式，例如考虑  $ARIMA(0, 1, 1) \times (1, 0, 1)_{12}$  模型，

$$Y_t - Y_{t-1} = \Phi(Y_{t-12} - Y_{t-13}) + e_t - \theta e_{t-1} - \Theta e_{t-12} + \theta\Theta e_{t-13} \quad (4.1)$$

可改写为

$$Y_t = Y_{t-1} + \Phi Y_{t-12} - \Phi Y_{t-13} + e_t - \theta e_{t-1} - \Theta e_{t-12} + \theta\Theta e_{t-13} \quad (4.2)$$

则我们可向前预测，从  $t$  点出发向前预测，

$$\hat{Y}_t(1) = Y_t + \Phi Y_{t-11} - \Phi Y_{t-12} - \theta e_t - \Theta e_{t-11} + \theta\Theta e_{t-12} \quad (4.3)$$

下一个为

$$\hat{Y}_t(2) = \hat{Y}_t(1) + \Phi Y_{t-10} - \Phi Y_{t-11} - \Theta e_{t-10} + \theta\Theta e_{t-11} \quad (4.4)$$

并以此类推，在前置时刻  $l = 1, 2, 3, \dots, 13$  以上，其中  $e_{t-13}, e_{t-12}, e_{t-11}, \dots, e_t$  为噪声项，但对于  $l > 13$ ，将代之以模型自回归部分，即

$$\hat{Y}_t(l) = \hat{Y}_t(l-1) + \Phi \hat{Y}_t(l-12) - \Phi \hat{Y}_t(l-13), l > 13 \quad (4.5)$$

为了便于理解模型预测的一般性质，我们先来考虑  $AR(1)_{12}$ ，和  $MA(1)_{12}$  这两个特殊的模型。以及乘积模型  $ARIMA(0, 0, 0) \times (0, 1, 1)_{12}$  和  $ARIMA(0, 1, 1) \times (0, 1, 1)_{12}$



$$Y_t = \Phi Y_{t-12} + e_t \quad (4.6)$$

很明显地, 有

$$\hat{Y}_t(l) = \Phi \hat{Y}_t(l-12) \quad (4.7)$$

然而, 对  $l$  向后迭代, 也可得出

$$\hat{Y}_t(l) = \Phi^{k+1} Y_{t+r-11} \quad (4.8)$$

这里的  $k$  和  $r$  是由式子  $l = 12k + r + 1$  所定义的, 其中  $0 \leq r < 12$  且  $k = 0, 1, 2, \dots$ . 也就是说  $k$  是  $(l-1)/12$  的整数部分, 而  $r/12$  是  $(l-1)/12$  的小数部分. 如果最后的观测点是 12 月份, 那么下一期 1 月份的预测值为  $\Phi$  乘以 1 月份的最后观测值, 2 月预测值为  $\Phi$  乘以 2 月份的最后观测值, 以此类推. 后推两期的 1 月预测值为  $\Phi^2$  乘以 1 月份的最后观测值. 当只关注 1 月份的数据可以发现, 未来预测值以指数方式衰减, 衰减速度取决于  $\Phi$  的大小.

当且仅当  $j$  为 12 倍数时,  $\psi$  权重非零, 即

$$\psi_j = \begin{cases} \Phi^{\frac{j}{12}} & j = 0, 12, 24, \dots \\ 0 & \text{others} \end{cases} \quad (4.9)$$

预测误差方差可以写为

$$\text{Var}(e_t(l)) = \left[ \frac{1 - \Phi^{2k+2}}{1 - \Phi^2} \right] \sigma_e^2 \quad (4.10)$$

其中同上面所说,  $k$  是  $(l-1)/12$  的整数部分.



对于季节  $MA(1)_{12}$  模型, 有

$$Y_t = e_t - \Theta e_{t-12} + \theta_0 \quad (4.11)$$

向前预测可以看到

$$\left. \begin{aligned} \hat{Y}_t(1) &= -\Theta e_{t-11} + \theta_0 \\ \hat{Y}_t(2) &= -\Theta e_{t-10} + \theta_0 \\ &\vdots \\ \hat{Y}_t(12) &= -\Theta e_t + \theta_0 \end{aligned} \right\} \quad (4.12)$$

且

$$\hat{Y}_t(l) = \theta_0, l > 12 \quad (4.13)$$

这里得出了第一年中个月份的不同预测值, 但是从这以后的所有预测都将由过程均值给出.

对该模型而言,  $\psi_0 = 1$ ,  $\psi_{12} = -\Theta$ , 其余情况下  $\psi_j = 0$ . 我们可以得到



$$\text{Var}(e_t(l)) = \begin{cases} \sigma_e^2 & 1 \leq l \leq 12 \\ (1 + \Theta^2)\sigma_e^2 & 12 < l \end{cases} \quad (4.12)$$

接下来我们考虑两个非平稳的乘法季节  $ARIMA$  模型,  $ARIMA(0, 1, 1) \times (0, 1, 1)_{12}$  和  $ARIMA(0, 0, 0) \times (0, 1, 1)_{12}$  模型.

$ARIMA(0, 1, 1) \times (0, 1, 1)_{12}$  模型为

$$Y_t = Y_{t-1} + Y_{t-12} - Y_{t-13} + e_t - \theta e_{t-1} - \Theta e_{t-12} + \theta \Theta e_{t-13} \quad (4.13)$$

预测值满足

$$\left. \begin{aligned} \hat{Y}_t(1) &= Y_t + Y_{t-11} - Y_{t-12} - \theta e_t - \Theta e_{t-11} + \theta \Theta e_{t-12} \\ \hat{Y}_t(2) &= \hat{Y}_t(1) + Y_{t-10} - Y_{t-11} - \Theta e_{t-10} + \theta \Theta e_{t-11} \\ &\vdots \\ \hat{Y}_t(12) &= \hat{Y}_t(11) + Y_t - Y_{t-1} - \Theta e_t + \theta \Theta e_{t-1} \\ \hat{Y}_t(13) &= \hat{Y}_t(12) + \hat{Y}_t(1) - Y_t + \theta \Theta e_t \end{aligned} \right\} \quad (4.14)$$

且

$$\hat{Y}_t(l) = \hat{Y}_t(l-1) + \hat{Y}_t(l-12) - \hat{Y}_t(l-13), l > 13 \quad (4.15)$$





$ARIMA(0,0,0) \times (0,1,1)_{12}$  模型为

$$Y_t - Y_{t-12} = e_t - \Theta e_{t-12} \quad (4.16)$$

因此进行迭代可得

$$\left. \begin{aligned} \hat{Y}_t(1) &= Y_{t-11} - \Theta e_{t-11} \\ \hat{Y}_t(2) &= Y_{t-10} - \Theta e_{t-10} \\ &\vdots \\ \hat{Y}_t(12) &= Y_t - \Theta e_t \end{aligned} \right\} \quad (4.17)$$

继而

$$\hat{Y}_t(l) = \hat{Y}_t(l-12), l > 12 \quad (4.18)$$

从而得到所有 1 月份的预测, 同样得到所有 2 月份的预测等.

如果反转此模型, 得到

$$Y_t = (1 - \Theta)(Y_{t-12} + \Theta Y_{t-24} + \Theta^2 Y_{t-36} + \cdots) + e_t$$

结果可以写出

$$\left. \begin{aligned} \hat{Y}_t(1) &= (1 - \Theta) \sum_{j=0}^{\infty} \Theta^j Y_{t-11-12j} \\ \hat{Y}_t(2) &= (1 - \Theta) \sum_{j=0}^{\infty} \Theta^j Y_{t-10-12j} \\ &\vdots \\ \hat{Y}_t(12) &= (1 - \Theta) \sum_{j=0}^{\infty} \Theta^j Y_{t-12j} \end{aligned} \right\} \quad (4.19)$$

从这种表示中可以发现，每年 1 月份的预测都是所有 1 月份观测值的某个指数加权滑动平均，每年其他月份类似。



廈門大學  
XIAMEN UNIVERSITY

Part 2

# 案例分析

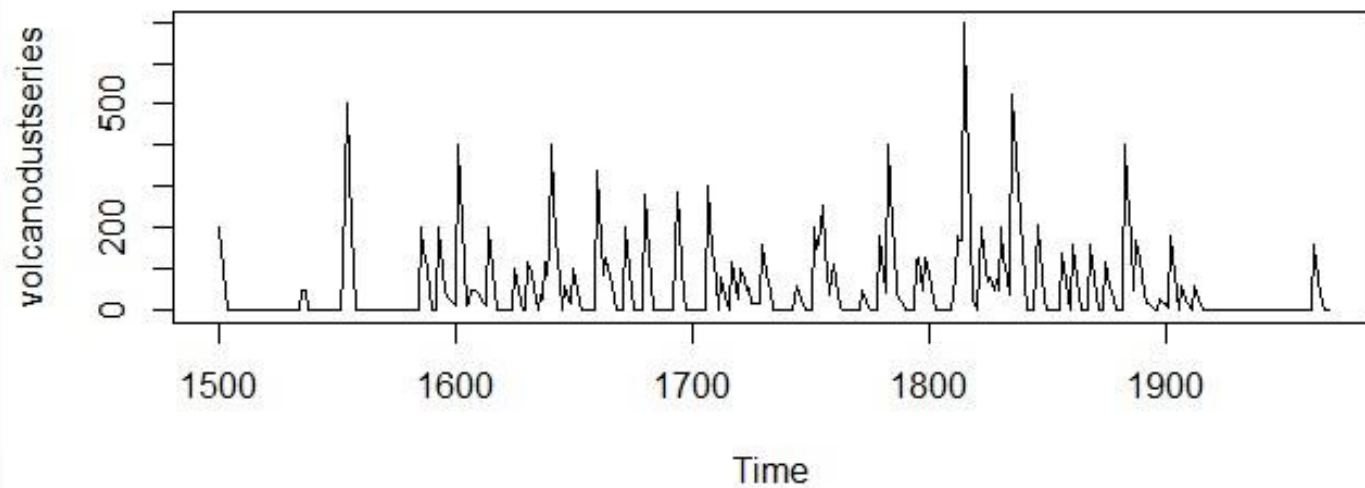


## 案例一

问题背景：现已知 1500 年到 1969 年北半球火山灰覆盖指数数据包含在数据文件 "dvi.dat" 中 (数据文件可在 [http : //robjhyndman.com/tsdldata/annual/dvi.dat](http://robjhyndman.com/tsdldata/annual/dvi.dat) 上进行下载).

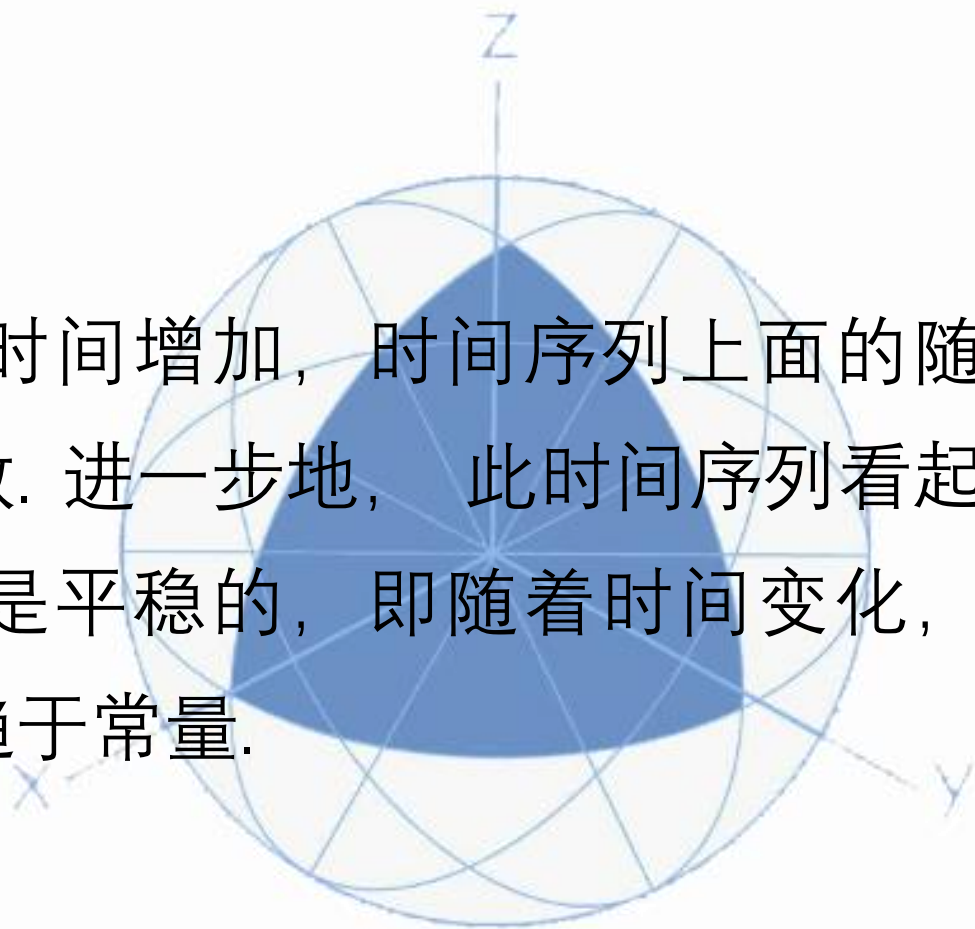


1、初步描述通过在软件中导入数据，我们可以绘出时序图如下所示：





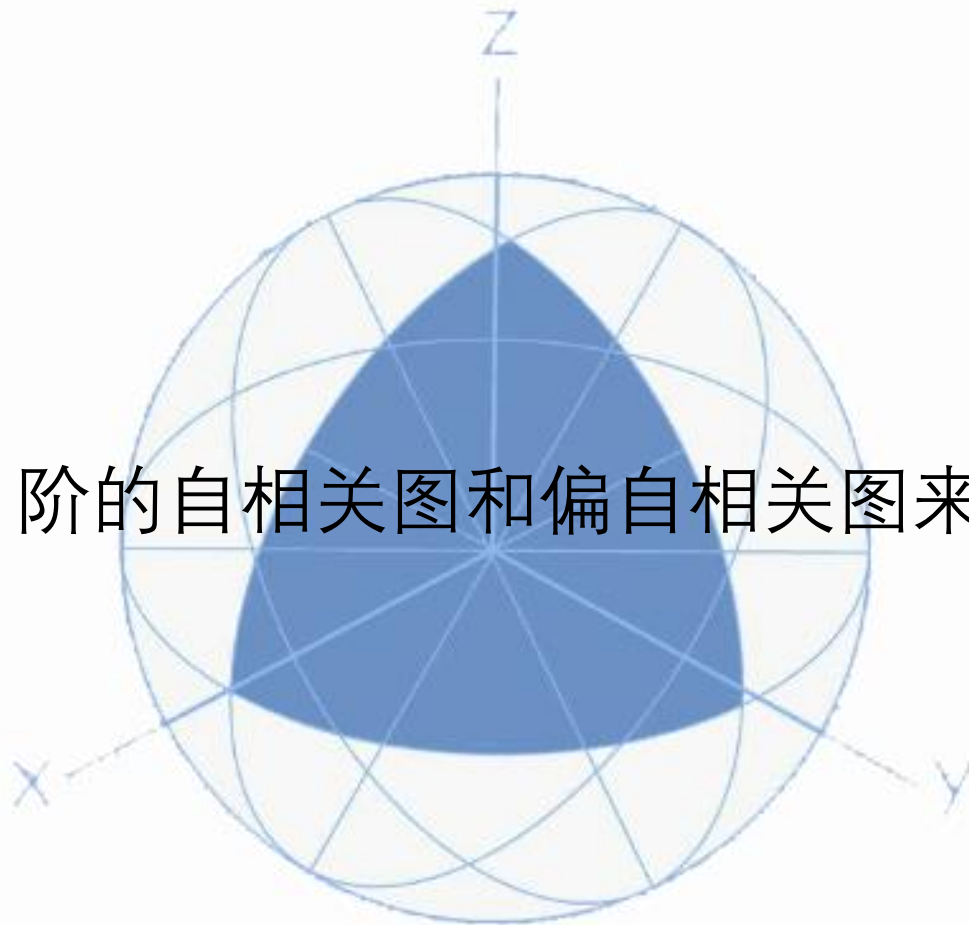
从图上可以看出，随着时间增加，时间序列上面的随机波动逐渐趋于一个常数. 进一步地，此时间序列看起来在平均值和方差上面是平稳的，即随着时间变化，他们的水平和方差大致趋于常量.





## 2、模型定阶

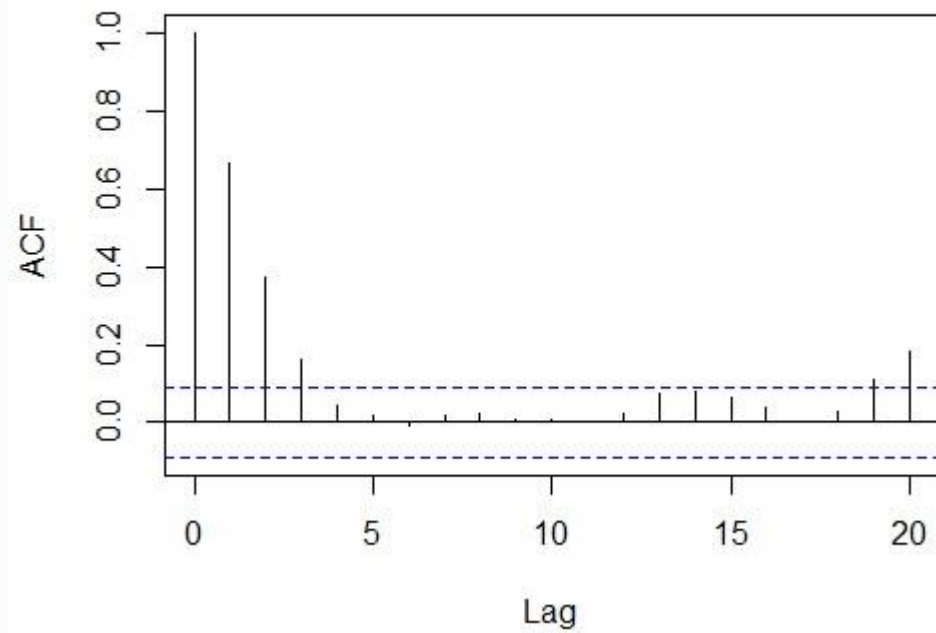
我们可以画出滞后 1-20 阶的自相关图和偏自相关图来初步进行模型识别：



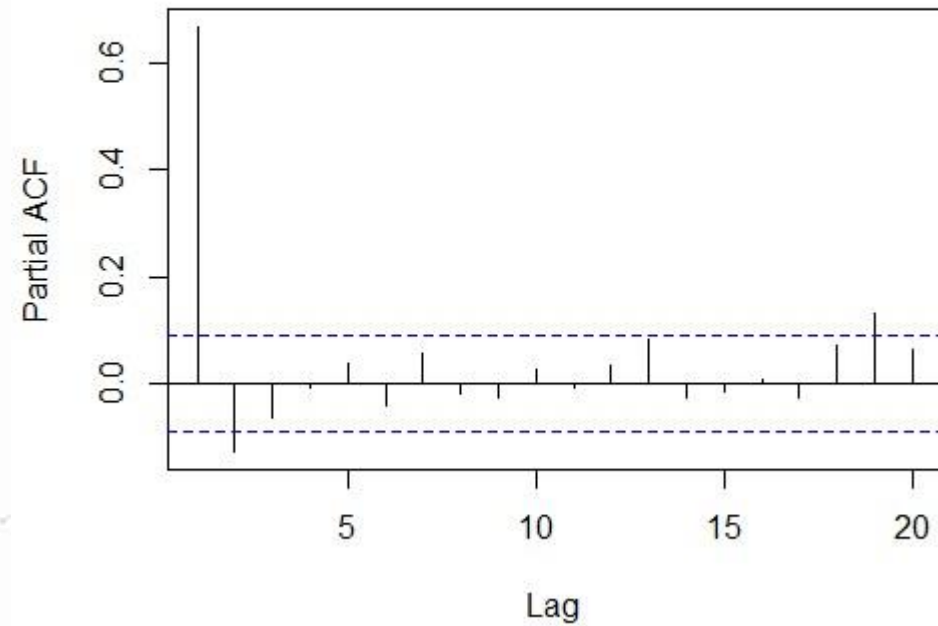


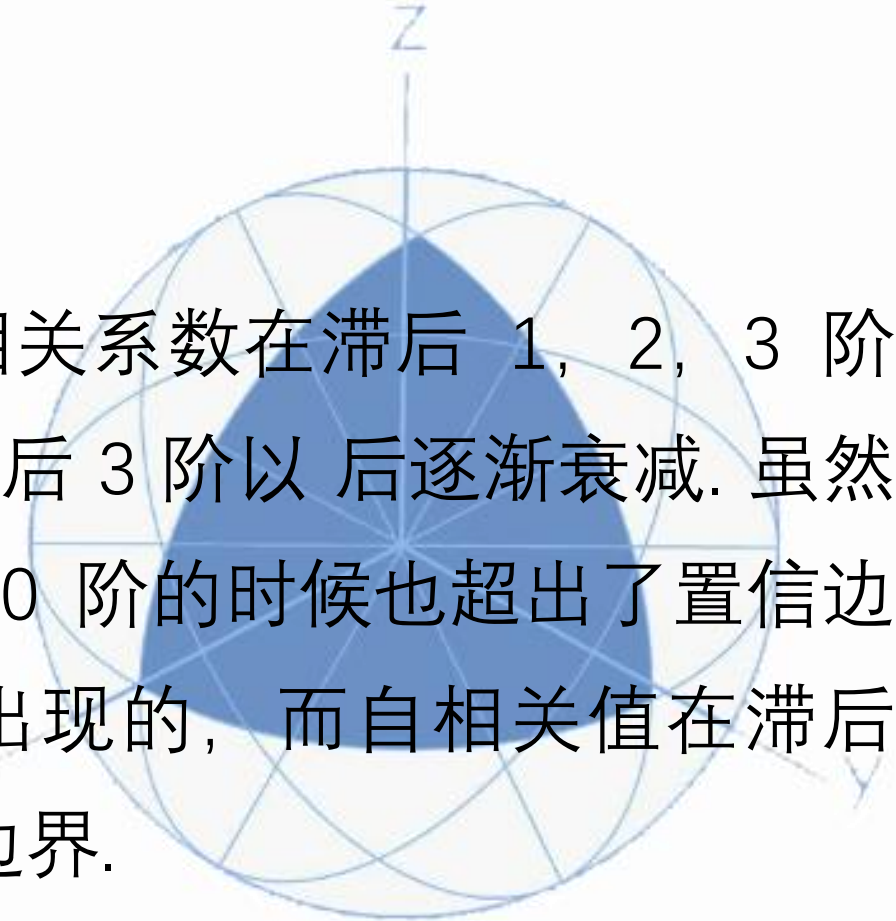
$Z$   
 $i$

Series volcanodustseries

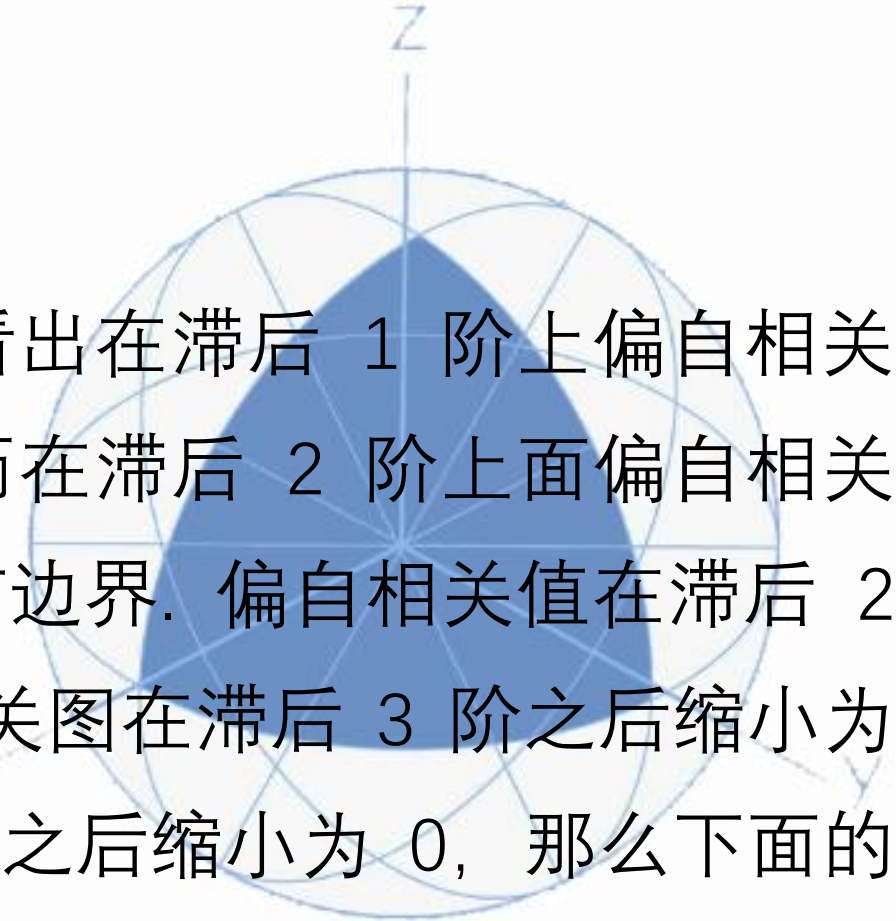


Series volcanodustseries





从自相关图可以看出，自相关系数在滞后 1, 2, 3 阶时超出了置信边界，且在滞后 3 阶以后逐渐衰减. 虽然自相关值在滞后 19 阶和 20 阶的时候也超出了置信边界，那么很可能属于偶然出现的，而自相关值在滞后 4-18 阶上都没有超出显著边界.



从偏自相关图中我们可以看出在滞后 1 阶上偏自相关为正且超出了显著边界，而在滞后 2 阶上面偏自相关是负的且也同样超出了置信边界。偏自相关值在滞后 2 阶之后缩小至 0。既然自相关图在滞后 3 阶之后缩小为 0，且偏相关图在滞后 2 阶之后缩小为 0，那么下面的模型可能适合此时间序列：

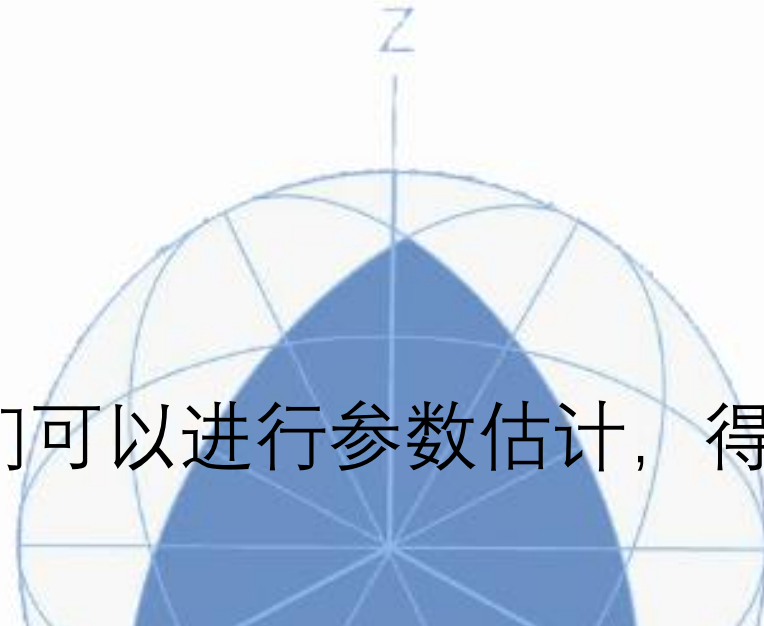




- ARMA(2,0) 模型，既然偏自相关图在滞后 2 阶之后缩小至 0，且自相关图在滞后 3 阶之后缩小至 0，且偏相关图在滞后 2 阶之后为 0.
- ARMA(0,3) 模型，既然自相关图在滞后 3 阶之后为 0，且偏相关缩小至 0(尽管这点对于此模型不太合适).
- ARMA(2,3) 混合模型，既然自相关图和偏相关图都缩小至 0(尽管自相关图缩小太突然对这个模型不太合适).

### 3、参数估计

通过建立的不同模型，我们可以进行参数估计，得到相应的系数，如下图所示：



```
Call:
arima(x = volcanodustseries, order = c(2, 0, 0), method = "ML")

Coefficients:
          ar1          ar2    intercept
      0.7533   -0.1268      57.3370
s.e.  0.0457    0.0458      8.5955

sigma^2 estimated as 4870:  log likelihood = -2662.54,  aic = 5333.09
```



Z  
|

Call:

```
arima(x = volcanodustseries, order = c(0, 0, 3), method = "ML")
```

Coefficients:

	ma1	ma2	ma3	intercept
	0.7439	0.4514	0.1917	57.4409
s.e.	0.0455	0.0502	0.0442	7.6544

sigma^2 estimated as 4852: log likelihood = -2661.69, aic = 5333.39

Call:

```
arima(x = volcanodustseries, order = c(2, 0, 3), method = "ML")
```

Coefficients:

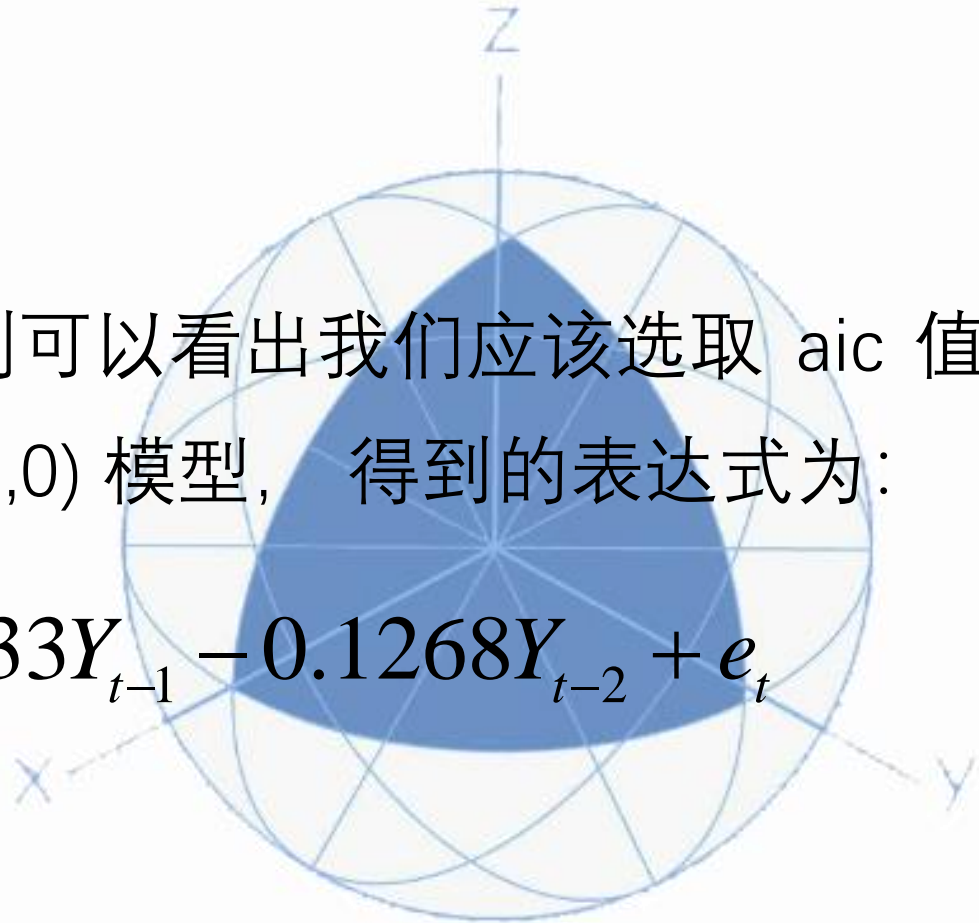
	ar1	ar2	ma1	ma2	ma3	intercept
	-0.2425	0.2471	0.9917	0.4117	0.1669	57.5018
s.e.	0.2547	0.1604	0.2538	0.1579	0.0726	8.2586

sigma^2 estimated as 4832: log likelihood = -2660.72, aic = 5335.43



通过相应的 aic 信息准则可以看出我们应该选取 aic 值最小的，也就是 ARMA(2,0) 模型，得到的表达式为：

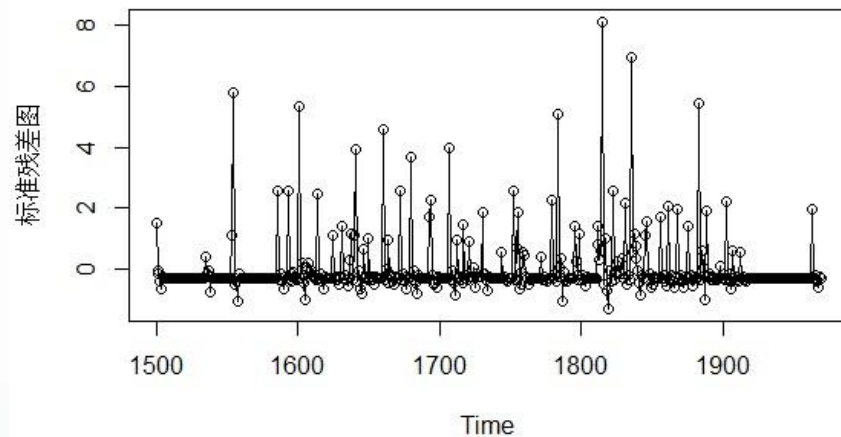
$$Y_t = 57.337 + 0.7533Y_{t-1} - 0.1268Y_{t-2} + e_t$$





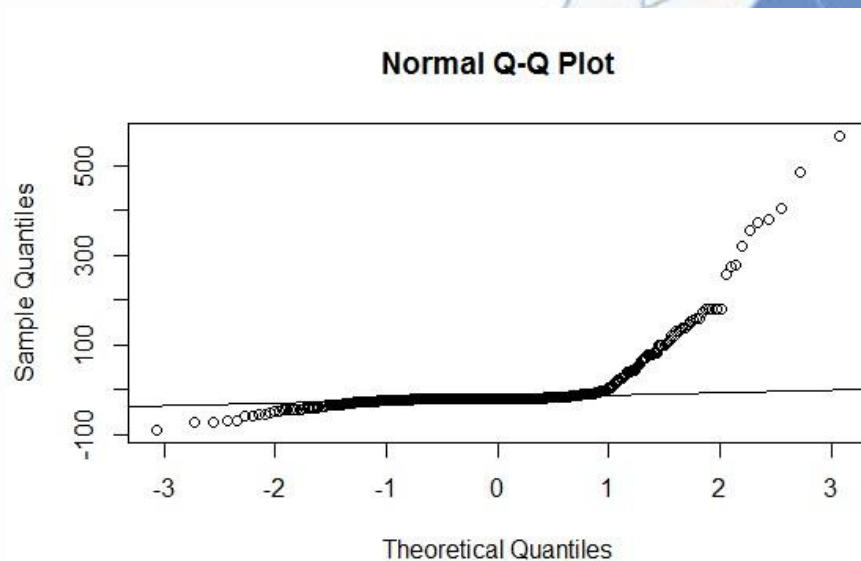
## 4、模型诊断

模型诊断我们主要需要检查残差的自相关性与正态性，首先我们可以绘出标准残差图如下所示：



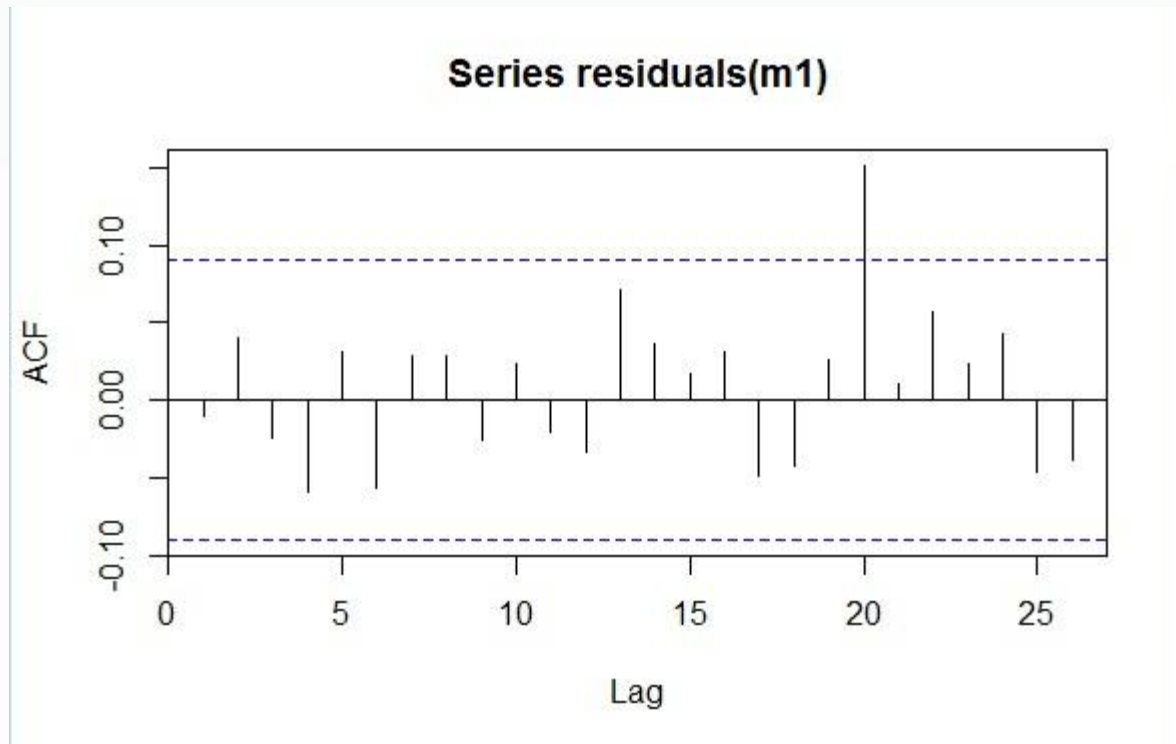


残差的正态性，一方面我们可以通过直观的 QQ 图来观察，另一方面我们也可以用 shapiro-walk 检验



另外正态性检验结果得  $p$  值远小于 0.05, 所以可以认为残差是满足正态性假设的. 残差的自相关性, 一般可以通过残差的自相关图来观察.

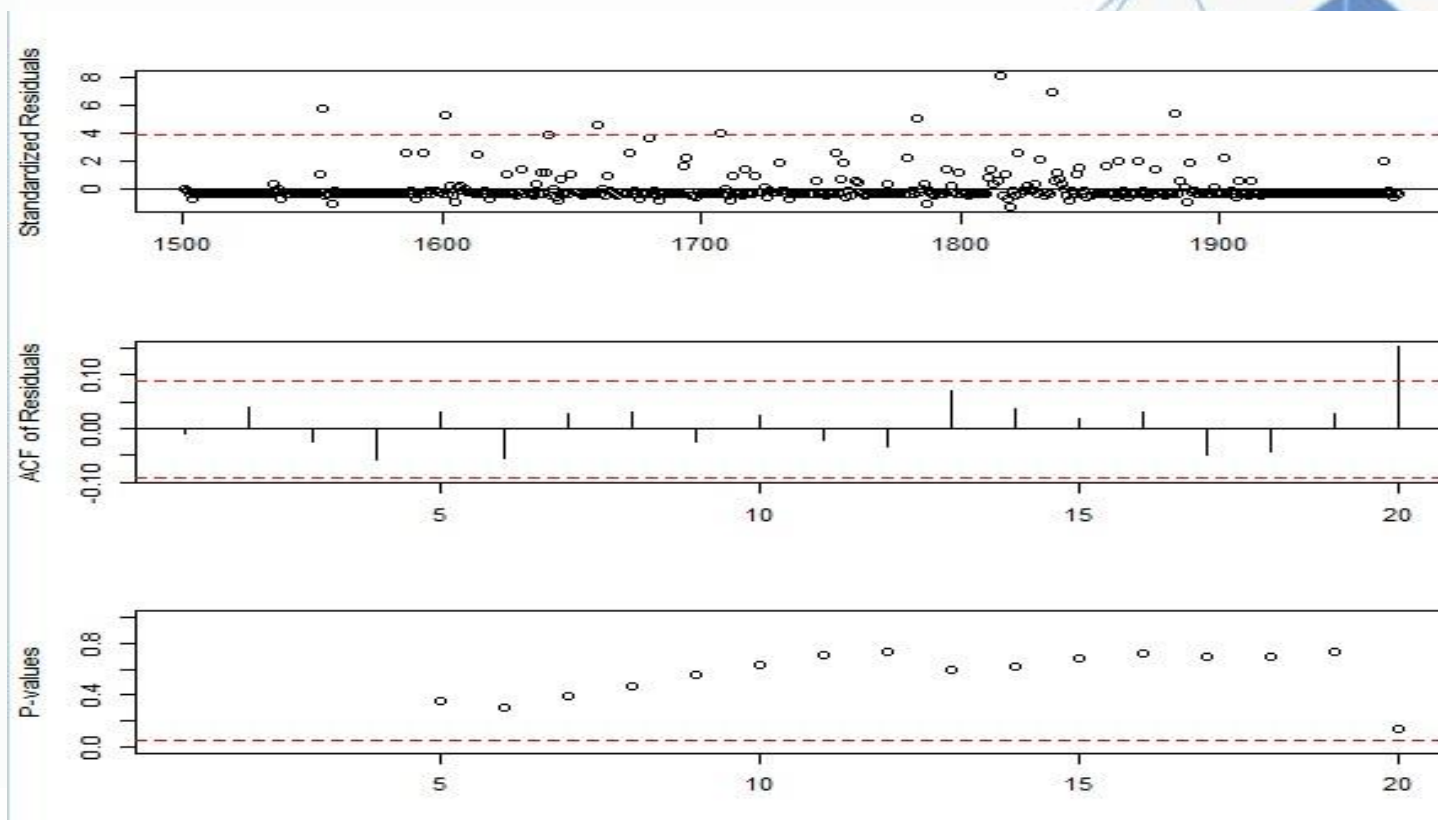




从自相关图我们可以看出残差的自相关系数在滞后 1-19 阶时都在置信边界范围以内，只有在滞后 20 阶远远的超出了，说明很有可能只是偶然因素导致的异常值的影响.

## 5、模型预测

通过模型诊断以后我们可以得出相应的预测图如下所示：



## 案例二

问题背景：图 1 中显示的是自 1994 年 1 月至 2004 年 12 月间加拿大西北部边境的阿勒 特地区的月度  $\text{CO}_2$  水平. 从图中我们可以看出， $\text{CO}_2$  含量有较强的上升趋势，同时也具有季节性特征.





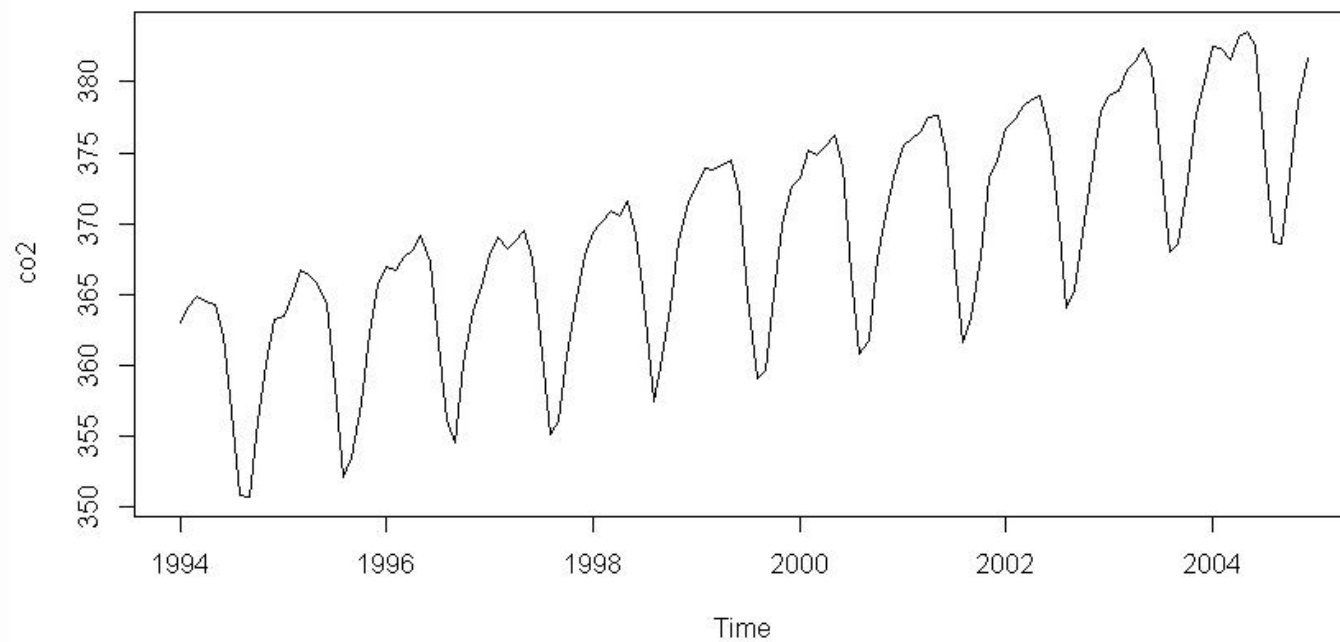


图 1 月 $CO_2$ 水平

图 2 中，用月度符号绘出了最后几年 $CO_2$ 水平变化的图象.

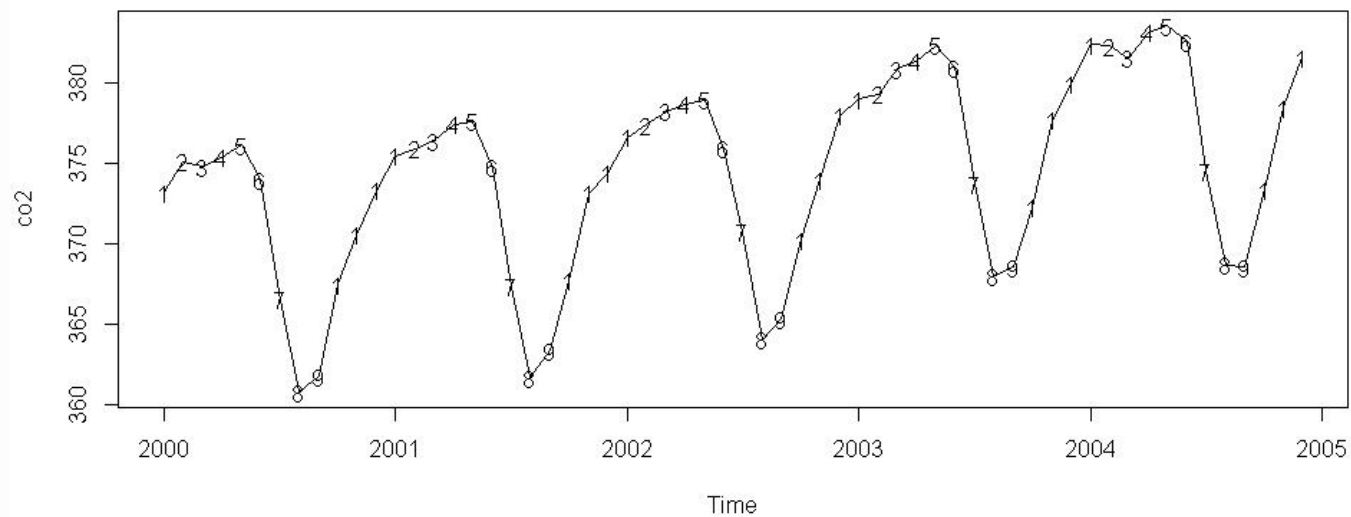
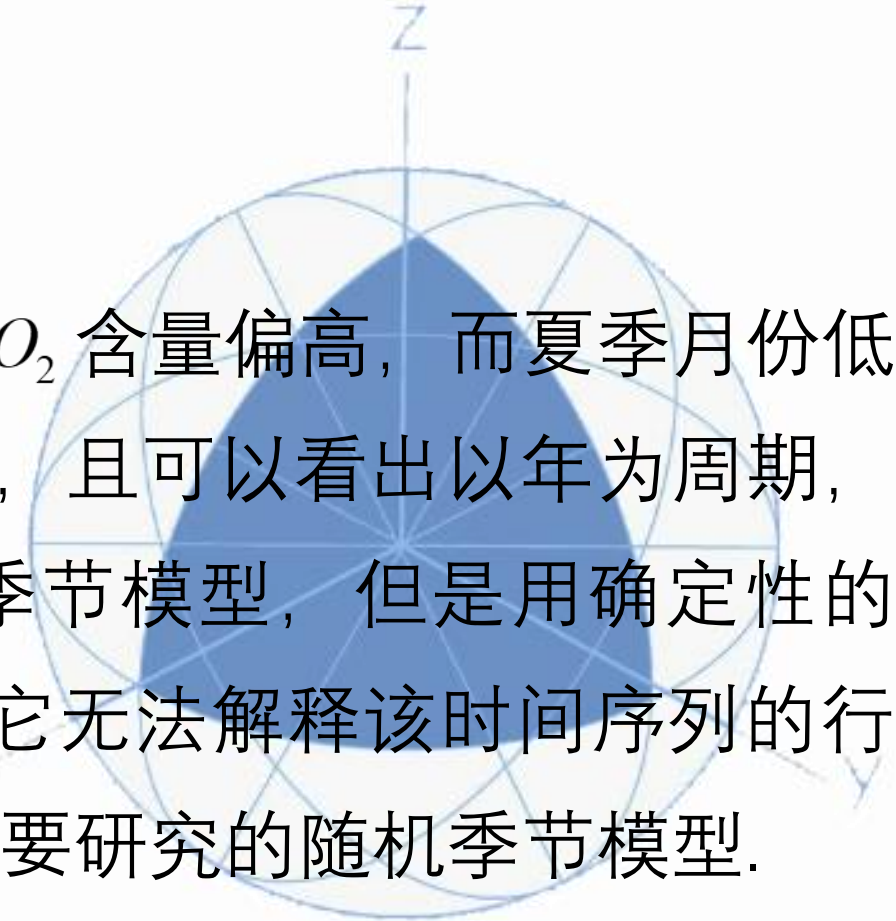


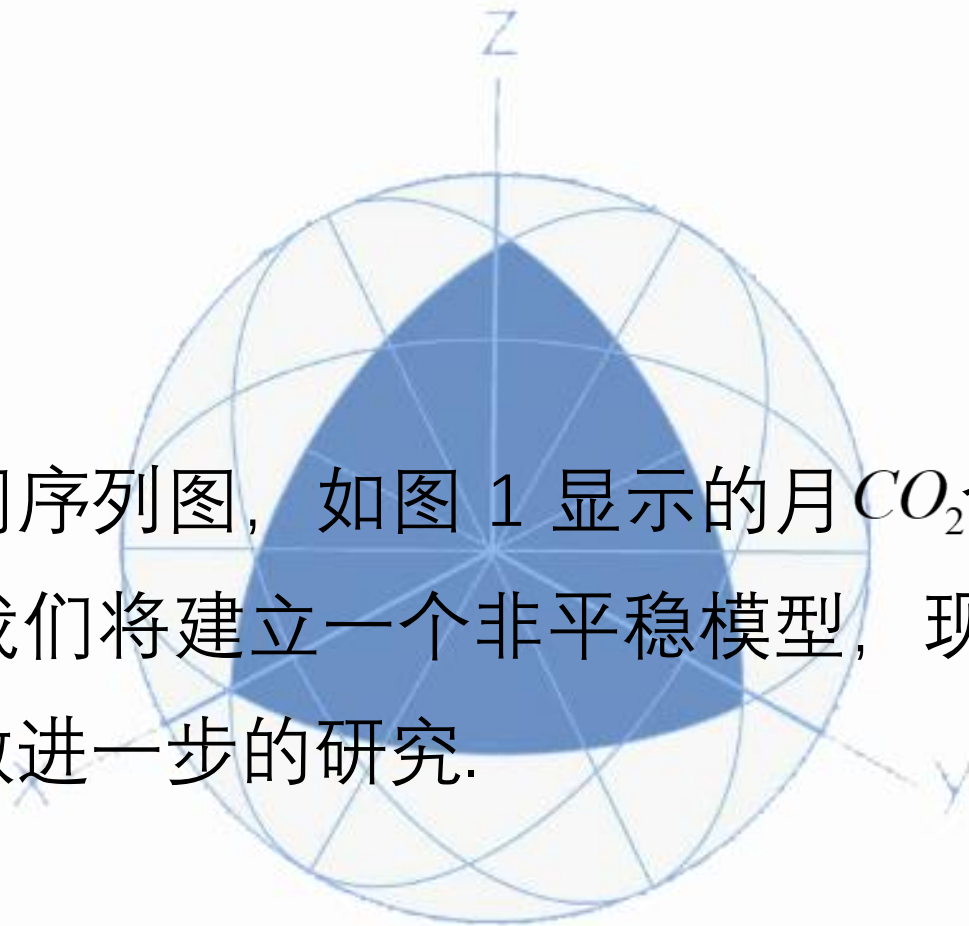
图 2 用月度符号表示的  $CO_2$  含量水平



如上两图所示，冬季月份  $CO_2$  含量偏高，而夏季月份低得多。图形呈现一种周期性，且可以看出以年为周期，这里我们就可以考虑应用季节模型，但是用确定性的季节模型，我们可以发现它无法解释该时间序列的行为，所以这就引出了我们将要研究的随机季节模型。

## 1、模型识别

第一步就是仔细观察时间序列图，如图 1 显示的月  $CO_2$  含量. 从图中的上升趋势将引导我们将建立一个非平稳模型，现在从序列的样本自相关函数来做进一步的研究.



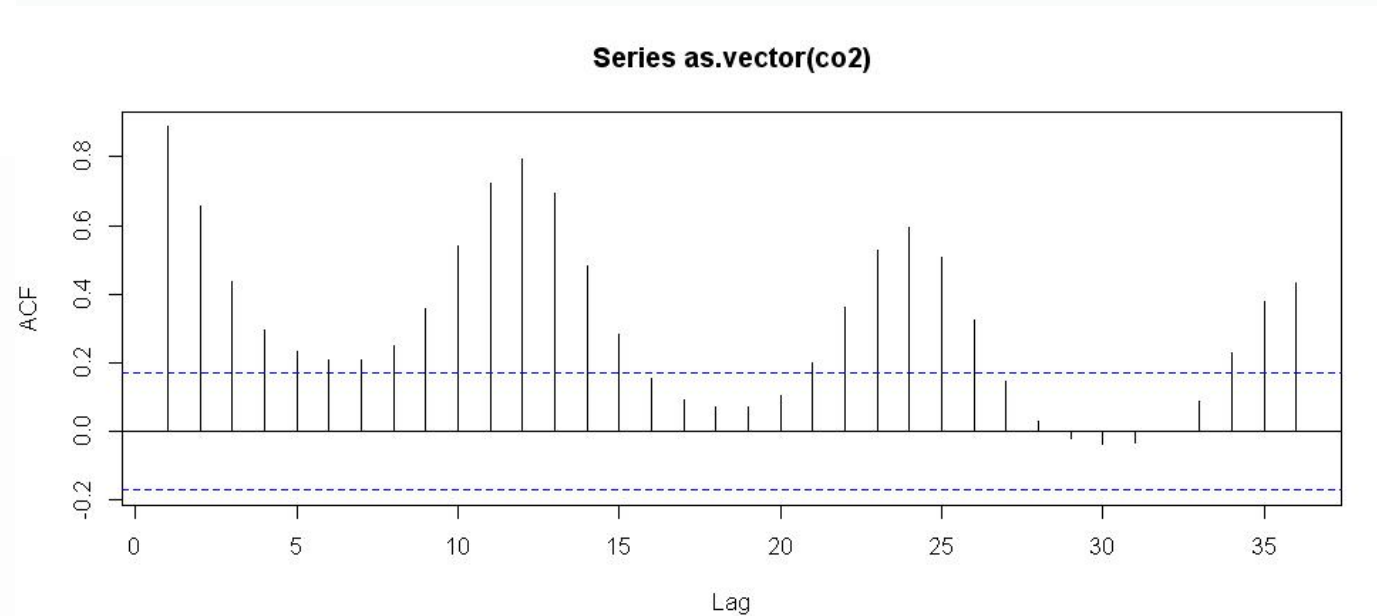


图 3  $CO_2$  含量样本的自相关函数

从图 3，我们可以看出季节自相关关系十分显著. 注意在滞后 12,24,36,... 上的强相关性. 至此我们对该序列进行一阶差分，观察进一步的情况.



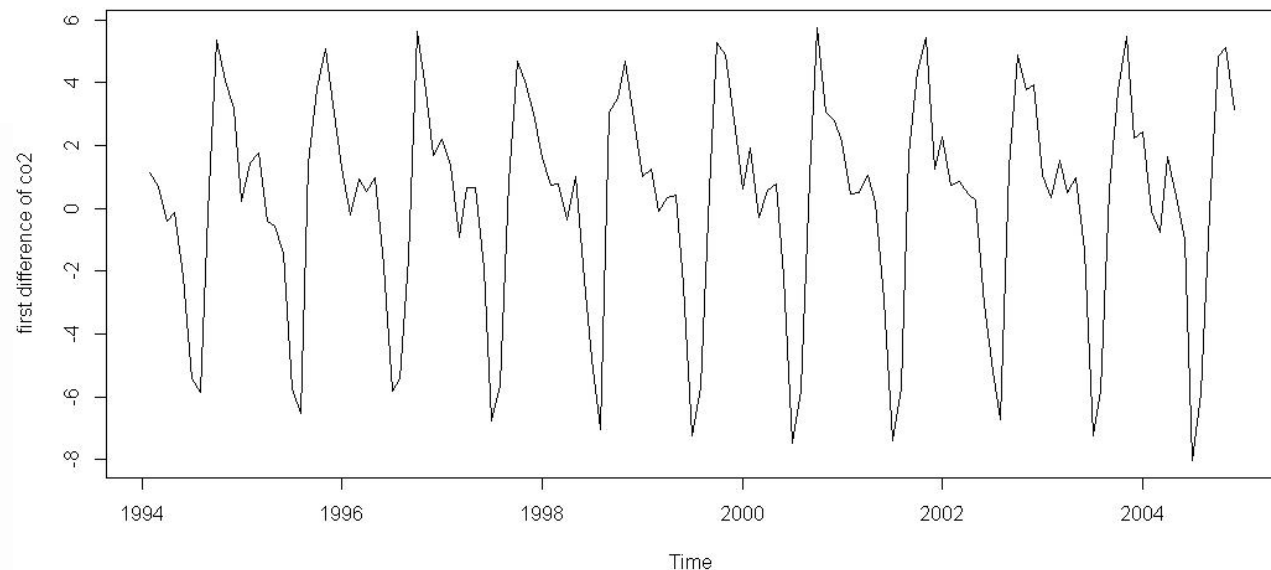


图 4  $CO_2$ 含量一次差分序列图

从图 4 我们可以看出，序列的上升趋势已然消失，但仍能看出带有周期性，我们继续进一步差分后序列的计算自相关函数.

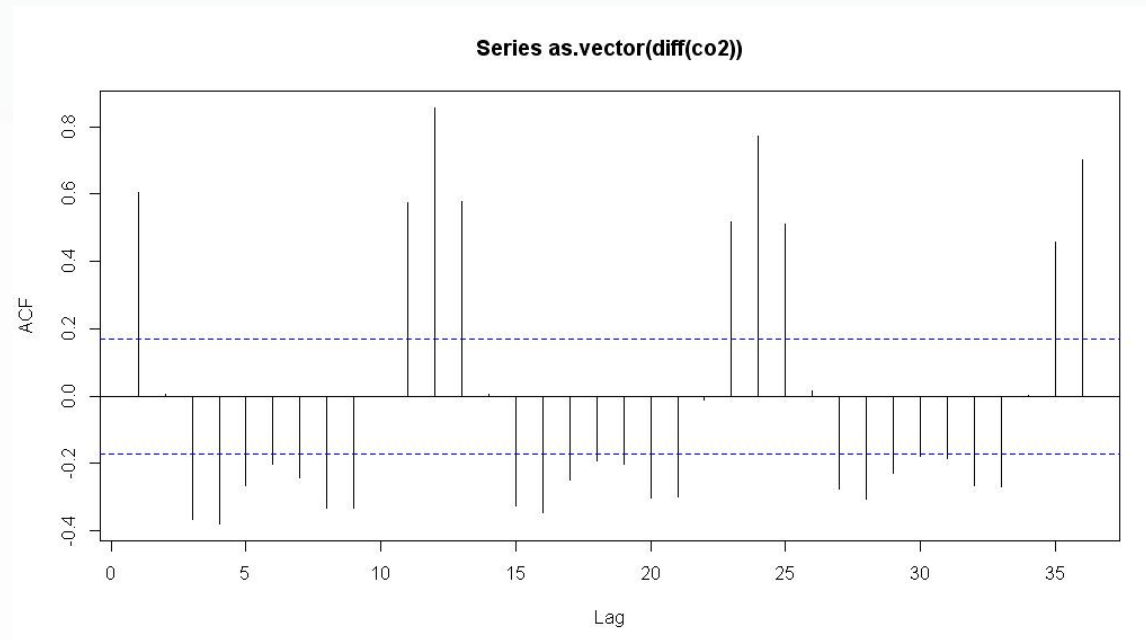


图 5  $CO_2$  含量一次差分序列的样本自相关函数

图 5 中我们能明显看出强烈的季节性，在滞后 12,24,36,... 上仍然具有强相关性。为此我们进一步做季节差分。

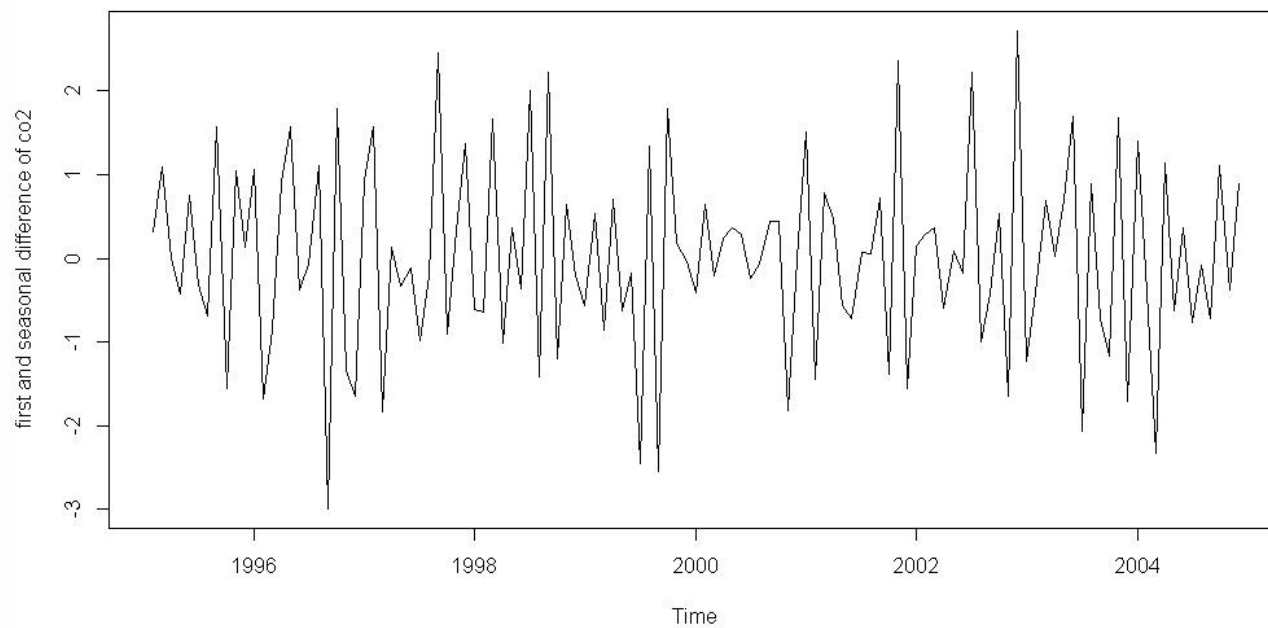


图 6  $CO_2$  含量一次差分 and 季节差分后的时间序列图

图 6 中，我们可得经过一次差分和季节差分后的时间序列的季节性基本消失了，紧接着我们继续做出此时的自相关函数图。



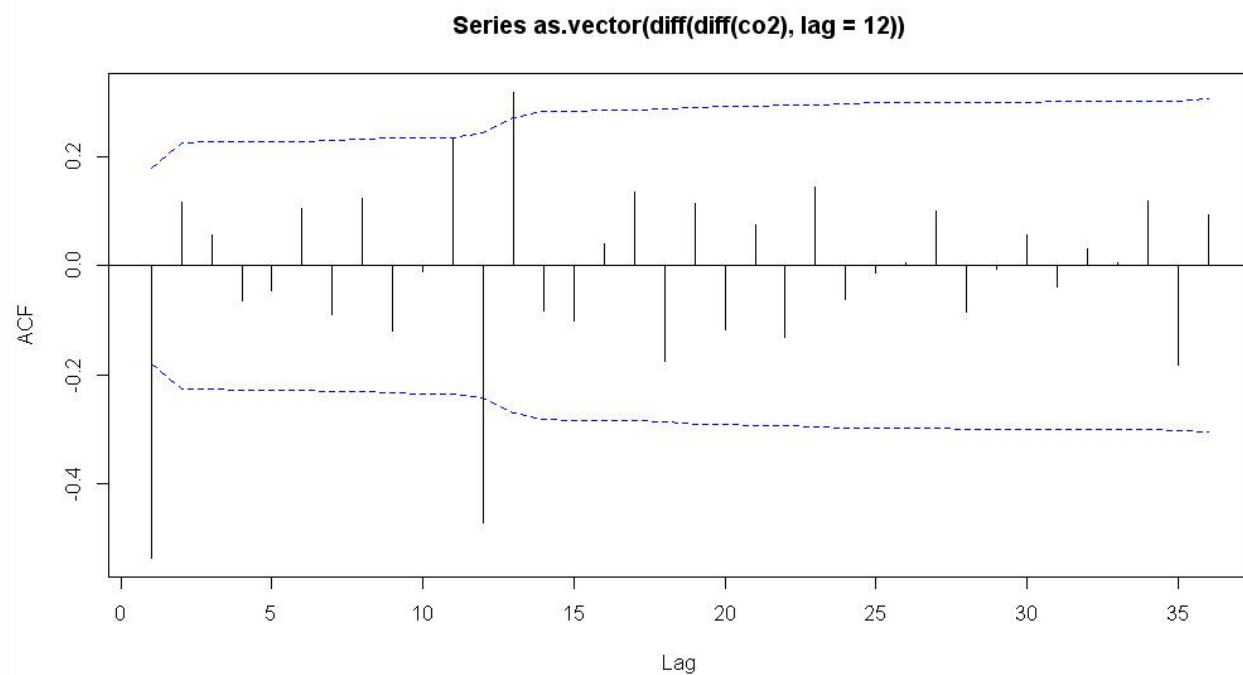


图 7  $CO_2$  含量一次差分 and 季节差分后的样本自相关函数



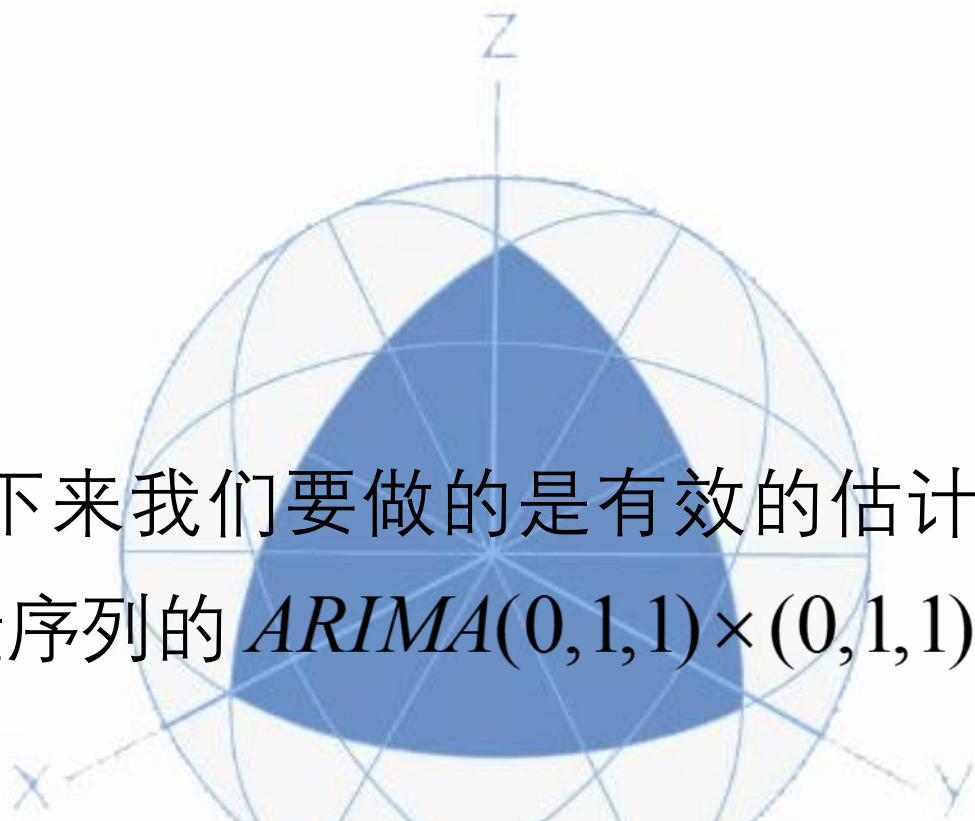
从图 7 中, 可以看出, 建立一个滞后 1 和 12 上具有自相关性的简单模型就够了. 考虑识别乘法季节  $ARIMA(0,1,1) \times (0,1,1)_{12}$  模型

$$\nabla_{12} \nabla Y_t = e_t - \theta e_{t-1} - \Theta e_{t-12} + \theta \Theta e_{t-13}$$

## 2、模型拟合

建立了季节模型后，接下来我们要做的是有效的估计出模型中的参数. 对 $CO_2$ 含量序列的  $ARIMA(0,1,1) \times (0,1,1)_{12}$  模型. 建立的模型如下

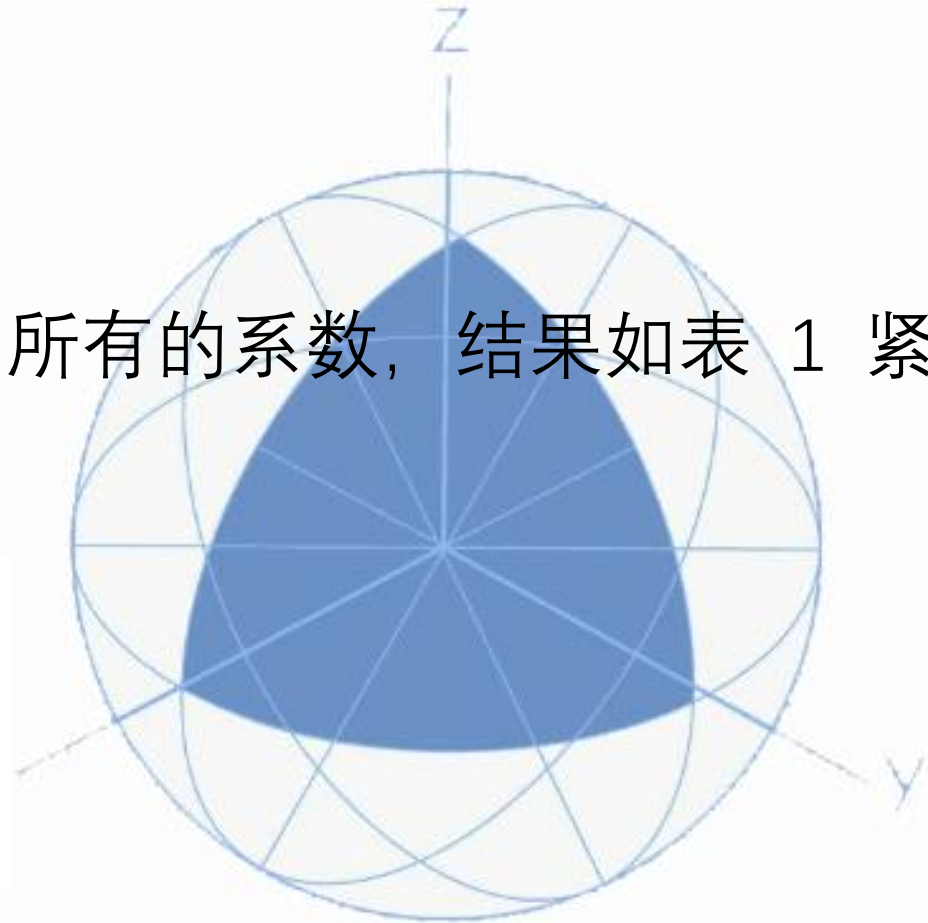
$$Y_t = Y_{t-1} + Y_{t-12} - Y_{t-13} + e_t - \theta e_{t-1} - \Theta e_{t-12} + \theta\Theta e_{t-13}$$



通过 R 软件我们可以算出所有的系数，结果如表 1 紧接着我们对模型加以检验.

表 10.1  $CO_2$  模型的参数估计

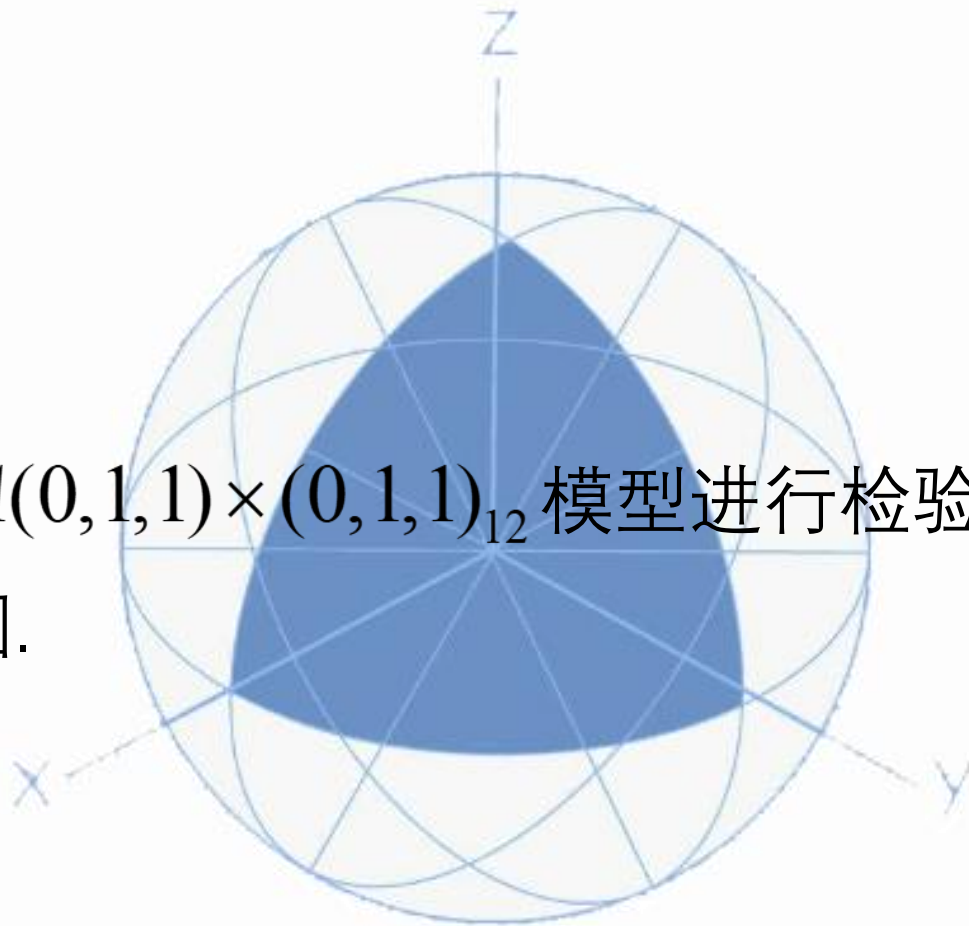
系数	$\theta$	$\Theta$
估计值	-0.5792	-0.8206
标准误差	0.0791	0.1137





### 3、模型检验—残差检验

为了对估计后的  $ARIMA(0,1,1) \times (0,1,1)_{12}$  模型进行检验，  
第一步，观察残差序列图.



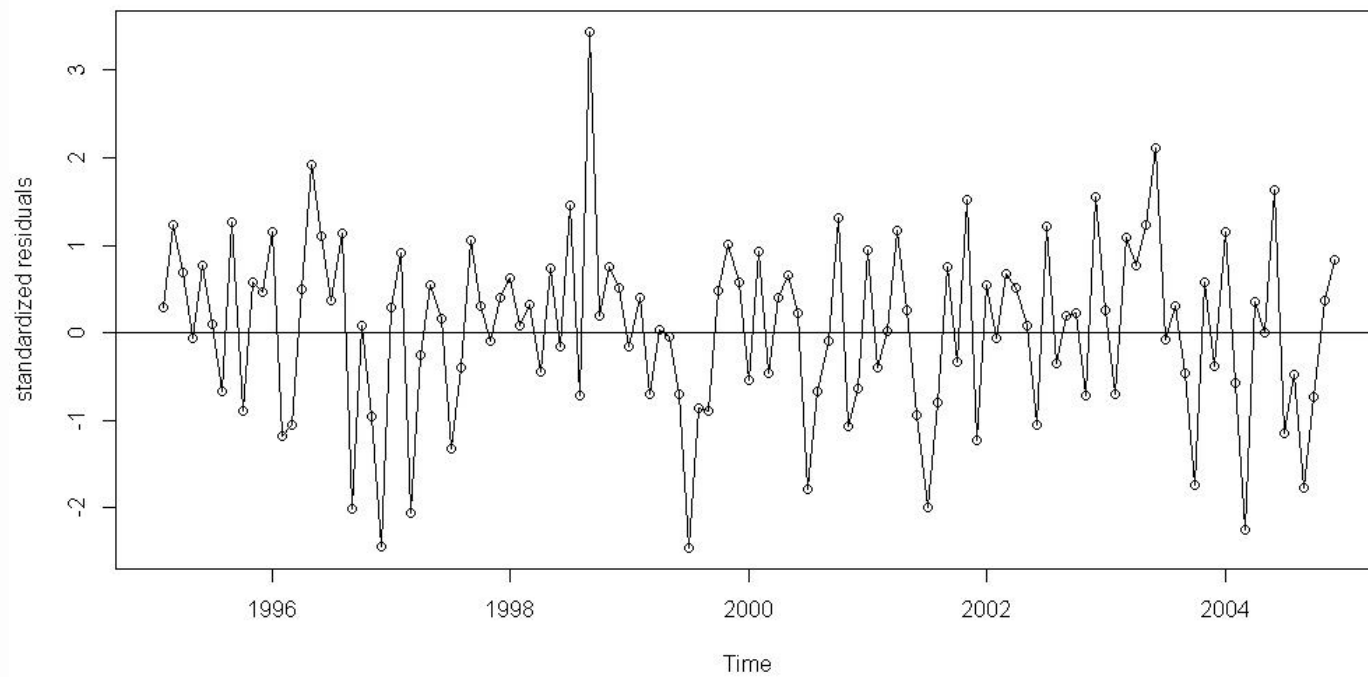


图 8  $ARIMA(0,1,1) \times (0,1,1)_{12}$  模型的残差



图 8 给出了标准残差图，除了序列中间某些异常行为外，此残差图并没有标明模型有任何主要的不规则性。我们进一步做残差的样本 ACF 以便进一步观察。



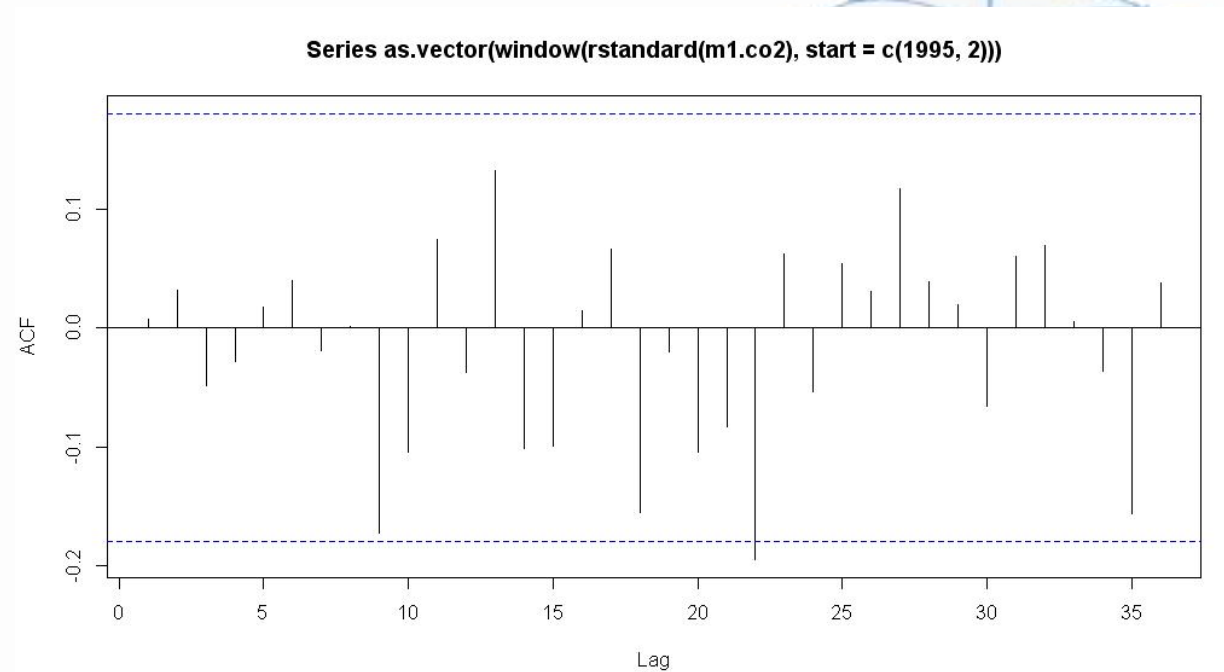


图 9  $ARIMA(0,1,1) \times (0,1,1)_{12}$  模型残差的 ACF

图 9 看出相关系数位于滞后 22, 其值仅为  $-0.17$ , 相关性非常小. 下面借助于残差来研究误差项的正态性问题.



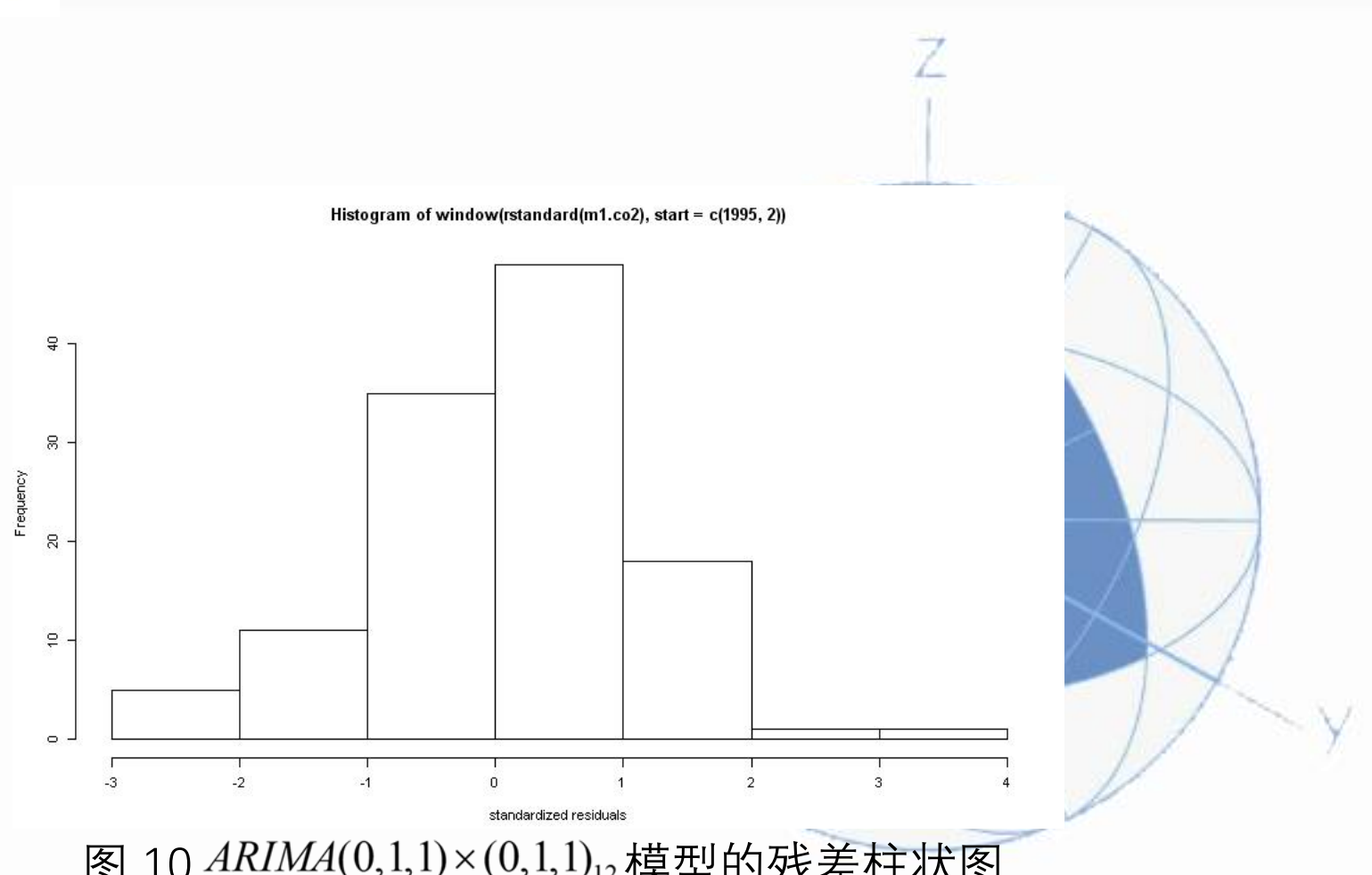


图 10  $ARIMA(0,1,1) \times (0,1,1)_{12}$  模型的残差柱状图

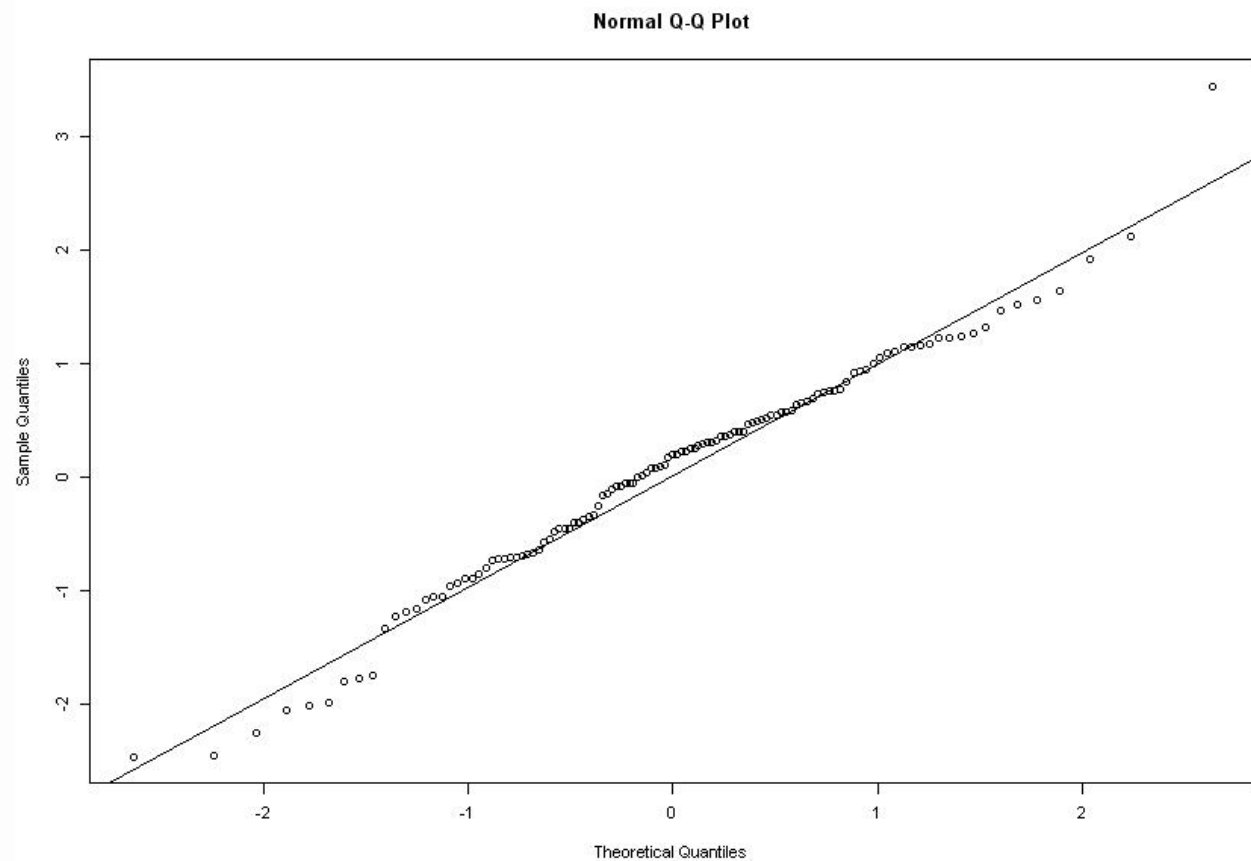


图 11  $ARIMA(0,1,1) \times (0,1,1)_{12}$  模型残差的 QQ 正态图



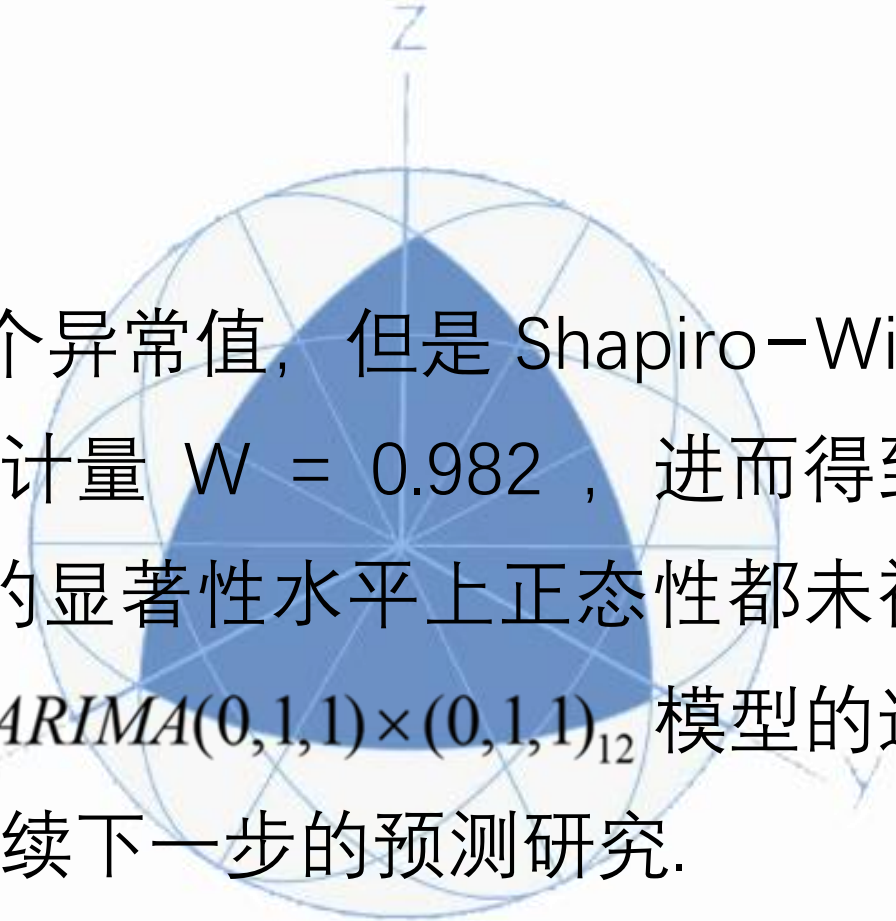


图 11 中的尾部，看到了一个异常值，但是 Shapiro-Wilk 正态性检验法给出的检验统计量  $W = 0.982$ ，进而得到  $p$  值为 0.11，且在任何通常的显著性水平上正态性都未被拒绝. 综上，我们可以判定  $ARIMA(0,1,1) \times (0,1,1)_{12}$  模型的选定是合理，从而我们可以继续下一步的预测研究.

## 4、模型预测

由方程 (4.14) 进行迭代做进一步的预测,  $\text{CO}_2$  模型的预测值与极限预测下图.

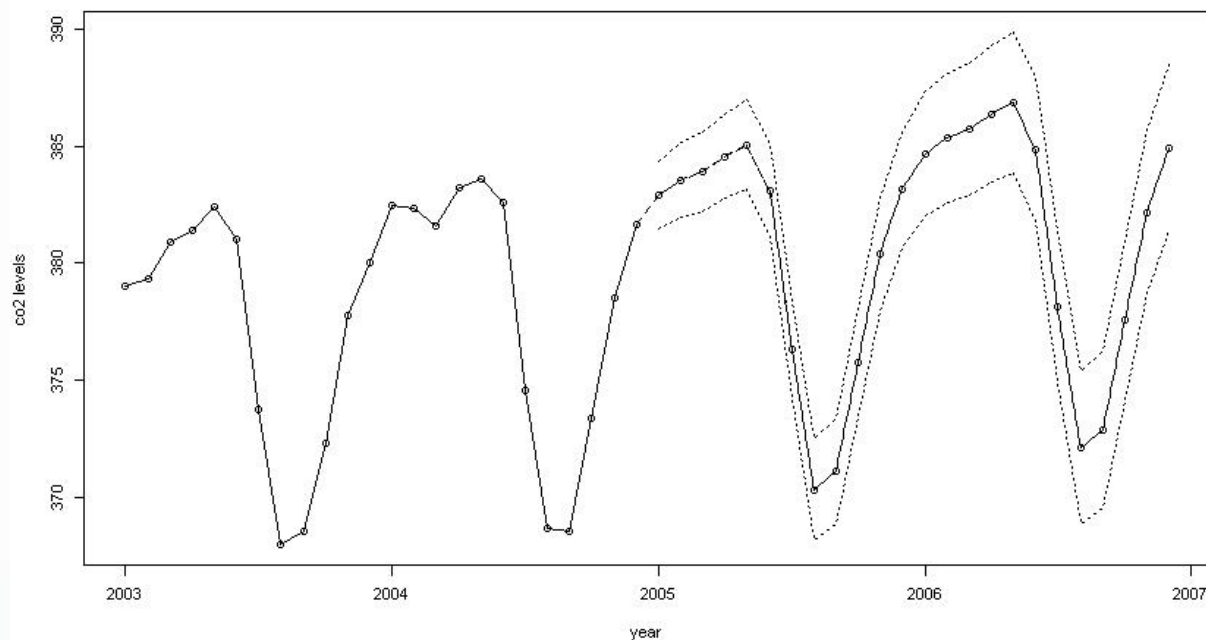
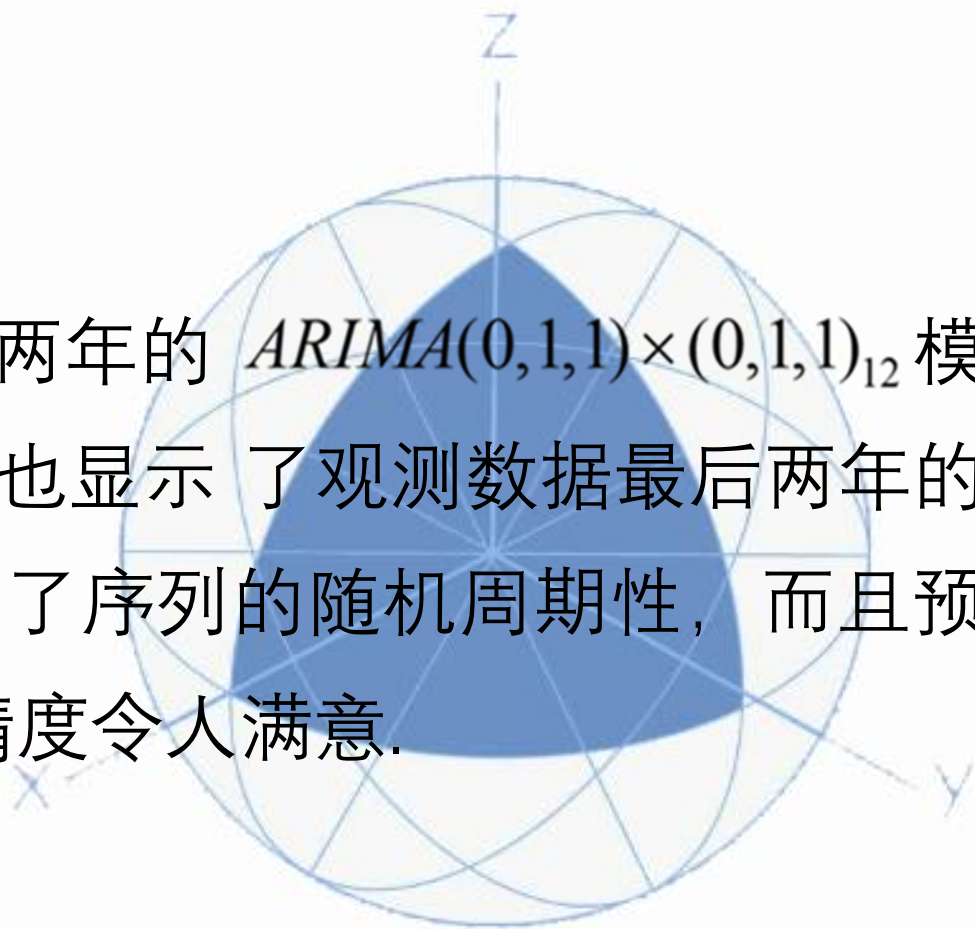


图 12 拟合的前置时间为两年的  $ARIMA(0,1,1) \times (0,1,1)_{12}$  模型的 95% 预测极限. 图中也显示了观测数据最后两年的值. 预测值很好地模仿出了序列的随机周期性, 而且预测极限也显示出预测之精度令人满意.



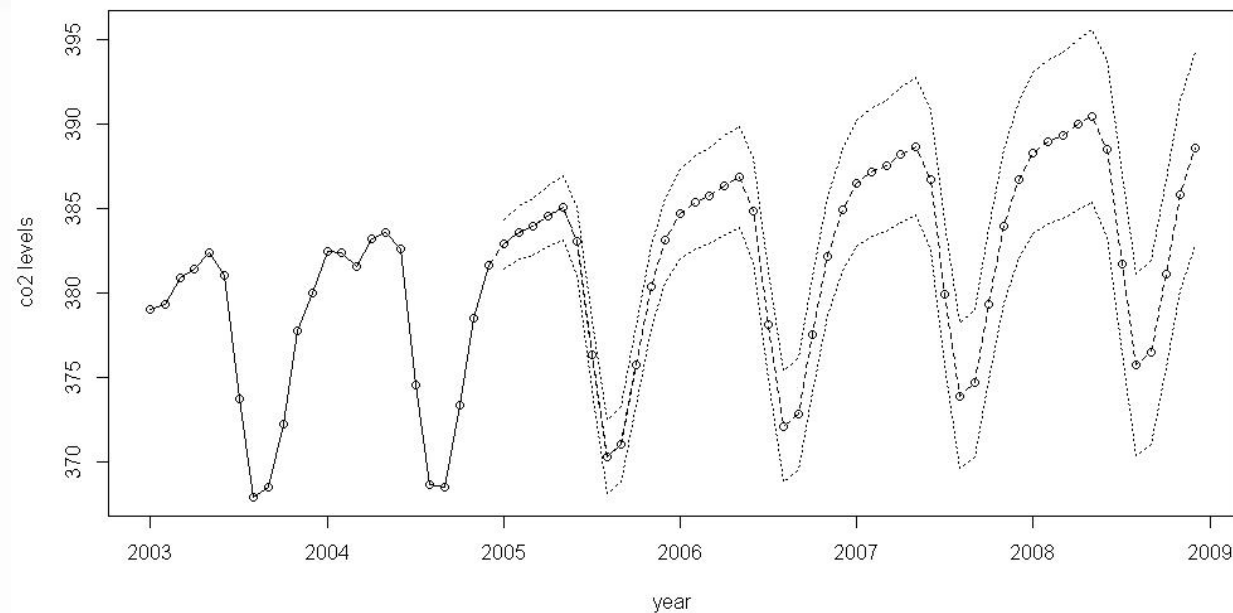


图 13 CO<sub>2</sub> 模型的长期预测

图 13 显示了最后一年的观测数据和后续 4 年的预测值，我们可以发现，预测极限越来越宽，这源于预测中较大的不确定性。即在迭代中的预测，随着预测的步长其精度也随之降低。



廈門大學  
XIAMEN UNIVERSITY

THANK YOU

