数学建模模型算法精讲课——

# 时间序列ARIMA模型

—— 江北老师

不为失败找借口，
只为成功找方法

# ARIMA模型

- 模型引出
- 模型求解
- 模型检验
- 具体代码

# ARIMA模型——模型介绍

## ➢ 时间序列

- 时间序列也称动态序列，是指将某种现象的指标数值按照时间顺序排列而成的数值序列。时间序列分析大致可分成三大部分，分别是描述过去、分析规律和预测未来，本讲将主要介绍时间序列分析中常用ARIMA模型。

## ➢ 时间序列数据

- 对同一对象在不同时间连续观察所取得的数据，它具备两个要素，第一个要素是时间要素，第二个要素是数值要素
  - ✓ 从出生到现在，你的体重的数据（每年生日称重一次）
  - ✓ 中国历年来GDP的数据
  - ✓ 在某地方每隔一小时测得的温度数据
- 时间序列根据时间和数值性质的不同，可以分为时期时间序列和时点时间序列
  - ✓ 时期序列中，数值要素反映现象在一定时期内发展的结果
  - ✓ 时点序列中，数值要素反映现象在一定时点上的瞬间水平

## ➢ 区分时期和时点时间序列

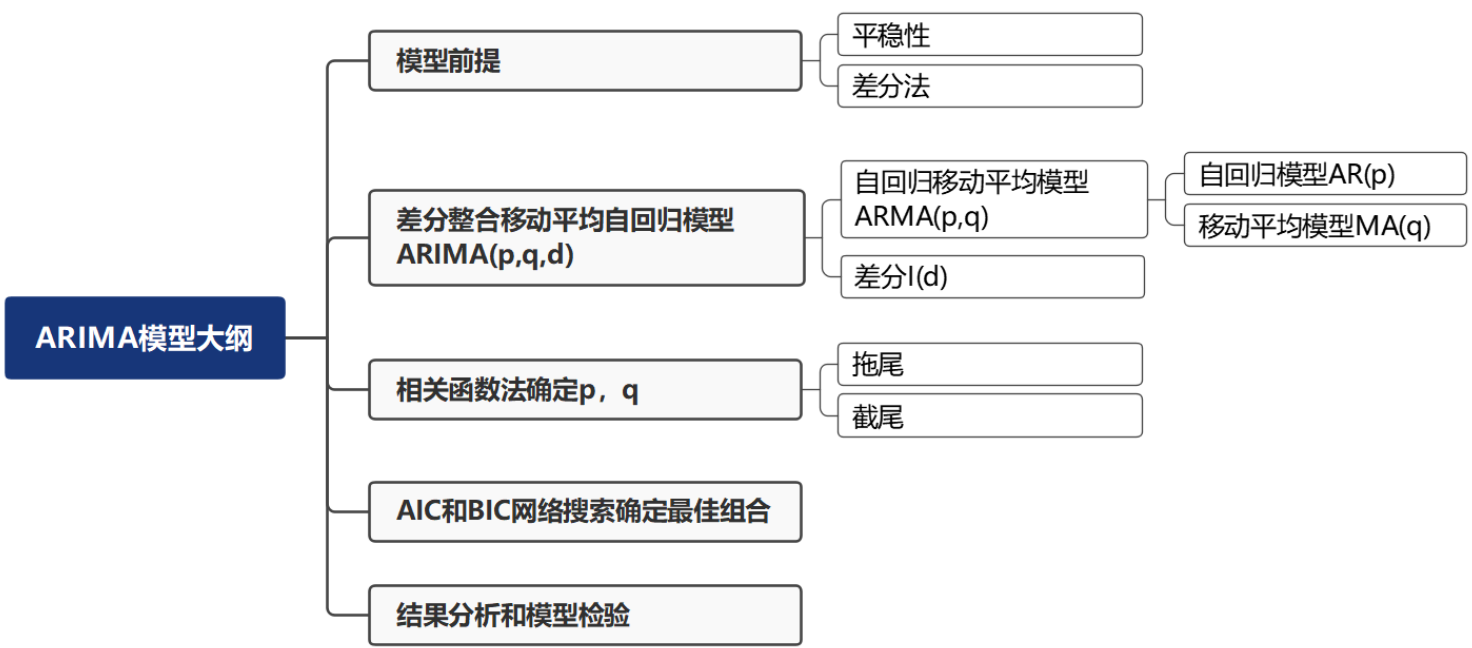
- ✓ 1）从出生到现在，你的体重的数据（每年生日称重一次）
- ✓ 2）中国历年来GDP的数据
- ✓ 3）在某地方每隔一小时测得的温度数据

- 1）和3）是时点时间序列；2）是时期时间序列
- 时期序列可加，时点序列不可加

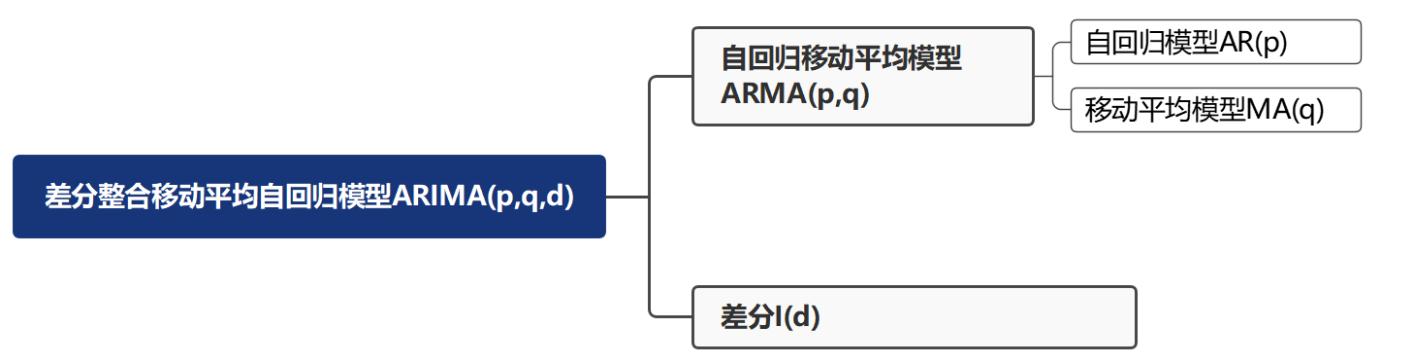
时期序列中的观测值反映现象在一段时期内发展过程的总量，不同时期的观测值可以相加，相加结果表明现象在更长一段时间内的活动总量；而时点序列中的观测值反映现象在某一瞬间上所达到的水平，不同时期的观测值不能相加，相加结果没有实际意义。

- 之前讲的灰色预测模型就有累加的过程。

# ARIMA模型——模型介绍



# ARIMA模型——模型介绍




- **自回归模型（$AR(p)$）**
  - 描述当前值和历史值之间的关系，用变量自身的历史数据对自身进行预测，其必须要满足平稳性要求，只适用于预测与自身前期相关的现象（时间序列的自相关性）
  - $p$阶自回归过程的公式定义：$y_t = \mu + \sum_{i=1}^{p} \gamma_i y_{t-i} + \epsilon_t$，$p$表示用几期的历史值来预测
  - $y_t$是当前值　$\mu$是常数项　$p$是阶数　$\gamma_i$是自相关系数

# ARIMA模型——模型介绍

- **移动平均模型（$MA(q)$）**
  - 移动平均模型关注的是自回归模型中误差项的累计
  - $q$阶自回归过程的公式定义：$y_t = \mu + \epsilon_t + \sum_{i=1}^{q} \theta_i \epsilon_{t-i}$
  - 即时间序列当前值与历史值没有关系，而只依赖于历史白噪声的线性组合
  - 移动平均法能有效地消除预测中的随机波动
- **自回归移动平均模型（$ARMA(p,q)$）**
  - 自回归与移动平均的结合
  - 公式定义：$y_t = \mu + \sum_{i=1}^{p} \gamma_i y_{t-i} + \epsilon_t + \sum_{i=1}^{q} \theta_i \epsilon_{t-i}$
  - 该式表明：
    - ✓ 一个随机时间序列可以通过一个自回归移动平均模型来表示，即该序列可以由其自身的过去或滞后值以及随机扰动项来解释。
    - ✓ 如果该序列是平稳的，即它的行为并不会随着时间的推移而变化，那么我们就可以通过该序列过去的行为来预测未来。

# ARIMA模型——模型介绍

## ➢ 差分自回归移动平均模型$ARIMA(p,d,q)$

- 将自回归模型$(AR)$、移动平均模型$(MA)$和差分法结合，我们就得到了差分自回归移动平均模型ARIMA($p$、$d$、$q$)
- $p$是自回归项，$q$为移动平均项数，$d$为时间序列成为平稳时所做的差分次数
- 原理：将非平稳时间序列转化为平稳时间序列然后将因变量仅对它的滞后值以及随机误差项的现值和滞后值进行回归所建立的模型

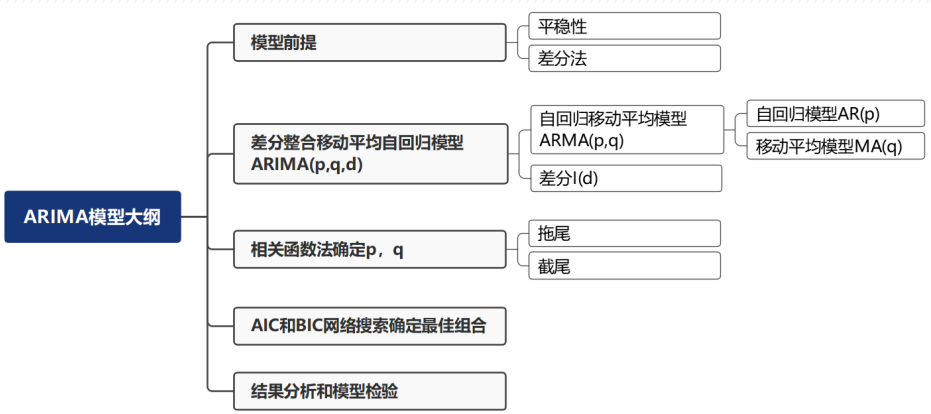
## ➢ $ARIMA$模型的建模步骤

- 对序列绘图，进行平稳性检验，观察序列是否平稳；对于非平稳时间序列要先进行 d 阶差分，转化为平稳时间序列；
- 经过第一步处理，已经得到平稳时间序列。要对平稳时间序列分别求得其自相关系数（ACF）和偏自相关系数（PACF），通过对自相关图和偏自相关图的分析，得到最佳的阶数p、q；
- 由以上得到的d、q、p ，得到 ARIMA 模型。然后开始对得到的模型进行模型检验。

# ARIMA模型——模型建立

## ➢ *ARIMA*模型的建模步骤

- 1)对序列绘图，进行平稳性检验，观察序列是否平稳；对于非平稳时间序列要先进行 d 阶差分，转化为平稳时间序列；
- 2)经过第一步处理，已经得到平稳时间序列。要对平稳时间序列分别求得其自相关系数（ACF）和偏自相关系数（PACF），通过对自相关图和偏自相关图的分析或通过AIC/BIC搜索，得到最佳的阶数p、q；
- 3)由以上得到的d、q、p ，得到 ARIMA 模型。然后开始对得到的模型进行模型检验。



# ARIMA模型——模型建立

## ➢ 股票价格预测

已知一个上市公司一段时期的开盘价，最高价，最低价，收盘价等信息，要求建立模型，预测股价

- 我们这里只需要股票的收盘价（close），我们可以把数据提取出来，并划分为训练集和测试集
- 本题我们把1-3月份的数据作为训练集，4-6月份的数据作为测试集

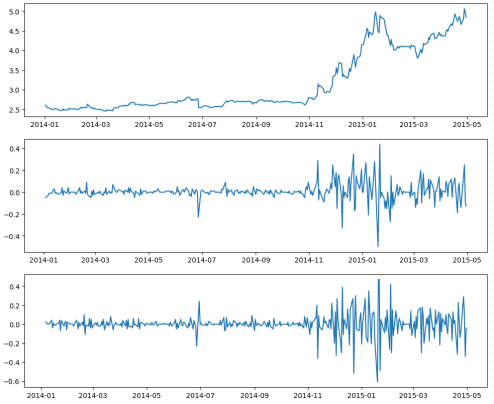
| | Date | Open | High | Low | Close | Volume |
|---|---|---|---|---|---|---|
| 1 | 2014-1-2 | 2.62 | 2.62 | 2.59 | 2.61 | 41632500 |
| 2 | 2014-1-3 | 2.6 | 2.61 | 2.56 | 2.56 | 45517700 |
| 3 | 2014-1-6 | 2.57 | 2.57 | 2.5 | 2.53 | 68674700 |
| 4 | 2014-1-7 | 2.51 | 2.52 | 2.49 | 2.52 | 53293800 |
| 5 | 2014-1-8 | 2.51 | 2.54 | 2.49 | 2.51 | 69087900 |
| 6 | 2014-1-9 | 2.51 | 2.53 | 2.49 | 2.5 | 45339800 |
| 7 | 2014-1-10 | 2.5 | 2.51 | 2.49 | 2.49 | 41009000 |
| 8 | 2014-1-13 | 2.5 | 2.52 | 2.5 | 2.52 | 29469300 |
| 9 | 2014-1-14 | 2.51 | 2.52 | 2.5 | 2.51 | 30626300 |
| 10 | 2014-1-15 | 2.51 | 2.52 | 2.5 | 2.51 | 50614100 |
| 11 | 2014-1-16 | 2.51 | 2.52 | 2.49 | 2.5 | 31381500 |
| 12 | 2014-1-17 | 2.5 | 2.5 | 2.48 | 2.48 | 41326400 |
| 13 | 2014-1-20 | 2.49 | 2.49 | 2.45 | 2.47 | 39797800 |
| 14 | 2014-1-21 | 2.46 | 2.48 | 2.46 | 2.47 | 33143500 |
| 15 | 2014-1-22 | 2.47 | 2.51 | 2.47 | 2.51 | 52102900 |
| 16 | 2014-1-23 | 2.5 | 2.5 | 2.48 | 2.48 | 40174200 |
| 17 | 2014-1-24 | 2.48 | 2.5 | 2.47 | 2.49 | 36289700 |
| 18 | 2014-1-27 | 2.48 | 2.49 | 2.46 | 2.48 | 43715200 |
| 19 | 2014-1-28 | 2.48 | 2.5 | 2.47 | 2.49 | 29807200 |
| 20 | 2014-1-29 | 2.5 | 2.53 | 2.49 | 2.53 | 39061000 |
| 21 | 2014-1-30 | 2.52 | 2.53 | 2.51 | 2.51 | 19595700 |
| 22 | 2014-1-31 | 2.51 | 2.51 | 2.51 | 2.51 | 0 |
| 23 | 2014-2-3 | 2.51 | 2.51 | 2.51 | 2.51 | 0 |
| 24 | 2014-2-5 | 2.51 | 2.51 | 2.51 | 2.51 | 0 |
| 25 | 2014-2-6 | 2.51 | 2.51 | 2.51 | 2.51 | 0 |
| 26 | 2014-2-7 | 2.5 | 2.5 | 2.48 | 2.49 | 21622600 |
| 27 | 2014-2-10 | 2.5 | 2.52 | 2.49 | 2.51 | 36473700 |
| 28 | 2014-2-11 | 2.51 | 2.56 | 2.5 | 2.55 | 60653200 |
| 29 | 2014-2-12 | 2.54 | 2.56 | 2.53 | 2.54 | 29736800 |
| 30 | 2014-2-13 | 2.54 | 2.57 | 2.53 | 2.55 | 48498800 |
| 31 | 2014-2-14 | 2.55 | 2.55 | 2.53 | 2.54 | 27342300 |
| 32 | 2014-2-17 | 2.54 | 2.56 | 2.54 | 2.55 | 31423800 |
| 33 | 2014-2-18 | 2.55 | 2.56 | 2.53 | 2.54 | 30007000 |
| 34 | 2014-2-19 | 2.54 | 2.67 | 2.53 | 2.63 | 1.06E+08 |
| 35 | 2014-2-20 | 2.62 | 2.64 | 2.6 | 2.61 | 45629300 |
| 36 | 2014-2-21 | 2.6 | 2.61 | 2.56 | 2.58 | 27924300 |
| 37 | 2014-2-24 | 2.57 | 2.58 | 2.51 | 2.53 | 29814500 |
| 38 | 2014-2-25 | 2.53 | 2.55 | 2.51 | 2.54 | 34158800 |

ChinaBank

# ARIMA模型——时间序列

## ➢ 平稳性

- 平稳性就是要求经由样本时间序列所得到的拟合曲线在未来的一段时间内仍然能够按照现有的形态延续下去
- 平稳性要求序列的均值和方差不发生明显变化
  - ✓ 严平稳：序列所有的统计性质（期望，方差）都不会随着时间的推移而发生变化
  - ✓ 宽平稳：期望与相关系数（依赖性）不变，就是说t时刻的值X依赖于过去的信息
- 实际数据大致上都是宽平稳
- 平稳性对于我们分析时间序列至关重要。如果一个时间序列不是平稳的，通常需要通过差分的方式将其转化为平稳时间序列。



## ➢ 差分法实现

- 时间序列在t和t−1时刻的差值。将非平稳序列变平稳。

$$\Delta yx = y(x+1) - y(x), (x = 0,1,2,\dots)$$

- 比如一组数列 [0, 1, 2, 3, 4, 5, 6, 7]
- 进行差分后就会得到新数列 [1, 1, 1, 1, 1, 1, 1]

# ARIMA模型——时间序列



## ➢ 自相关系数（$ACF$）

- 有序的随机变量序列与其自身相比较。自相关系数反映了统一序列在不同时序的取值之间的相关性，对于时间序列$y_t$，$y_t$与$y_{t-k}$的相关系数称为$y_t$间隔k的自相关系数。
- 公式：$ACF(k)=\rho_k=\frac{\mathrm{Cov}(y_t,y_{t-k})}{Var(y_t)}$　　取值范围为$[-1,1]$

## ➢ 偏自相关系数（$PACF$）

- 对于一个平稳$AR(p)$模型，求出滞后k自相关系数$\rho(k)$时，实际上得到并不是$x(t)$与$x(t-k)$之间单纯的相关关系。
- 因为$x(t)$同时还会受到中间$k-1$个随机变量$x(t-1)$，$x(t-2)$，…，$x(t-k+1)$的影响，而这$k-1$个随机变量又都和$x(t-k)$具有相关关系，所以自相关系数里面实际掺杂了其他变量对$x(t)$与$x(t-k)$的影响。
- 为了能单纯测度$x(t-k)$对$x(t)$的影响，引进偏自相关系数$(PACF)$的概念。对于平稳时间序列$\{x(t)\}$，所谓滞后k偏自相关系数指在剔除了中间$k-1$个随机变量$x(t-1)$，$x(t-2)$，…，$x(t-k+1)$的干扰之后，$x(t-k)$对$x(t)$影响的相关程度。
- 公式：$PACF(k)=\frac{COV[(Z_t-\bar{Z}_t),(Z_{t-k}-\bar{Z}_{t-k})]}{\sqrt{var(Z_t-\bar{Z}_t)}\sqrt{var(Z_{t-k}-\bar{Z}_{t-k})}}$

# ARIMA模型——时间序列



## ➢ ADF检验

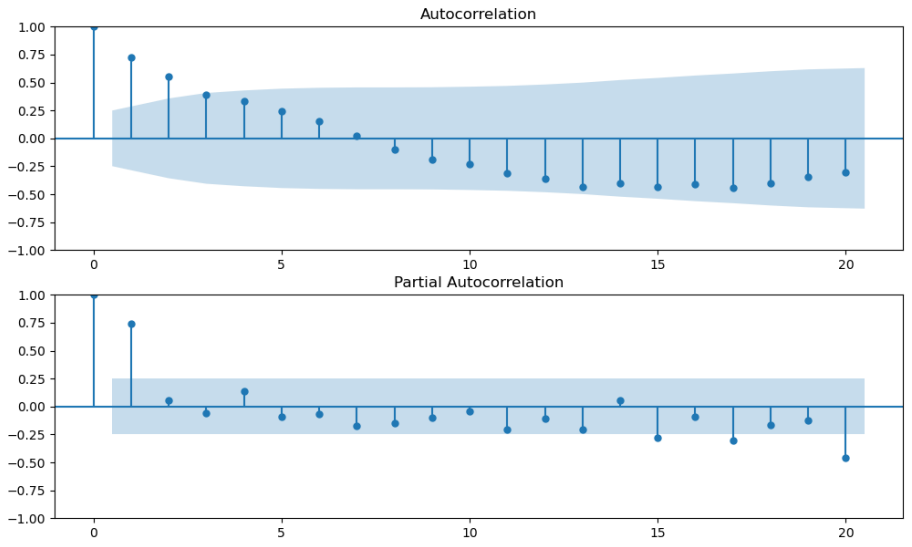
- 对于一个时间序列，如何确定它是否满足平稳性要求？通常采用图检验法（通过时间序列趋势图或者自相关函数图判断）或ADF 检验
- ADF大致的思想就是基于随即游走（不平稳的一个特殊序列）的，对其进行回归，如果发现$p=1$，说明序列满足随机游走，就是非平稳的

```
timeseries_adf :  (0.527919808483182, 0.9856974415734416, 9, 335, {'1%': -3.4500219858626227, '5%': -2.870206553997666, '10%': -2.571387268879483}, -734.0738716811488)
timeseries_diff1_adf :  (-6.177185544979001, 6.58710923976123e-08, 8, 336, {'1%': -3.449962981927952, '5%': -2.870180642420163, '10%': -2.5713734527352607}, -735.8436797171294)
timeseries_diff2_adf :  (-9.202545123160352, 1.9841232339615405e-15, 13, 331, {'1%': -3.4502615951739393, '5%': -2.8703117734117742, '10%': -2.5714433728242714}, -717.2833732193085)
```

- $ADF$检验的结果共有五个参数:
  - ✓ 第一个值： 表示Test Statistic， 即T检验，表示T统计量，假设检验值
  - ✓ 第二个值：p-value，即p值，表示T统计量对应的概率值
  - ✓ 第三/四个值：Lags Used，即表示延迟和测试的次数
  - ✓ 第五个参数{'10%': xxx, '1%': xxx, '5%': xxx}：不同程度拒绝原假设的统计值
- 如何确定该序列是否平稳呢?
  - ✓ 1%、%5、%10不同程度拒绝原假设的统计值和 ADF 假设检验值比较，ADF 假设检验值同时小于1%、5%、10%即说明非常好地拒绝该假设
  - ✓ P-value是否非常接近0

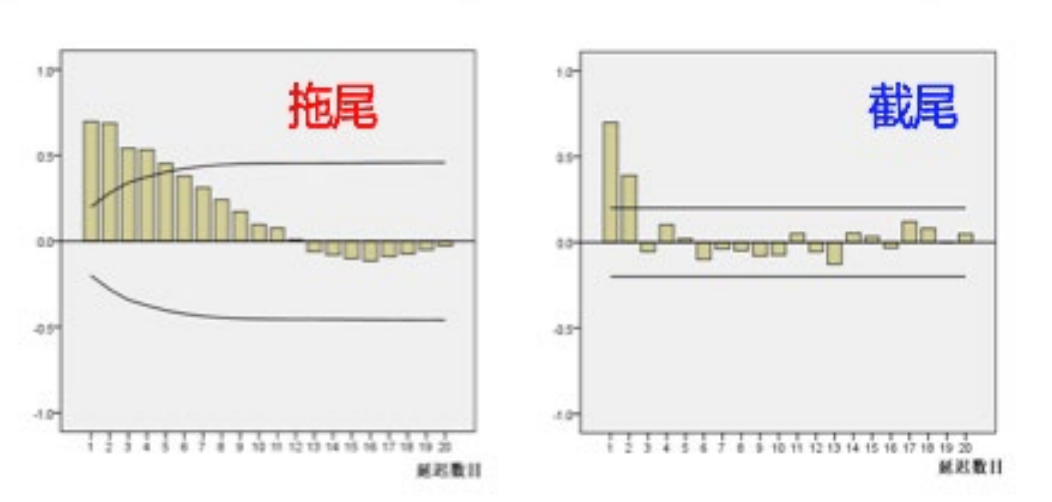
## ➢ 股票价格预测-平稳性检验

- 右图是训练集的ACF和PACF图，由图形可以看出，大部分的值都落在了置信区间内，可以把训练集本身作为平稳序列，无需差分
- 也可通过观察序列趋势图或者ADF检验进行平稳性检验



## ➢ 股票价格预测-确定$p, q$

- 拖尾和截尾：拖尾指序列以指数率单调递减或震荡衰减，而截尾指序列从某个时点变得非常小

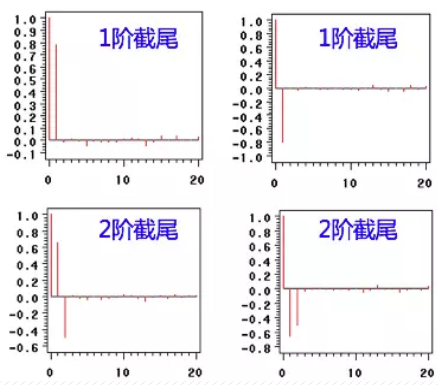


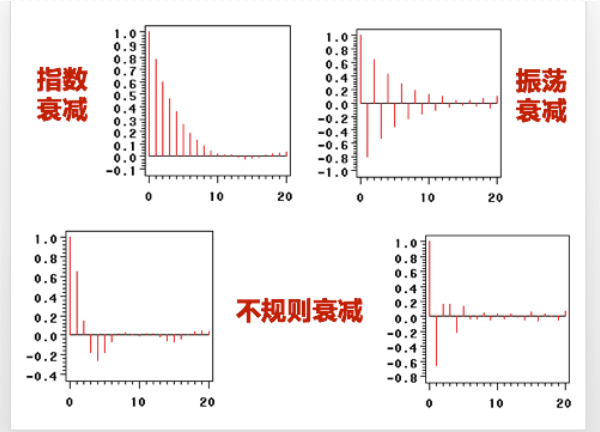
# ARIMA模型——模型建立



➢ **股票价格预测-确定$p, q$**

- **截尾（出现以下情况，通常视为(偏)自相关系数d阶截尾）**
  - ✓ 1）在最初的d阶明显大于2倍标准差范围
  - ✓ 2）之后几乎95%的(偏)自相关系数都落在2倍标准差范围以内
  - ✓ 3）且由非零自相关系数衰减为在零附近小值波动的过程非常突然
- **拖尾（出现以下情况，通常视为(偏)自相关系数拖尾）**
  - ✓ 1）如果有超过5%的样本(偏)自相关系数都落入2倍标准差范围之外
  - ✓ 2）或者是由显著非0的(偏)自相关系数衰减为小值波动的过程比较缓慢或非常连续

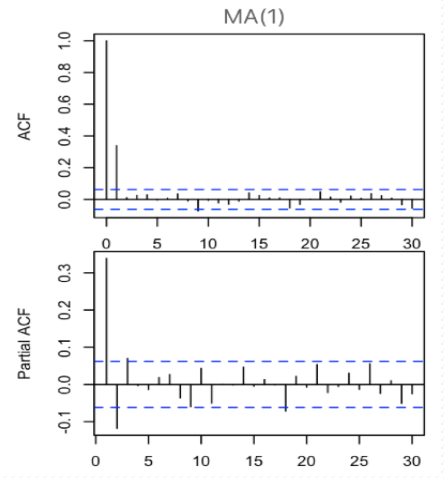


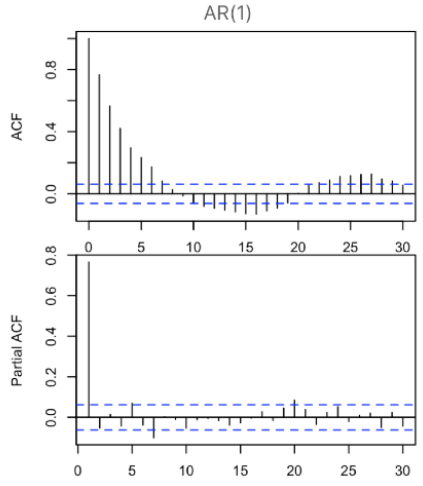
# ARIMA模型——模型建立

➢ **股票价格预测-确定$p, q$**

| | ACF拖尾 | ACF截尾 |
|---|---|---|
| PACF拖尾 | ARMA(p,q) | MA(q) |
| PACF截尾 | AR(p) | 序列本身不存在明显的自相关性，ARMA类模型可能不适用 |

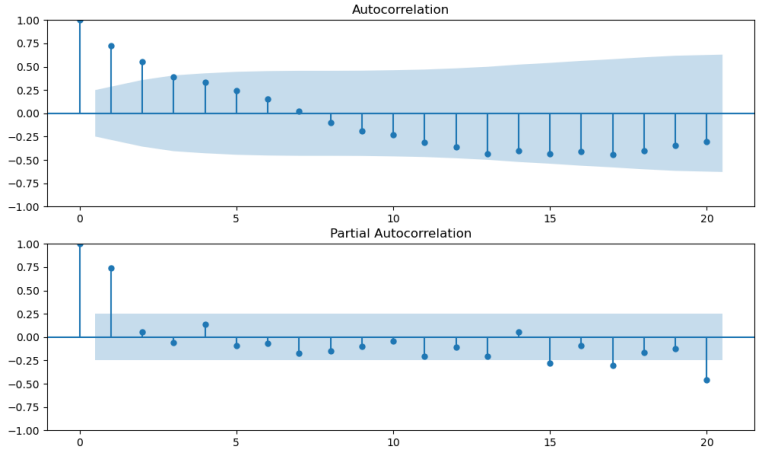



✓ 以超出2倍标准差的最大滞后阶数作为MA模型的阶数，此时ACF超出蓝线的最大滞后阶数为1，MA(1)模型。




✓ 最后一个超出2倍标准差（蓝线）的阶数为1（有1根纵向线超过2倍标准差），故为AR(1)模型。

## ➢ 股票价格预测-确定$p, q$




- 上图为训练集数据对应的ACF和PACF图
- 由图可以确认，ACF拖尾，PACF 1阶截尾，所以选择AR(1)模型，或者说是ARIMA（1，0，0）模型

## ➢ 股票价格预测-确定$p,q$

- 通过拖尾和截尾对模型定阶，具有很强的主观性。回顾一下我们对于模型参数估计得方法，是通过对损失和正则项的加权评估。我们在参数选择的时候，需要平衡预测误差与模型复杂度。我们可以根据信息准则函数法，来确定模型的阶数。这里介绍 AIC、BIC准则
- AIC 准则全称是最小化信息量准则（Akaike Information Criterion）:
  - ✓ AIC=-2ln(L)+2K ，其中 L 表示模型的极大似然函数， K 表示模型参数个数
- AIC 准则存在一定的不足。当样本容量很大时，在 AIC 准则中拟合误差提供的信息就要受到样本容量的放大，而参数个数的惩罚因子却和样本容量没关系（一直是2），因此当样本容量很大时，使用 AIC 准则的模型不收敛于真实模型，它通常比真实模型所含的未知参数个数要多
- BIC（Bayesian InformationCriterion）贝叶斯信息准则弥补了 AIC 的不足:
  - ✓ BIC=-2ln(L)+Kln(n) ，其中 n 表示样本容量。
- 显然，这两个评价指标越小越好。我们通过网格搜索，确定 AIC、BIC 最优的模型（p、q）

# ARIMA模型——模型建立

## ➢ 股票价格预测-确定$p,q$

- 我们以BIC准则为例，确定$p,q$的取值范围为$[0,5]$，通过循环网格搜索所有组合的BIC的值，得到结果如下图


- 可以看到，BIC最小值的组合为('AR1', 'MA0')

```
results_bic.stack().idxmin()
[41]  ✓ 0.0s
..  ('AR1', 'MA0')
```

# ARIMA模型——模型建立

## ➢ 股票价格预测-确定$p,q$

- 我们也可以使用 AIC 和 BIC 准则对训练数据 train 进行 ARMA 模型阶数的选择

```
train_results = sm.tsa.arma_order_select_ic(train, ic=['aic', 'bic'], trend='n', max_ar=8, max_ma=8)

print('AIC', train_results.aic_min_order)
print('BIC', train_results.bic_min_order)
```

✓ 20.6s

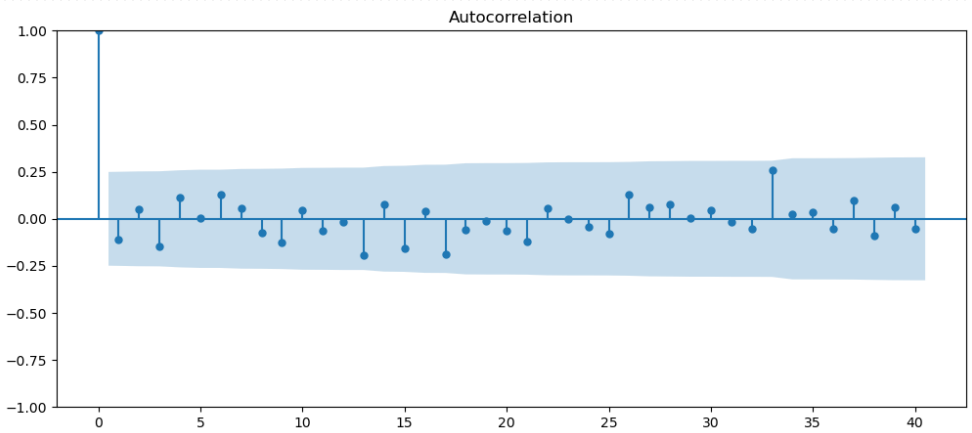
- 这里限制了自回归项（AR）和移动平均项（MA）的最大阶数，将其设置为8

```
AIC (2, 1)
BIC (1, 0)
```

- 最终结果为选用AIC组合，$p,q$为2,1，BIC组合，$p,q$为1,0

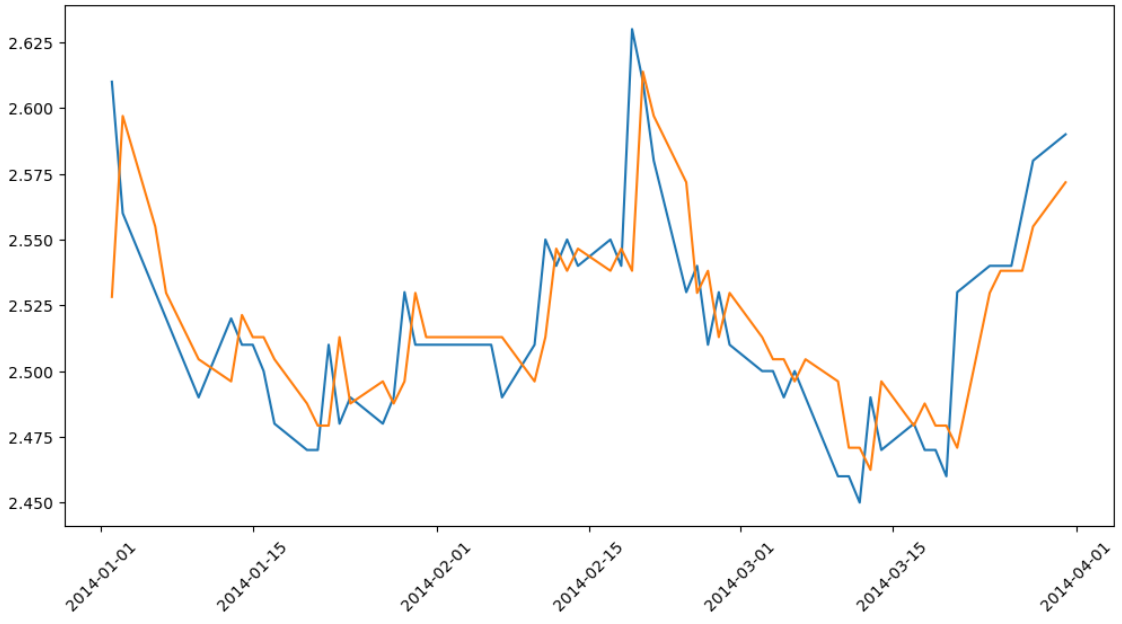
## 股票价格预测-模型检验

- 检验参数估计的显著性（t检验）
- 检验残差序列的随机性，即残差之间是独立的$e_i = y_i - \hat{y}_i$
- 残差序列的随机性可以通过自相关函数法来检验，即做残差的自相关函数图
- 从ACF图中可以看出残差之间独立性比较高



# ARIMA模型——模型建立

- 股票价格预测-模型预测



# ARIMA模型——相关代码

- 见文件ARIMA.ipynb

# 欢迎关注数模加油站

## THANKS