TimeSeriesAnalysis

Castor

2024-11-04

Table of contents

Preface

关于本网页

本书为北京大学数学科学学院金融数学系金融数学应用硕士《金融时间序列分析》（李东风）的阅读笔记，原文示例
R 语言进行编写，本文实现了 Python 语言版本。

原文链接：《金融时间序列分析》

本书内容

- □ 收益率，债券，波动率，金融数据示例，收益率分布性质，金融数据可视化，统计分布复习;
- □ 线性时间序列：款平稳，自相关系数函数，ACF的白噪声检验， AR,
  偏相关系数，定阶与参数估计，预测，MA， ARMA, ARIMA，单位根过程，单位根检验，
  指数平滑方法，季节模型， 回归模型的序列相关误差项，协整，
  长记忆模型，模型比较与模型平均 线性时间序列的案例研究;
- □ 资产波动率，ARCH效应，ARCH模型，GARCH模型，IGARCH模型， GARCH-
  M模型，EGARCH模型, TGARCH模型, APARCH模型， 非对称GARCH模型，随机波动率模型
  波动率模型案例研究;
- □ 多元时间序列的基础知识和VAR模型， 协整和协整检验，格兰杰因果性;
- □ 状态空间模型介绍。

Chapter 1

金融数据及其特征

课程采用蔡瑞胸（Ruey S. Tsay）的《金融数据分析导论：基于R语言》（Tsay 2013）（An Introduction to Analysis of Financial Data with R）作为主要教材之一。 这是第一章金融数据及其特征的授课笔记。

Chapter 2

线性时间序列模型

2.1 介绍

课程采用蔡瑞胸（Ruey S. Tsay）的《金融数据分析导论：基于R语言》(Tsay 2013) （An Introduction to Analysis of Financial Data with R）作为主要教材之一。“线性时间序列模型”这一部分是教材的第二章和第三章的授课笔记，本章讲授时间序列的线性模型，包括：

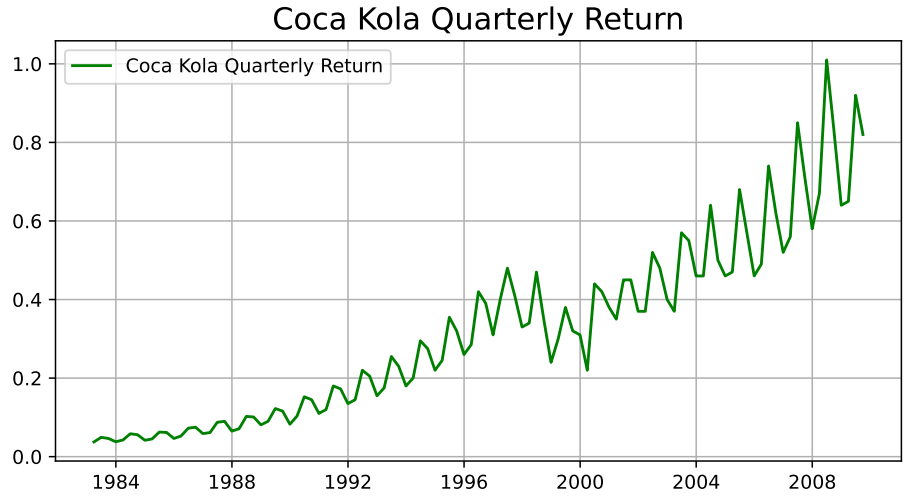
- ☐ 一些基本概念
- ☐ AR, MA, ARMA模型
- ☐ 单位根过程
- ☐ 指数平滑
- ☐ 季节模型
- ☐ 回归模型中误差项有序列相关的处理
- ☐ 长记忆的分数阶差分模型
- ☐ 模型比较
- ☐ 实例分析

2.1.1 例子：可口可乐公司盈利数据

序列仍体现出缓慢的、非单调的上升趋势，又有明显的每年的周期变化（称为季节性），还有短期的波动：

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np

raw_data = []
with open("../ftsdata/q-ko-earns8309.txt", "r", encoding="utf-8") as file:
    for line in file.readlines():
        line = line.strip("\n").strip(" ").replace("\t", " ").split(" ")
```

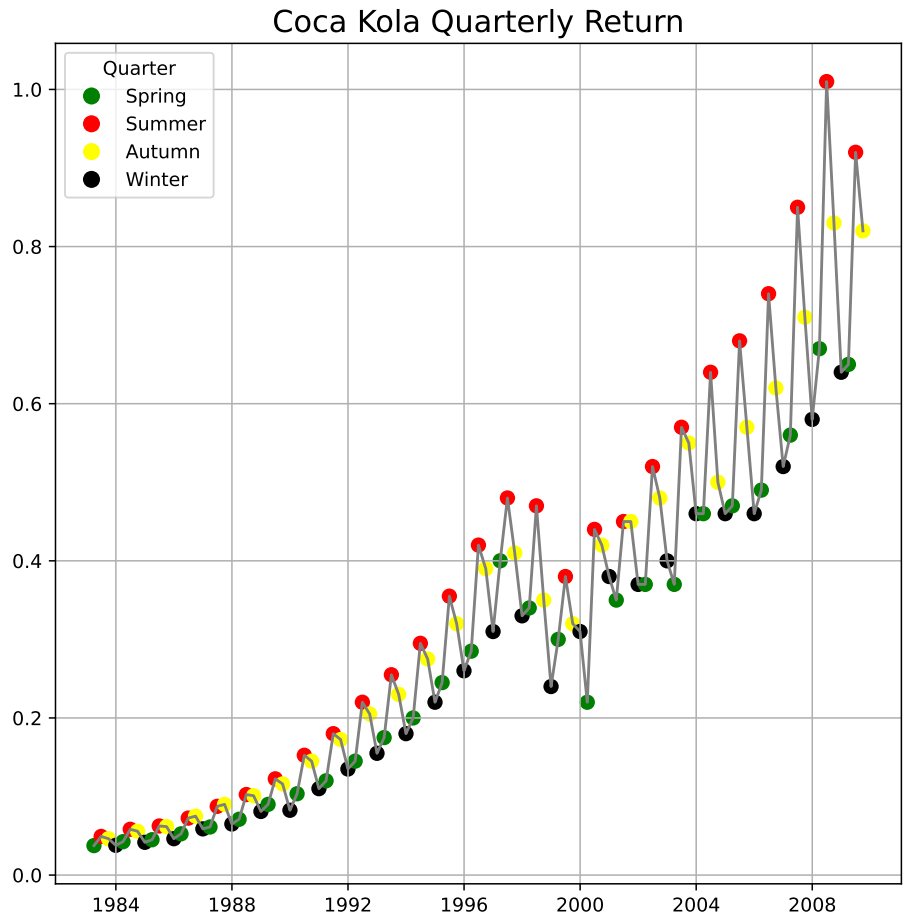
```
        line = list(filter(lambda x: x != "", line))
        raw_data.append(line)
data = pd.DataFrame(raw_data[1:], columns=raw_data[0])


data["pends"] = pd.to_datetime(data["pends"], format="%Y%m%d")
data["anntime"] = pd.to_datetime(data["anntime"], format="%Y%m%d")
data["value"] = pd.to_numeric(data["value"])

plt.figure(figsize=(8, 4))
plt.plot(data["pends"], data['value'], label='Coca Kola Quarterly Return', color='green
plt.title('Coca Kola Quarterly Return', fontsize=16)
plt.grid(True)
plt.legend()
plt.show()
```



标出不同季节的数据点，可以看出，每年一般冬季和春季最低， 夏季最高，秋季介于夏季和冬季之间：

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np

raw_data = []
with open("../ftsdata/q-ko-earns8309.txt", "r", encoding="utf-8") as file:
    for line in file.readlines():
        line = line.strip("\n").strip(" ").replace("\t", " ").split(" ")
        line = list(filter(lambda x: x != "", line))
        raw_data.append(line)
data = pd.DataFrame(raw_data[1:], columns=raw_data[0])
```

```
data["pends"] = pd.to_datetime(data["pends"], format="%Y%m%d")
data["anntime"] = pd.to_datetime(data["anntime"], format="%Y%m%d")
data["value"] = pd.to_numeric(data["value"])

data['Date'] = pd.to_datetime(data['pends'])
data.set_index('Date', inplace=True)

data['Year'] = data.index.year
data['Quarter'] = data.index.quarter

cpal = ['green', 'red', 'yellow', 'black']

plt.figure(figsize=(8, 8))

plt.plot(data.index, data['value'], label='Coca Kola Quarterly Return', color='gray')

for i, row in data.iterrows():
    plt.scatter(row.name, row['value'], color=cpal[row['Quarter'] - 1], s=50)

plt.title('Coca Kola Quarterly Return', fontsize=16)
plt.grid(True)

quarter_labels = ['Spring', 'Summer', 'Autumn', 'Winter']
plt.legend([plt.Line2D([0], [0], marker='o', color='w', markerfacecolor=cpal[i], markersize=10) f
           quarter_labels,
           title='Quarter')

plt.show()
```

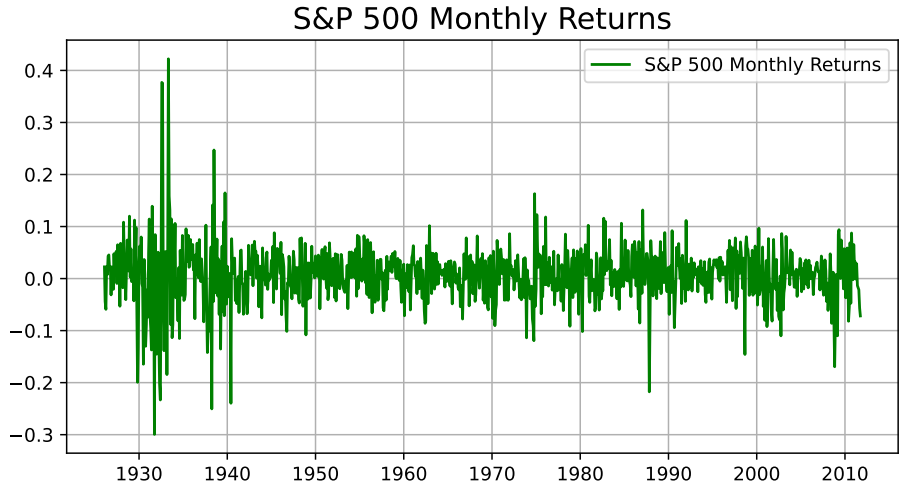


2.1.2 例子：标普500指数月对数收益率

收益率在0上下波动，除了个别时候基本在某个波动范围之内：

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np

raw_data = []
with open("../ftsdata/m-ibmsp-2611.txt", "r", encoding="utf-8") as file:
    for line in file.readlines():
        line = line.strip("\n").strip(" ").replace("\t", " ").split(" ")
        line = list(filter(lambda x: x != "", line))
        raw_data.append(line)
data = pd.DataFrame(raw_data[1:], columns=raw_data[0])
```

```
data["date"] = pd.to_datetime(data["date"], format="%Y%m%d")
data["ibm"] = pd.to_numeric(data["ibm"])
data["sp"] = pd.to_numeric(data["sp"])
data.head()

plt.figure(figsize=(8, 4))
plt.plot(data["date"], data['sp'], label='S&P 500 Monthly Returns', color='green')
plt.title('S&P 500 Monthly Returns', fontsize=16)
plt.grid(True)
plt.legend()
plt.show()
```



2.2 平稳性

例子：标普500指数月对数收益率所展示的收益率曲线基本呈现一个在水平线（一般是0）上下波动，且波动范围基本不变，这样的表现是时间序列【弱平稳序列】的表现。

例子：可口可乐公司盈利数据所展示的价格曲线呈现出水平的上下起伏，如果分成几段的话，各段的平均值差距较大

时间序列：设有随机变量序列 $\{x_t, t = \dots, -2, -1, 0, 1, 2, \dots\}$, 称其为一个时间序列。其中$x_t$是一个随机变量，也可以$X(t, \omega), t \in \mathbb{Z}$($\mathbb{Z}$表示所有整数组成的集合)，$\omega \in \Omega$，$\Omega$表示在一定的条件下所有可能的试验结果的集合。经济和金融Ω对应的结果，称为一条“轨道”。而针对随机变量的许多理论性质都是在$\omega \in \Omega$上讨论的，比如$EX_t = \int X_t(\omega)P(d\omega)$是$X_t(\omega)$对$\omega \in \Omega$的平均。

为了能够用一条轨道的观测样本得到所有$\omega \in \Omega$的性质，需要时间序列满足“遍历性”。

时间序列的样本：设$\{x_t, t = 1, 2, \dots, T\}$是时间序列中的一段。仍将$x_t$看成随机变量，也可以写成大写的$X_t$。如果有了具体数值，那么样本就是一条轨道中的一段。

自协方差函数：时间序列$\{X_t\}$中两个随机变量的协方差 $Cov(X_s, X_t)$叫做自协方差。如果

$Cov(X_s, X_t) = \gamma_{|t-s|}$仅依赖于$t-s$, 则称

$$\gamma_k = \text{Cov}(X_{t-k}, X_t), k = 0, 1, 2, ...$$

为时间序列$\{X_t\}$的自协方差函数。因为$\text{Cov}(X_s, X_t) = \text{Cov}(X_t, X_s)$, 所以$\gamma_{-k} = \gamma_k$。易见 $\gamma_0 = \text{Var}(X_t)$。

由Cauchy-Schwartz不等式:

$$|\gamma_k| = |E[(X_{t-k} - \mu)(X_t - \mu)]| \leq \left(E(X_{t-k} - \mu)^2 \, E(X_t - \mu)^2\right)^{1/2} = \gamma_0$$

弱平稳序列(宽平稳序列, weakly stationary time series): 如果时间序列$\{X_t\}$存在有限的二阶矩且满

1. $EX_t = \mu$与t无关;
2. $\text{Var}(X_t) = \gamma_0$与$t$无关;
3. $\gamma_k = \text{Cov}(X_t - k, X_t), \ k = 1, 2, ...$与$t$无关,

则称$\{X_t\}$为弱平稳序列。

适当条件下可以用时间序列的样本估计自协方差函数, 这是用一条轨道的信息推断所有实验结果Ω, 估计式为:

$$\hat{\gamma}_k = \frac{1}{T} \sum_{t=k+1}^{T} (x_{t-k} - \bar{x})(x_t - \bar{x}), k = 0, 1, ..., T-1$$

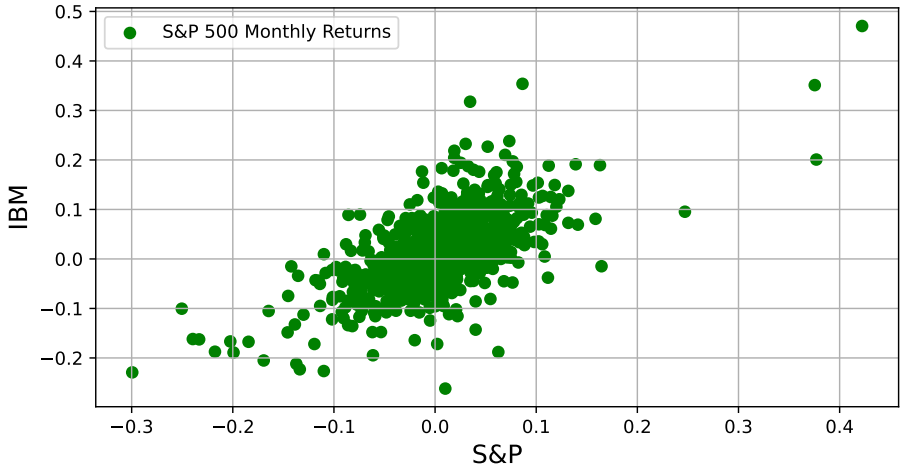
称$\hat{\gamma}_k$为样本自协方差。 注意这里用了$1/T$而不是$1/(T-k)$, 用$1/(T-k)$在获得无偏性的同时会造成一 些理论上的困难。

2.3 相关系数和自相关系数

2.3.1 相关系数

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np

raw_data = []
with open("../ftsdata/m-ibmsp-2611.txt", "r", encoding="utf-8") as file:
    for line in file.readlines():
        line = line.strip("\n").strip(" ").replace("\t", " ").split(" ")
        line = list(filter(lambda x: x != "", line))
        raw_data.append(line)
data = pd.DataFrame(raw_data[1:], columns=raw_data[0])

data["date"] = pd.to_datetime(data["date"], format="%Y%m%d")
data["ibm"] = pd.to_numeric(data["ibm"])
data["sp"] = pd.to_numeric(data["sp"])
```

```
data.head()

plt.figure(figsize=(8, 4))
plt.scatter(data["sp"], data['ibm'], label='S&P 500 Monthly Returns', color='green')
plt.xlabel("S&P", fontsize=14)
plt.ylabel("IBM", fontsize=14)
plt.grid(True)
plt.legend()
plt.show()
```



上图是IBM股票月度简单收益率对标普500收益率的散点图，可以看出两者有明显的正向相关关系。

两个随机变量X和Y的相关系数定义为：

$$\rho(X,Y) = \rho_{xy} = \frac{\mathsf{Cov}(X,Y)}{\sqrt{\mathsf{Var}(X)\mathsf{Var}(Y)}} = \frac{E[(X-\mu_x)(Y-\mu_y)]}{\sqrt{E(X-\mu_x)^2 E(Y-\mu_y)^2}}$$

如果有(X,Y)的独立同分布样本$(x_t, y_t),\ t = 1, 2, \ldots, T,$ 可估计相关系数（皮尔逊，Pearson）为：

$$\hat{\rho}_{xy} = \frac{\sum_{t=1}^{T}(x_t - \bar{x})(y_t - \bar{y})}{\sqrt{\sum_{t=1}^{T}(x_t - \bar{x})^2 \sum_{t=1}^{T}(y_t - \bar{y})^2}}$$

对于不独立的样本，比如时间序列样本，也可以计算相关系数，其估计合理性需要一些模型假设。
对于联合分布非正态的情况，有时相关系数不能很好地反映X和Y的正向或者负向的相关。
斯皮尔曼（Spearman）相关系数是计算X的样本的秩(名次)与Y的样本的秩之间的相关系数，也称为Spearmam rank correlation。

另一种常用的非参数相关系数是肯德尔tau(Kendall’s τ)系数，反映了一致数对和非一致数对之间的差对随机向量(X,Y),设$(X_1,Y_1),(X_2,Y_2)$相互独立且联合分布与(X,Y)联合分布相同，定义X和Y的肯德尔tau系数为：

$$\tau = P\left[(X_1 - X_2)(Y_1 - Y_2) > 0\right] - P\left[(X_1 - X_2)(Y_1 - Y_2) < 0\right]$$

即两个观测的分量次序一致的概率减去分量次序相反的概率。一致的概率越大，说明两个的正向相关性越强。

对 IBM 收益率与标普收益率数据计算这三种相关系数：

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np

raw_data = []
with open("../ftsdata/m-ibmsp-2611.txt", "r", encoding="utf-8") as file:
    for line in file.readlines():
        line = line.strip("\n").strip(" ").replace("\t", " ").split(" ")
        line = list(filter(lambda x: x != "", line))
        raw_data.append(line)
data = pd.DataFrame(raw_data[1:], columns=raw_data[0])

data["date"] = pd.to_datetime(data["date"], format="%Y%m%d")
data["ibm"] = pd.to_numeric(data["ibm"])
data["sp"] = pd.to_numeric(data["sp"])
data.head()

pearson_corr = data['ibm'].corr(data['sp'])
spearman_corr = data['ibm'].corr(data['sp'], method='spearman')
kendall_corr = data['ibm'].corr(data['sp'], method='kendall')

print("Pearson correlation:", pearson_corr)
print("Spearman correlation:", spearman_corr)
print("Kendall correlation:", kendall_corr)
```

```
Pearson correlation: 0.6395978546773113
Spearman correlation: 0.6065788974589758
Kendall correlation: 0.4328065703413303
```

2.3.2 自相关函数与白噪声

设$\{X_t\}$为弱平稳序列，$\{\gamma_k\}$为自协方差函数。则

$$\rho(X_{t-k}, X_t) = \frac{\mathsf{Cov}(X_{t-k}, X_t)}{\sqrt{\mathsf{Var}(X_{t-k})\,\mathsf{Var}(X_t)}} = \frac{\gamma_k}{\sqrt{\gamma_0\gamma_0}} = \frac{\gamma_k}{\gamma_0},\ k = 0, 1, \ldots,\ \forall t$$

记$\rho_k = \gamma_k/\gamma_0$, 这是$X_t - k$与$X_t$的相关系数且与$t$无关，称$\{\rho_k, k = 0, 1, ...\}$为时间序列$\{X_t\}$的自相关函数（Autocorrelation function, ACF)。$\rho_0 = 1$。

如果弱平稳序列$\{X_t\}$满足$\rho_k = 0, k = 1, 2, ...$, 称$\{X_t\}$为白噪声序列。如果随机变量序列$\{X_t\}$独立且期望和方差不随

适当条件下ρ_k可以从时间序列样本估计为：

$$\hat{\rho}_k = \frac{\hat{\gamma}_k}{\hat{\gamma}_0},\ k = 0, 1, ...$$

$\hat{\rho}_0 = 1$。称$\hat{\rho}_k, k = 1, 2, ...$为样本自相关函数。

如果时间序列严平稳遍历，则$\hat{\rho}_k$是ρ_k的强相合估计。

若$\{X_t\}$为有二阶矩的独立同分布随机变量列，则 $\hat{\rho}_k (k > 0)$渐近服从$\mathrm{N}(0, \frac{1}{T})$。

如果$\{\varepsilon_t\}$是零均值独立同分布白噪声，q为非负整数，$\{\psi_j, j = 0, 1, ..., q\}$是数列，$\psi_0 = 1$,

$$X_t = \mu + \sum_{j=0}^{q} \psi_j \varepsilon_{t-j},\ t \in \mathbb{Z},$$

则从$\{X_t, t = 1, ..., T\}$计算的ACF满足：当$k > q$时，$\sqrt{T}\hat{\rho}_k$渐近服从$\mathbb{N}(0, 1 + 2\sum_{j=0}^{q} \rho_j^2)$, 这称为Bartlett公式。

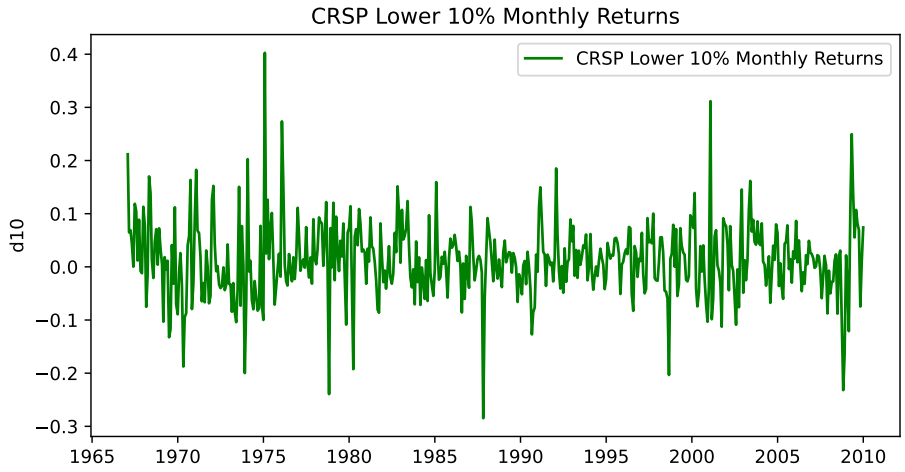
2.3.2.1 例子：CRSP 的第10分位组合的月度收益率

第10分位组合是NYSE、AMEX、NASDAQ市值最小的10%股票组成的投资组合，每年都重新调整。

- ☐ CRSP是Center for Research in Security Prices, 位于Chicago Booth。
- ☐ NYSE(The New York Stock Exchange, 纽约证券交易所)，
- ☐ AMEX(American Stock Exchange, 美国证券交易所，在纽约华尔街附近)，
- ☐ NASDAQ(National Association of Securities Dealers Automated Quotations，纳斯达克，位于纽约)。

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np

raw_data = []
with open("../ftsdata/m-dec12910.txt", "r", encoding="utf-8") as file:
    for line in file.readlines():
        line = line.strip("\n").strip(" ").replace("\t", " ").split(" ")
        line = list(filter(lambda x: x != "", line))
        raw_data.append(line)
data = pd.DataFrame(raw_data[1:], columns=raw_data[0])
```

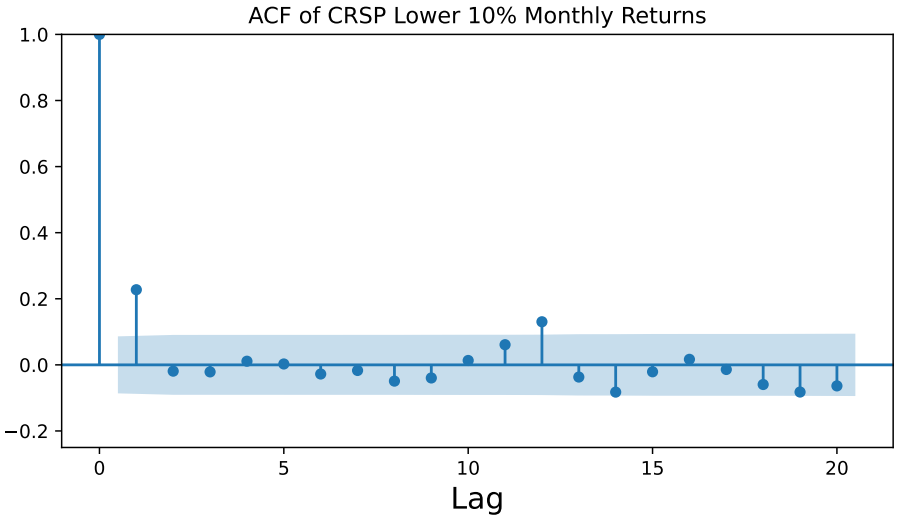
```
data["date"] = pd.to_datetime(data["date"], format="%Y%m%d")
data.set_index("date", inplace=True)
data = data.apply(pd.to_numeric)
data.head()

plt.figure(figsize=(8, 4))
plt.plot(data["dec10"], label='CRSP Lower 10% Monthly Returns', color="green")
plt.title('CRSP Lower 10% Monthly Returns')
plt.ylabel('d10')
plt.legend()
plt.show()
```



绘制时间序列的自相关函数图（ACF）：

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np

raw_data = []
with open("../ftsdata/m-dec12910.txt", "r", encoding="utf-8") as file:
    for line in file.readlines():
        line = line.strip("\n").strip(" ").replace("\t", " ").split(" ")
        line = list(filter(lambda x: x != "", line))
        raw_data.append(line)
data = pd.DataFrame(raw_data[1:], columns=raw_data[0])

data["date"] = pd.to_datetime(data["date"], format="%Y%m%d")
data.set_index("date", inplace=True)
data = data.apply(pd.to_numeric)
data.head()
```

```
from statsmodels.graphics.tsaplots import plot_acf

plt.figure(figsize=(8, 4))
ax = plt.gca()  #
plot_acf(data["dec10"], ax=ax, lags=20)
ax.set_ylim(-0.25, 1)
ax.set_xlabel('Lag', fontsize=16)  #
plt.title('ACF of CRSP Lower 10% Monthly Returns')
plt.show()
```



在 ACF 图像中，Lag 相当于 k，ACF 相当于 $\hat{\rho}_k$。

ACF 图中横轴上下两条水平下是在独立同分布白噪声假设下的加减两倍标准差，即 $\pm\frac{2}{\sqrt{T}}$。如果独立同分布白噪声假设成立，每个 $\hat{\rho}_k$ 有 95% 以上的概率落入这两条线之间。

ACF 图 $k=0$ 处总对应 $\hat{\rho}_0=1$。

上图的 $\hat{\rho}_1$ 和 $\hat{\rho}_{12}$ 都超出了界限(因为是月度数目，横轴的单位是1/12为一个时间点)。从此图可以认为此投资组合的

2.3.3 用单个自相关系数作白噪声检验

如果$\{X_t\}$是独立同分布白噪声，则 $\hat{\rho}_k(k \geq 1)$ 近似 $N(0,1/T)$。若 H_0 是序列为白噪声，取统计量：

$$t=\sqrt{T}\hat{\rho}_k$$

如果$|t| > \text{qnorm}(1-\alpha/2)$，则拒绝白噪声零假设。实际中常取$\alpha = 0.05$, $\text{qnorm}(1-\alpha/2)\approx 2$, 当$\hat{\rho}_1$超出$\pm 2/\sqrt{T}$则拒绝$H_0$, 有多个$\hat{\rho}_k$超出$\pm 2/\sqrt{T}$也可拒绝$H_0$, 有一个$t$统计量值很大(比如超出

在判断$\{X_t\}$是否$X_t = \mu + \sum_{j=0}^{q} \psi_j \varepsilon_{t-j}$这样的模型时，根据 Bartlet 公式，可取：

$$t = \frac{\hat{\rho}_k}{\sqrt{\frac{1}{T}\left(1 + 2\sum_{j=1}^{k-1} \hat{\rho}_j^2\right)}}, k > q$$

当t超出 qnorm$(1-\alpha/2)$时拒绝这样的模型。

2.3.3.1 TODO：例子 3.3

2.3.4 Ljung-Box 白噪声检验

为了检验时间序列样本是否来自白噪声序列，可以检验 $\rho_k = 0, k = 1, 2, \cdots$ 的零假设。前面检验单个 ρ_k 的做法如果正对多个进行检验就有多重检验的第一类错误增大的问题。

Box 和 Pierce(G. Box & Pierce, 1970) 提出了混成统计量 (Portmanteau statistic)：

$$Q_*(m) = T\sum_{j=1}^{m} \hat{\rho}_j^2$$

用来检验：

$$H_0 : \rho_1 = \cdots = \rho_m = 0 \longleftrightarrow H_a : \text{不全为零}$$

在$\{X_t\}$是独立白噪声序列条件下，$Q_*(m)$近似服从$\chi^2(m)$分布。给定检验水平α, 当$Q_*(m) >$qchisq$(1-\alpha, m)$ 时拒绝H_0, 否定白噪声假设。如果检验的序列是线性时间序列估计的残差序列，则卡方自由度应改

Ljung 和 Box(Ljung & Box, 1978) 对此检验方法进行了改进。统计量改为：

$$Q(m) = T(T+2)\sum_{j=1}^{m} \frac{\hat{\rho}_j^2}{T-j}$$

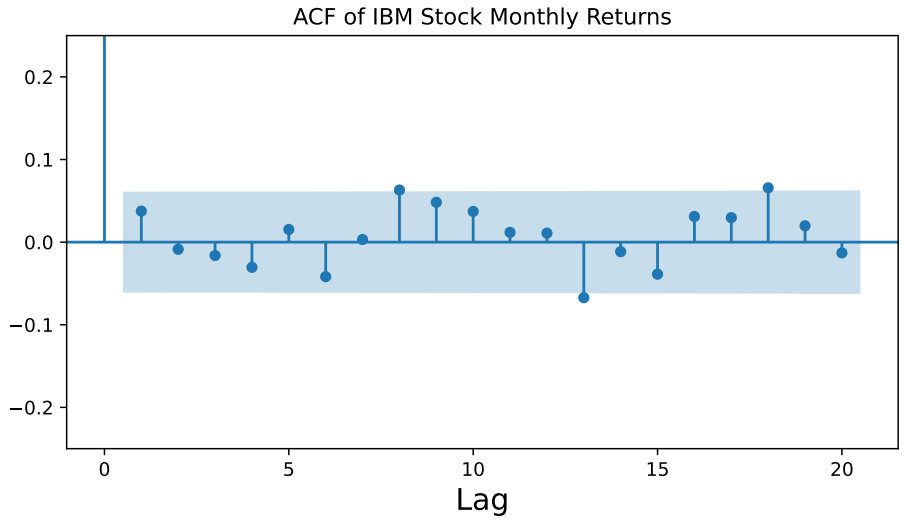
在独立同分布白噪声假设下仍近似服从$\chi^2(m)$分布。当$Q(m) >$qchisq$(1-\alpha, m)$时拒绝H_0, 否定白噪声假设。这个检验称为 Ljung-Box 白噪声检验。如果检验的序列是线性时间序 ARMA(p, q)模型建模的残差作白噪声检验，卡方自由度应改为$m-(p+q)$。

2.3.4.1 例子：检验 IBM 股票月收益率是否为白噪声

绘制数据的 ACF 图像：

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
```

```
raw_data = []
with open("../ftsdata/m-ibmsp-2611.txt", "r", encoding="utf-8") as file:
    for line in file.readlines():
        line = line.strip("\n").strip(" ").replace("\t", " ").split(" ")
        line = list(filter(lambda x: x != "", line))
        raw_data.append(line)
data = pd.DataFrame(raw_data[1:], columns=raw_data[0])

data["date"] = pd.to_datetime(data["date"], format="%Y%m%d")
data.set_index("date", inplace=True)
data = data.apply(pd.to_numeric)
data.head()

from statsmodels.graphics.tsaplots import plot_acf

plt.figure(figsize=(8, 4))
ax = plt.gca()  #
plot_acf(data["ibm"], ax=ax, lags=20)
ax.set_ylim(-0.25, 0.25)
ax.set_xlabel('Lag', fontsize=16)  #
plt.title('ACF of IBM Stock Monthly Returns')
plt.show()
```



从 ACF 图来看，月度收益率是白噪声。

作 Ljung-Box 白噪声检验，分别取 $m = 12$ 和 $m = 24$：

```
import pandas as pd
import matplotlib.pyplot as plt
```

```
import numpy as np

raw_data = []
with open("../ftsdata/m-ibmsp-2611.txt", "r", encoding="utf-8") as file:
    for line in file.readlines():
        line = line.strip("\n").strip(" ").replace("\t", " ").split(" ")
        line = list(filter(lambda x: x != "", line))
        raw_data.append(line)
data = pd.DataFrame(raw_data[1:], columns=raw_data[0])

data["date"] = pd.to_datetime(data["date"], format="%Y%m%d")
data.set_index("date", inplace=True)
data = data.apply(pd.to_numeric)
data.head()

from statsmodels.stats.diagnostic import acorr_ljungbox


#   Ljung-Box
lb_test_12 = acorr_ljungbox(data["ibm"], lags=[12], return_df=True)
print(lb_test_12)

lb_test_24 = acorr_ljungbox(data["ibm"], lags=[24], return_df=True)
print(lb_test_24)
```

```
      lb_stat  lb_pvalue
12  13.097984   0.361959
      lb_stat  lb_pvalue
24  35.384127   0.062905
```

可以看出，在 0.05 水平下均不拒绝零假设，支持 IBM 月度简单收益率是白噪声的零假设。

从简单收益率计算对数收益率，并进行 LB 白噪声检验：

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np

raw_data = []
with open("../ftsdata/m-ibmsp-2611.txt", "r", encoding="utf-8") as file:
    for line in file.readlines():
        line = line.strip("\n").strip(" ").replace("\t", " ").split(" ")
        line = list(filter(lambda x: x != "", line))
        raw_data.append(line)
data = pd.DataFrame(raw_data[1:], columns=raw_data[0])

data["date"] = pd.to_datetime(data["date"], format="%Y%m%d")
```

```
data.set_index("date", inplace=True)
data = data.apply(pd.to_numeric)
data.head()

#   Ljung-Box
lb_test_12 = acorr_ljungbox(np.log(data["ibm"] + 1), lags=[12], return_df=True)
print(lb_test_12)

lb_test_24 = acorr_ljungbox(np.log(data["ibm"] + 1), lags=[24], return_df=True)
print(lb_test_24)
```

```
      lb_stat  lb_pvalue
12  12.814366   0.382677
      lb_stat  lb_pvalue
24  34.505798   0.076073
```

可以看出，在 0.05 水平下不拒绝零假设。

2.3.4.2 例子：检验 CRSP 的第10分位组合的月度收益率是否为白噪声

这是［例子：CRSP 的第10分位组合的月对数收益率］的继续，对其数据作白噪声检验。

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np

raw_data = []
with open("../ftsdata/m-dec12910.txt", "r", encoding="utf-8") as file:
    for line in file.readlines():
        line = line.strip("\n").strip(" ").replace("\t", " ").split(" ")
        line = list(filter(lambda x: x != "", line))
        raw_data.append(line)
data = pd.DataFrame(raw_data[1:], columns=raw_data[0])

data["date"] = pd.to_datetime(data["date"], format="%Y%m%d")
data.set_index("date", inplace=True)
data = data.apply(pd.to_numeric)
data.head()

from statsmodels.stats.diagnostic import acorr_ljungbox


#   Ljung-Box
lb_test_12 = acorr_ljungbox(data["dec10"], lags=[12], return_df=True)
print(lb_test_12)
```

```
lb_test_24 = acorr_ljungbox(data["dec10"], lags=[24], return_df=True)
print(lb_test_24)
```

```
      lb_stat  lb_pvalue
12  41.059699   0.000048
      lb_stat  lb_pvalue
24  56.245617   0.000212
```

可以看出，在 0.05 水平下拒绝零假设，认为 CRSP 第10分位组合的月度简单收益率不是白噪声。

有效市场假设认为收益率是不可预测的，也就不会有非零的自相关。但是，股价的决定方式和指数收盗

常见的白噪声检验还有 TREVOR S. BREUSCH (1978) 和 LESLIE G. GODFREY (1978) 提出的拉格朗日乘子法检验(LM 检验)。零假设为白噪声，对立假设为 AR、MA 或者 ARMA。

2.4 线性时间序列

设$\{X_t\}$是独立同分布的二阶矩有限的随机变量，称$\{X_t\}$为独立同分布白噪声(white noise)。最常用的白噪声一般假设均值为零。如果$\{X_t\}$独立同$N(0,\sigma^2)$分布，称$\{X_t\}$为高斯(Gaussian) 白噪声或正态白噪声。

白噪声序列的自相关函数为零($\rho_0 = 1$除外)。

实际应用中如果样本自相关函数近似为零(ACF 图中都位于控制线之内或基本不超出控制线)，则可认为该噪声的样本。

设$\{\varepsilon_t\}$是零均值独立同分布白噪声，$\mathrm{Var}(\varepsilon_t) = \sigma^2$，数列$\{\psi_j\}$满足$\sum_j \psi_j^2 < \infty, \psi_0 = 1$，令：

$$X_t = \mu + \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}, t \in \mathbb{Z}$$

则称$\{X_t\}$是(因果)线性时间序列，ε_t代表了在时刻t增加的变动信息，称为新息(innovation)或者扰动(shock)。因果线性时间序列满足新息$\varepsilon_{t+j}(j \geq 1)$与历史的$X_t, X_{t-1}$独立。而且，$\{X_t - 1, X_{t-2}, ...\}$与$\{\varepsilon_t - 1, \varepsilon_{t-2}, ...\}$可以互相线性地表示。

这个定义可以放宽到$\{\varepsilon_t\}$是宽平稳的不相关列(宽白噪声)的情形。这时，称(3.1)为$\{X_t\}$的 Wold 分解，称$\{X_t\}$ 为纯非决定性序列。

许多弱平稳时间序列是线性时间序列，后面讲到的 AR 模型、MA 模型、ARMA 模型都属于线性时间序列。另外的 许多弱平稳时间序列可以被线性时间序列近似。非平稳的时间序列不

$\{\psi_j\}$称为模型的ψ权重。易见：

$$EX_t = \mu, \quad \mathrm{Var}(X_t) = \sigma^2 \sum_{j=0}^{\infty} \psi_j^2,$$

自协方差函数为：

$$\begin{aligned}\gamma_k =&\mathtt{Cov}(X_t, X_{t-k}) = E\left[\left(\sum_{i=0}^{\infty}\psi_i\varepsilon_{t-i}\right)\left(\sum_{i=0}^{\infty}\psi_i\varepsilon_{t-i-l}\right)\right]\\ =&E\left(\sum_{i,j=0}^{\infty}\psi_i\psi_j\varepsilon_{t-i}\varepsilon_{t-j-k}\right) = \sigma^2\sum_{i,j=0}^{\infty}\psi_i\psi_j\delta_{i-j-k}\\ =&\sigma^2\sum_{j=0}^{\infty}\psi_j\psi_{j+k}\end{aligned}$$

其中δ_k当$k=0$时为 1，当$k\neq 0$时为0。

自相关函数为：

$$\rho_k = \frac{\gamma_k}{\gamma_0} = \frac{\sum_{j=0}^{\infty}\psi_j\psi_{j+k}}{1+\sum_{j=1}^{\infty}\psi_j^2},\ k\geq 0$$

线性时间序列模型满足$\psi_j\to 0, j\to\infty$，所以历史上的扰动的影响会逐渐消失。另外$\rho_k\to 0, k\to\infty$，所以相距较远的观测之间的相关性很小。

不是所有的弱平稳时间序列都有这样的性质。非平稳序列更是不需要满足这些性质。

2.5 附录：补充知识

Chapter 3

自回归模型

3.1 自回归模型的概念

如果$\rho_1 \neq 0$, 则X_t与X_{t-1}相关，可以用X_{t-1}预测X_t。最简单的预测为线性组合，如下模型：

$$X_t = \phi_0 + \phi_1 X_{t-1} + \varepsilon_t$$

称为一阶自回归模型（Autoregression model）, 记作 AR(1)模型。其中$\{\varepsilon_t\}$是零均值独立同分布白噪声序列，方差为σ
1。更一般的定义中仅要求$\{\varepsilon_t\}$是零均值白噪声，不要求独立同分布。

这个模型与一元线性回归模型$Y_i = \phi_0 + \phi_1 x_i + \varepsilon_i$在某些方面类似，比如，$\varepsilon_t$都是起到误差或者扰动的作用。但是，
在自回归模型中，自变量X_{t-1}在时刻$t-1$时作为因变量，所以自回归模型中因变量和自变量不是两个变量而是同一个

AR(1)模型也是马尔可夫（Markov）过程：X_t在$X_{t-1}, X_{t-2}, \ldots$条件下的条件分布，只与X_{t-1}有关。已知X_{t-1}后，用X

条件期望和条件方差：

$$E(X_t|X_{t-1}) = \phi_0 + \phi_1 X_{t-1}, \quad \text{Var}(X_t|X_{t-1}) = \sigma^2$$

即在$X_t - 1 = x_{t-1}$已知后，X_t条件的条件分布是期望为$\phi_0 + \phi_1 x_{t-1}$, 方差为σ^2的分布。可以证明：

$$\text{Var}(X_t) = \frac{\sigma^2}{1 - \phi_1^2}$$

因为$|\phi_1| < 1$, 所以X_t在$X_t - 1 = x_{t-1}$已知条件下的条件方差小于其无条件方差，也就是说用$X_t -$
1的信息去预测X_t, 可以使得X_t的波动减小，能够达到预测的效果。

AR(1)模型的推广是 AR(p)模型：

$$X_t = \phi_0 + \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + \varepsilon_t$$

其中$\{\varepsilon_t\}$是零均值独立同分布白噪声序列，方差为σ^2,且ε_t与$X_t-1, X_{t-2}, ...$独立。系数$\phi_1, ..., \phi_p$需要满

$$1 - \phi_1 z - \cdots - \phi_p z^p = 0$$

的所有复根z_*都满足$|z_*| > 1$,上述方程的左边的多项式称为 AR(p) 模型的特征多项式，特征多项式的所有
S. Tsay, 2013) 中就是以倒数为特征根。对 AR(1) 就是要求$|\phi_1| < 1$.

更一般的定义中仅要求$\{\varepsilon_t\}$是零均值白噪声，不要求独立同分布。

3.2 滞后算子

设$\{\xi_t, t \in \mathbb{Z}\}$为弱平稳时间序列或者常数列，定义如下的滞后算子$B$：

$$B\xi_t = \xi_{t-1}, \quad B^j\xi_t = \xi_{t-j},\ j \in \mathbb{Z}$$

考虑复变函数$P(z) = \sum_{j=0}^{\infty} a_j z^j$,其中$\{a_j\}$为实数列，设$\sum_{j=0}^{\infty} |a_j| < \infty$
(称为系数绝对可和),有限阶多项式为$P(z)$的特例。定义：

$$P(B)\xi_t = \sum_{j=0}^{\infty} a_j \xi_{j-t}.$$

$P(B)$也称为一个(常系数时齐)线性滤波器，$P(B)\xi_t$是$\{\xi_t\}$序列的一个滑动平均变换。

设$Q(z) = \sum_{j=0}^{\infty} b_j z^j, \{b_j\}$为实数列，绝对可和，则$C(z) = P(z)Q(z) = \sum_{j=0}^{\infty} c_j z^j$,其中：

$$c_j = \sum_{i=0}^{j} a_i b_{j-i},\ j = 0, 1, ...$$

且$\{c_j\}$绝对可和，称数列$\{c_j\}$是数列$\{a_j\}$和数列$\{b_j\}$的离散卷积。对任意弱平稳列或者常数列$\{\xi_t\}$均有：

$$P(B)Q(B)\xi_t = Q(B)P(B)\xi_t = C(B)\xi_t,\ t \in \mathbb{Z}$$

若$\xi_t \equiv \xi$与t无关，则$B^j\xi = \xi$。比如，$B^j 1 = 1, P(B)1 = P(1)$。

令$D(z) = \frac{P(z)}{Q(z)}$,若$Q(z) \neq 0$对任意满足$|z| \leq 1$的复数z均成立，则$D(z) = \sum_{j=0}^{\infty} d_j z^j, \{d_j\}$绝对可和，且：

$$P(B)\xi_t = Q(B)D(B)\xi_t = D(B)Q(B)\xi_t,\ t \in \mathbb{Z}$$

3.3 TODO: AR(1) 模型的性质

3.4 TODO: AR(1) 模型的自相关函数

3.5 TODO: AR(2) 模型的性质

3.5.0.1 TODO: 例子：美国的国民生产总值（GNP）经过季节调整后的季度增长率

3.6 AR(p) 模型的性质

对于一般的AR(p) 模型，其 ACF 的性质以及序列的随机周期，也由其特征根决定。ACF 可以是单调衰减、震荡衰减、正负交替衰减、呈周期震荡衰减。在有复特征根根或者有接近-1的特征根时时间序列呈现出一定的随机周期变化。

由平稳性得：

$$\mu = \frac{\phi_0}{1 - \phi_1 - \cdots - \phi_p}$$

自相关函数 (ACF) 满足如下的递推(差分方程)：

$$(1 - \phi_1 B - \cdots - \phi_p B^p)\rho_j = 0, \quad j = 1, 2, ...$$

AR(p)模型的平稳解是线性时间序列。

3.7 偏自相关函数

实际数据用 AR 模型建模时，阶数p是未知的，确定p的问题称为定阶。一般常用偏自相关函数和 AIC 准则。

设$X_1, ..., X_n, Y$为随机变量：

$$L(Y|X_1, \ldots, X_n) = \underset{\hat{Y}=b_0+b_1X_1+\cdots+b_nX_n}{\operatorname{argmin}} E(Y - \hat{Y})^2$$

称为用$X_1, ..., X_n$对Y的最优线性预测。$Y - L(Y|X_1, ..., X_n)$与$Z - L(Z|X_1, ..., X_n)$的相关系数称为Y和Z在扣除$X_1, ..., X_n$影响后的偏相关系数。

对平稳线性时间序列，对$n = 1, 2, ...,$有：

$$L(X_t|X_{t-1}, ..., X_{t-n}) = \phi_{n0} + \phi_{n1}X_{t-1} + ... + \phi_{nn}X_{t-n}$$

其中$\phi_{nj}, j = 0, 1, ..., n$与$t$无关。称$\phi_{nn}$为时间序列$\{X_t\}$的偏自相关系数，$\{\phi_{nn}\}$序列称为时间序列$\{X_t\}$的偏自相关

ϕ_{nn}实际是X_t与X_t-n在扣除$X_t-2,...,X_t-n+1$的影响后的偏相关系数。$\phi_1 1$就是ρ_1。

ϕ_{nn}用样本进行估计，得到的估计值$\hat{\phi}_{nn}, n = 1, 2, ...$称为样本偏自相关函数。

如果$\{X_t\}$服从如下 AR(p)模型：

$$X_t = \phi_0 + \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + \varepsilon_t,\ \phi_p \neq 0$$

这意味着用$X_{t-1}, X_{t-2}, ...$的线性组合预测X_t时，只需要用到$X_t-1, ..., X_{t-p}$, 增加$X_{t-p-1}, X_{t-p-2}, ...$不能
$0, k > p$。这种性质叫做 AR 模型的偏自相关函数截尾性。

AR(p) 序列的样本偏自相关函数$\hat{\phi}_{kk}$满足如下性质：

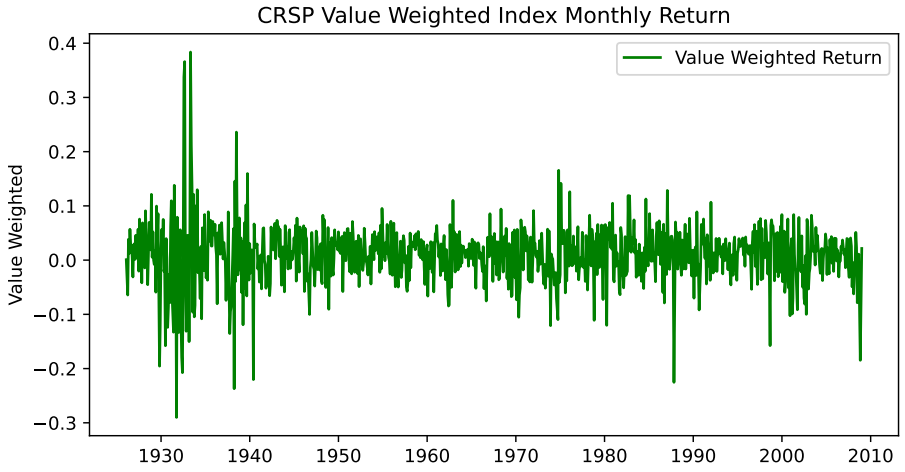
- $T \to \infty$ 时$\hat{\phi}_p p \to \phi_p \neq 0$。
- 对$k > p, \hat{\phi}_{kk} \to 0 (\hat{T} \to \infty)$。
- 对$k > p, \hat{\phi}_{kk}$渐近方差为$\frac{1}{T}$。

这样，可以用类似对 ACF 的白噪声检验那样给 PACF 图画出$\pm\frac{2}{\sqrt{T}}$的上下界限，以此判断 PACF 在哪里截尾。

3.7.0.1 例子：CRSP 价值加权指数月度收益率

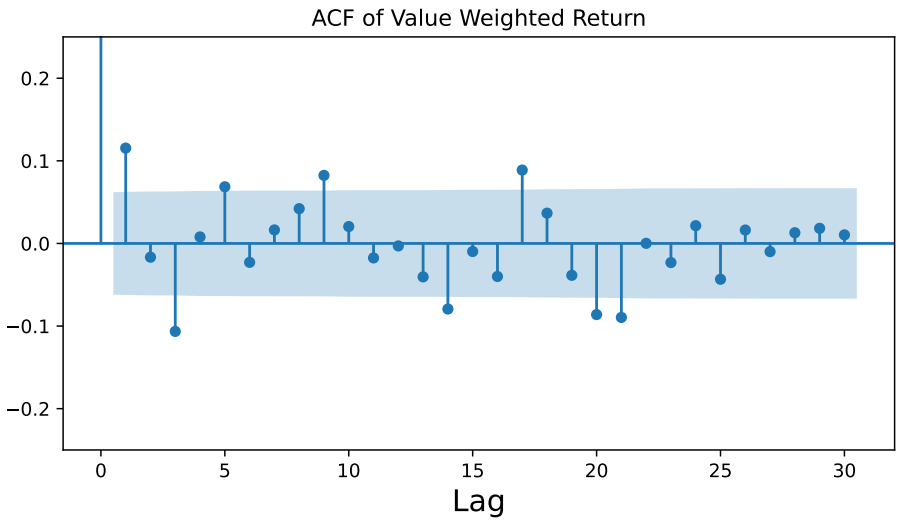
查看数据：

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np

raw_data = []
with open("../ftsdata/m-ibm3dx2608.txt", "r", encoding="utf-8") as file:
    for line in file.readlines():
        line = line.strip("\n").strip(" ").replace("\t", " ").split(" ")
        line = list(filter(lambda x: x != "", line))
        raw_data.append(line)
data = pd.DataFrame(raw_data[1:], columns=raw_data[0])

data["date"] = pd.to_datetime(data["date"], format="%Y%m%d")
data.set_index("date", inplace=True)
data = data.apply(pd.to_numeric)
data.head()

plt.figure(figsize=(8, 4))
plt.plot(data["vwrtn"], label='Value Weighted Return', color="green")
plt.title('CRSP Value Weighted Index Monthly Return')
plt.ylabel('Value Weighted')
plt.legend()
plt.show()
```



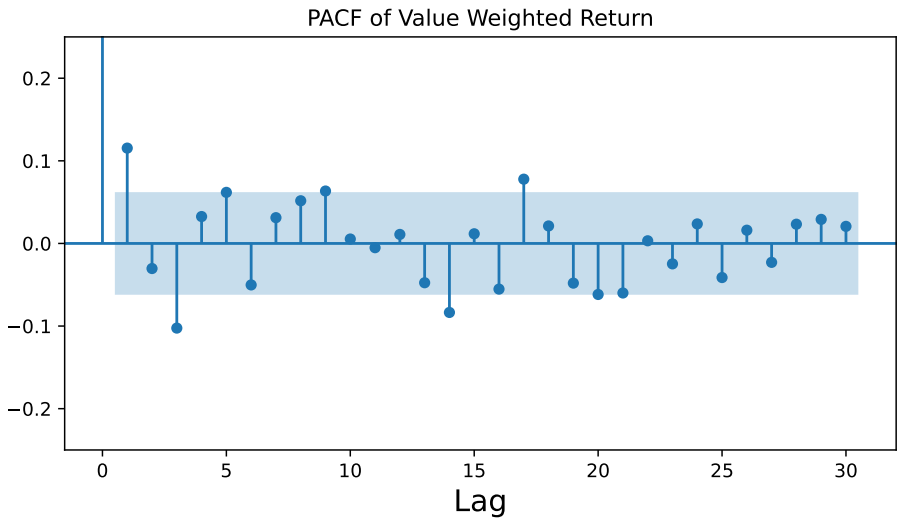
绘制 ACF 图像：

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np

raw_data = []
with open("../ftsdata/m-ibm3dx2608.txt", "r", encoding="utf-8") as file:
    for line in file.readlines():
        line = line.strip("\n").strip(" ").replace("\t", " ").split(" ")
        line = list(filter(lambda x: x != "", line))
        raw_data.append(line)
data = pd.DataFrame(raw_data[1:], columns=raw_data[0])

data["date"] = pd.to_datetime(data["date"], format="%Y%m%d")
data.set_index("date", inplace=True)
data = data.apply(pd.to_numeric)
data.head()

from statsmodels.graphics.tsaplots import plot_acf

plt.figure(figsize=(8, 4))
ax = plt.gca()  #
plot_acf(data["vwrtn"], ax=ax, lags=30)
ax.set_ylim(-0.25, 0.25)
ax.set_xlabel('Lag', fontsize=16)  #
plt.title('ACF of Value Weighted Return')
plt.show()
```



发现 ACF 到 $k = 21$ 还没有截尾，绘制 PACF 图：

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np

raw_data = []
with open("../ftsdata/m-ibm3dx2608.txt", "r", encoding="utf-8") as file:
    for line in file.readlines():
        line = line.strip("\n").strip(" ").replace("\t", " ").split(" ")
        line = list(filter(lambda x: x != "", line))
        raw_data.append(line)
data = pd.DataFrame(raw_data[1:], columns=raw_data[0])

data["date"] = pd.to_datetime(data["date"], format="%Y%m%d")
data.set_index("date", inplace=True)
data = data.apply(pd.to_numeric)
data.head()

from statsmodels.graphics.tsaplots import plot_pacf

plt.figure(figsize=(8, 4))
ax = plt.gca()  #
plot_pacf(data["vwrtn"], ax=ax, lags=30)
ax.set_ylim(-0.25, 0.25)
ax.set_xlabel('Lag', fontsize=16)  #
plt.title('PACF of Value Weighted Return')
plt.show()
```



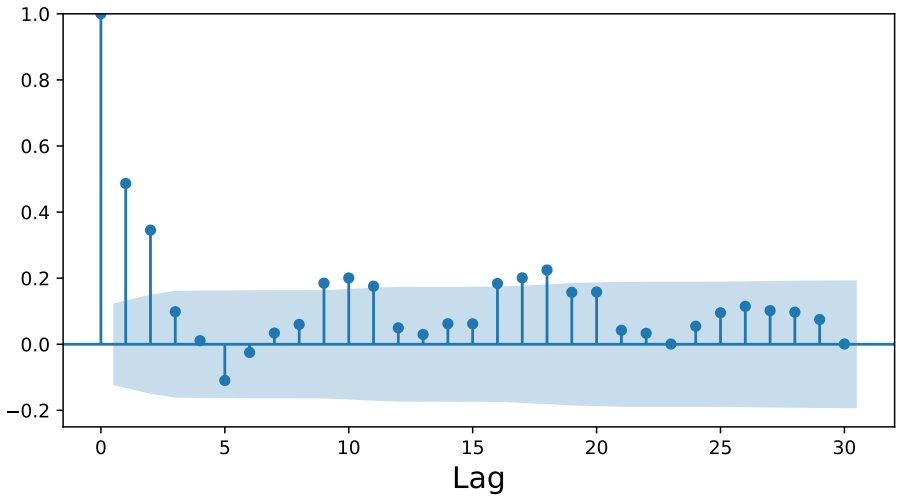
查看 PACF 图像，可以近似认为 $p = 3$，但实际上 PACF 在 $k = 17$ 处仍未截尾。

3.7.0.2 例子：美国的国民生产总值（GNP）经过季节调整后的季度增长率

这是 TODO: 例子：美国的国民生产总值（GNP）经过季节调整后的季度增长率的继续，判断美国 GNP 经过季节调整后的季度增长率的 p 的取值。

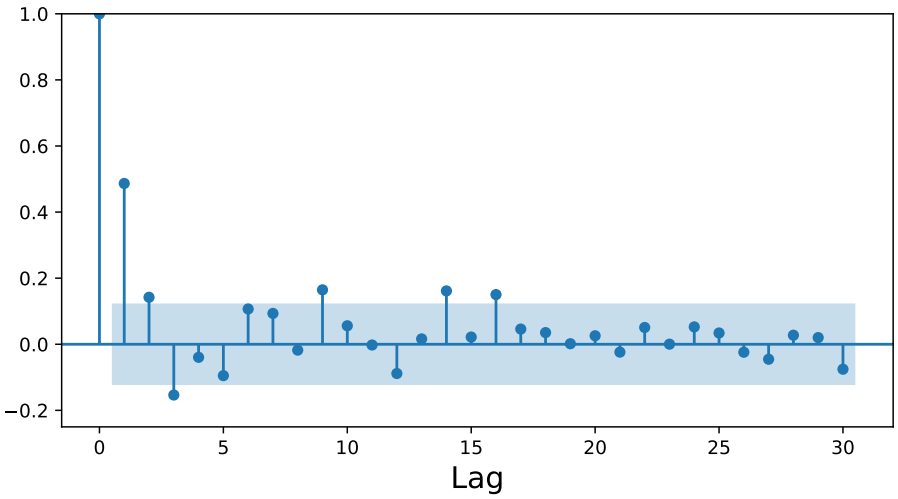
绘制 ACF 图：

```
import pandas as pd

#
da = pd.read_csv("../ftsdata/q-gnp4710.txt", sep='\s+', dtype=float)
data = pd.DataFrame(da["VALUE"].values, index=pd.date_range(start="1947-01", periods=len(da), fre
data.head()

from statsmodels.graphics.tsaplots import plot_acf

plt.figure(figsize=(8, 4))
ax = plt.gca()  #
plot_acf(np.diff(np.log(data["value"])), ax=ax, lags=30)
ax.set_ylim(-0.25, 1)
ax.set_xlabel('Lag', fontsize=16)  #
plt.title('')
plt.show()
```



ACF 图明显不截尾，绘制 PACF 图：

```
import pandas as pd

#
da = pd.read_csv("../ftsdata/q-gnp4710.txt", sep='\s+', dtype=float)
data = pd.DataFrame(da["VALUE"].values, index=pd.date_range(start="1947-01", periods=le
data.head()

from statsmodels.graphics.tsaplots import plot_pacf

plt.figure(figsize=(8, 4))
ax = plt.gca()  #
plot_pacf(np.diff(np.log(data["value"])), ax=ax, lags=30)
ax.set_ylim(-0.25, 1)
ax.set_xlabel('Lag', fontsize=16)  #
plt.title('')
plt.show()
```



PACF 虽然在 $k = 3, 9, 14, 16$ 等位置超出界限，但是超出不多，可考虑用 AR(3) 建模。

3.8 信息准则

信息准则是统计建模中常用的模型比较工具，其基本思想是模型拟合数据的拟合优度与模型简单化的折衷。AIC 准则(Akaike's Information Criterion)：

$$\text{AIC} = -\frac{2}{T}\ln(\text{似然函数值}) + \frac{2}{T}(\text{参数个数})$$

其中似然函数值是在参数最大似然估计处的似然函数值。当模型为高斯AR(p)，即$\{\epsilon_t\}$是独立同 N$(0, \sigma^2)$序列时的AR(p)模型时，AIC 公式为：

$$\text{AIC}(k) = \ln\tilde{\sigma}_k^2 + \frac{2k}{T}$$

其中k是模型的阶，$\tilde{\sigma}_k^2$是阶为k的条件下ε_t的方差的最大似然估计。$\ln\tilde{\sigma}_k^2$代表了模型对数据的拟合优劣，此值越大拟合
AIC(k)最小，就达成了拟合优度与模型简单程度的折衷。

另一个常用的信息准则是 BIC 准则 (Bayesian Information Criterion)，高斯 AR 模型为：

$$\text{BIC}(k) = \ln\tilde{\sigma}_k^2 + \frac{k\ln T}{T}$$

BIC 倾向于取比 AIC 更低阶的模型。

可以取 $k = 0, 1, \cdots, P_0$ 计算 AIC 或 BIC，去最小值点的 k。P_0 可以取为 $10 \log_{10} T$。

3.9 AR 模型的参数估计方法

AR 模型有多种估计方法，比如，用普通线性回归的最小二乘法估计，假设正态分布用最大似然估计，Yu Walken 递推计算，Burg 递推计算，等等。

设ϕ_i的估计为$\hat{\phi}_i$, 则拟合值为：

$$\hat{x}_t = \hat{\phi}_0 + \hat{\phi}_1 x_{t-1} + \cdots + \hat{\phi}_p x_{t-p},\ t = p+1, ..., T$$

残差为：

$$e_t = x_t - \hat{x}_t,\ t = p+1, ..., T$$

相应的新息方差σ^2 =Var(ε_t)的估计为：

$$\hat{\sigma}^2 = \frac{1}{T-2p-1} \sum_{t=p+1}^{T} e_t^2$$

如果使用高斯条件最大似然估计(认为$x_1, ..., x_p$固定), 则$\hat{\phi}_i$估计不变，但是新息方差得估计变成了：

$$\tilde{\sigma}^2 = \frac{1}{T-p} \sum_{t=p+1}^{T} e_t^2 = \frac{T-2p-1}{T-p} \sum_{t=p+1}^{T} e_t^2$$

3.10 TODO: AR 模型拟合优度指标

3.11 TODO: 用估计的 AR 模型进行预测

Chapter 4

移动平均模型

4.1 移动平均模型的概念

移动平均模型是具有q步外不相关性质的平稳列的模型；对于高阶的AR模型，有些可以用低阶的MA模型更好地描述。一般的AR模型也可以用高阶MA模型近似。

理论上，AR模型也可以是无穷阶的：

$$X_t = \phi_0 + \sum_{j=1}^{\infty} \phi_j X_{t-j} + \varepsilon_t.$$

其中$\{\phi_j\}$应绝对可和。一个特例为：

$$X_t = \phi_0 - \sum_{j=1}^{\infty} (-\theta_1)^j X_{t-j} + \varepsilon_t.$$

其中$0 < |\theta| < 1$。将模型写成：

$$X_t + \sum_{j=1}^{\infty} (-\theta_1)^j X_{t-j} = \phi_0 + \varepsilon_t.$$

以$t-1$代入，并乘以$-\theta_1$, 有：

$$\sum_{j=1}^{\infty} (-\theta_1)^j X_{t-j} = -\phi_0\theta_1 - \theta_1\varepsilon_{t-1}.$$

代入到上式中得：

$$X_t = \phi_0(1+\theta_1) + \varepsilon_t + \theta_1\varepsilon_{t-1}.$$

这样的模型称为MA(1)模型。

一般地，若$\{\varepsilon_t\}$是零均值独立同分布白噪声，方差为$\sigma^2, |\theta_1| < 1$, 令

$$X_t = \theta_0 + \varepsilon_t + \theta_1\varepsilon_{t-1},$$

易见$\{X_t\}$为线性时间序列形式的弱平稳列，称为MA(1)序列。

类似地，MA(2)序列的模型为：

$$X_t = \theta_0 + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2}.$$

MA(q)序列的模型为：

$$X_t = \theta_0 + \varepsilon_t + \theta_1\varepsilon_{t-1} + \cdots + \theta_q\varepsilon_{t-q}.$$

此模型也有特征多项式：

$$1 + \theta_1 z + \cdots + \theta_q z^q.$$

特征方程的根称为特征根，特征根都在单位圆外的条件称为MA模型的可逆条件。平稳性并不需要特征根的 条件。

上面从AR(∞) 导 出 MA(1) 的 过 程 ， 实 际 用 了 滞 后 算 子 的 一 些 运 算 法 则 ： 设 P(z) = Σ_{j= 0}^∞□_jz^j$和 $Q(z) = \sum_{j=0}^{\infty}\theta_j z^j$, $\sum_{j=0}^{\infty}|\phi_j| < \infty, \sum_{j=0}^{\infty}|\theta_j| < \infty$, 则：

$$P(z)Q(z) = R(z) = \sum_{j=0}^{\infty} r_j z^j,$$

其中：

$$r_j = \sum_{i=0}^{j}\phi_i\theta_{j-i},$$

且对弱平稳列$\{\xi_t\}$有：

$$P(B)Q(B)\xi_t = R(B)\xi_t.$$

4.2 移动平均模型的性质

以 MA(1) 和 MA(2) 为例讨论 MA 序列的性质，一般 MA(q) 序列类似讨论即可。

4.2.1 平稳性与自相关函数性质

以MA(1)为例。$X_t = \theta_0 + \varepsilon_t + \theta_1 \varepsilon_{t-1}$，其中$\{\varepsilon_t\}$是零均值独立同分布白噪声，$\theta_0, \theta_1$是任意实数，平稳性不需要特征根
易见：

$$EX_t = \theta_0, \ \forall t, \quad \mathsf{Var}(X_t) = \sigma^2(1+\theta_1^2) = \gamma_0.$$

而：

$$\gamma_1 = E[(X_t - \theta_0)(X_{t-1} - \theta_0)] = E[(\varepsilon_t + \theta_1\varepsilon_{t-1})(\varepsilon_{t-1} + \theta_1\varepsilon_{t-2})] = \theta_1 E\varepsilon_{t-1}^2 = \sigma^2\theta_1.$$

对$k > 1$有

$$\gamma_k = E[(\varepsilon_t + \theta_1\varepsilon_{t-1})(\varepsilon_{t-k} + \theta_1\varepsilon_{t-k-1})] = 0 \ (k > 1).$$

因为$k > 1$, 所以$t - k - 1 < t - k < t - 1 < t$, 求协方差时均不相关。
所以，对于MA(1)序列，有：

$$\gamma_k = \begin{cases} \sigma^2(1+\theta_1^2), & k = 0, \\ \sigma^2\theta_1, & k = 1, \\ 0, & k > 1. \end{cases}$$

相应地，MA(1)的自相关函数为：

$$\rho_k = \begin{cases} 1, & k = 0, \\ \frac{\theta_1}{1+\theta_1^2}, & k = 1, \\ 0, & k > 1. \end{cases}$$

这就验证了MA(1)序列是弱平稳列。 MA(1)的自相关函数在$k > 1$后为零的性质叫做MA序列的自相关函数截尾性。
对于MA(q)序列：

$$X_t = \theta_0 + \varepsilon_t + \theta_1\varepsilon_{t-1} + \cdots + \theta_q\varepsilon_{t-q},$$

易见：

$$EX_t = \theta_0, \quad \mathsf{Var}(X_t) = \sigma^2(1 + \theta_1^2 + \cdots + \theta_q^2) = \gamma_0.$$

其自相关函数ρ_k也满足q后截尾性，即$\rho_k = 0, \forall k > q$。如果$\theta_q \neq 0$, 则$\rho_q \neq 0$。这样，MA$(q)$序列的两个时间点的观测$X_s$和$X_t$ 当$|s-t| > q$时不相关，代表了一种特殊的“有限记忆”

4.2.2 可逆性

对MA(1)模型，当$|\theta_1| < 1$时，根据本章开始的推导可得：

$$\varepsilon_t = -\phi_0 + X_t + \sum_{j=1}^{\infty} (-\theta_1)^j X_{t-j}.$$

其中的级数是可以在a. s. 意义和均方意义下收敛的。这表明新息ε_t可以用当前的观测X_t以及历史观测X_t $1, 2, ...$的线性组合表示，而且历史观测X_{t-j}所在时刻离t时刻越远，其作用越小。这种性质叫做模型的可 $\mathbb{Z}$成立。

4.3 移动平均模型定阶

MA(q)序列的理论自相关函数ρ_k在q后截尾，$\rho_q \neq 0, \rho_k = 0, k > q$。

在$\{X_t\}$为独立同分布白噪声列的条件下，$k > 0$的$\hat{\rho}_k$渐近$\mathbb{N}(0, \frac{1}{T})$分布，所以查看ACF图，最后一个显著不

实际上，如$\{X_t\}$是MA(q)序列，则对$k > q, \sqrt{T}\hat{\rho}_k$渐近服从正态分布，渐近均值为零，渐近方差为：

$$1 + 2\rho_1^2 + \cdots + 2\rho_q^2.$$

也可以用AIC定阶：

$$\text{AIC}(k) = \ln \hat{\sigma}_k^2 + \frac{2k}{T}.$$

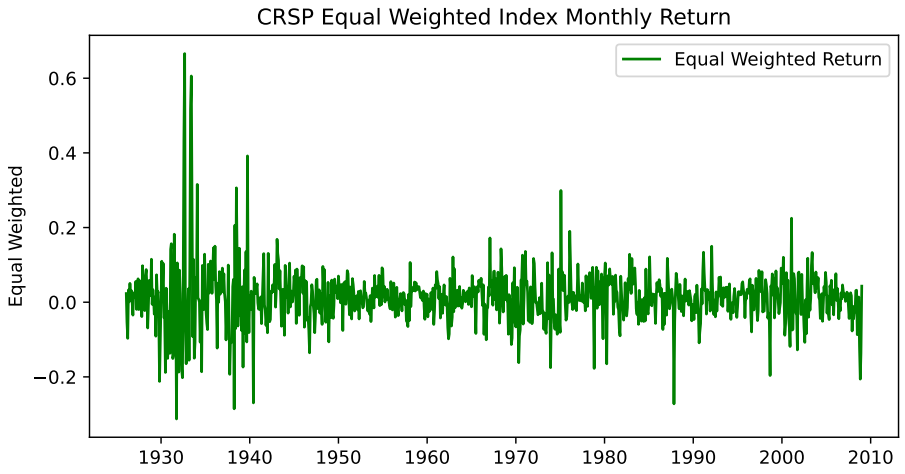
其中$\hat{\sigma}_k^2$是用MA(k)建模时新息方差的最大似然估计。

4.3.0.1 例子：CRSP 等权指数月度收益率

查看数据：

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np

raw_data = []
with open("../ftsdata/m-ibm3dx2608.txt", "r", encoding="utf-8") as file:
    for line in file.readlines():
        line = line.strip("\n").strip(" ").replace("\t", " ").split(" ")
        line = list(filter(lambda x: x != "", line))
```

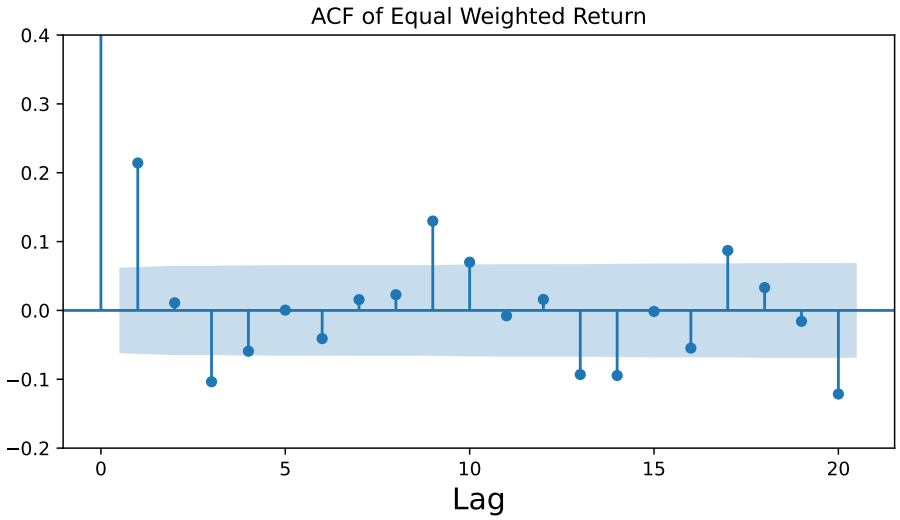
```
        raw_data.append(line)
data = pd.DataFrame(raw_data[1:], columns=raw_data[0])

data["date"] = pd.to_datetime(data["date"], format="%Y%m%d")
data.set_index("date", inplace=True)
data = data.apply(pd.to_numeric)
data.head()

plt.figure(figsize=(8, 4))
plt.plot(data["ewrtn"], label='Equal Weighted Return', color="green")
plt.title('CRSP Equal Weighted Index Monthly Return')
plt.ylabel('Equal Weighted')
plt.legend()
plt.show()
```



绘制 ACF 图像：

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np

raw_data = []
with open("../ftsdata/m-ibm3dx2608.txt", "r", encoding="utf-8") as file:
    for line in file.readlines():
        line = line.strip("\n").strip(" ").replace("\t", " ").split(" ")
        line = list(filter(lambda x: x != "", line))
        raw_data.append(line)
data = pd.DataFrame(raw_data[1:], columns=raw_data[0])

data["date"] = pd.to_datetime(data["date"], format="%Y%m%d")
```

```
data.set_index("date", inplace=True)
data = data.apply(pd.to_numeric)
data.head()

from statsmodels.graphics.tsaplots import plot_acf

plt.figure(figsize=(8, 4))
ax = plt.gca()  #
plot_acf(data["ewrtn"], ax=ax, lags=20)
ax.set_ylim(-0.2, 0.4)
ax.set_xlabel('Lag', fontsize=16)  #
plt.title('ACF of Equal Weighted Return')
plt.show()
```



可以看出，ACF 在 $k=1$ 时很大，$k=3, k=9$ 也较明显。可以考虑拟合 MA(3) 或 MA(9)。

4.4 移动平均模型的估计

MA模型参数的估计方法有：

- ☐ 矩估计法，利用$\{\gamma_k\}$与$\{\theta_k\}$、σ^2的关系求非线性方程组解；
- ☐ 逆相关函数法，将MA模型转换为长阶自回归模型，用估计自回归模型的方法估计，能保证可逆性
- ☐ 新息估计法；
- ☐ 条件最大似然估计法；
- ☐ 精确最大似然估计法。

条件最大似然估计法和完全最大似然估计法都假定$\{\varepsilon_t\}$为高斯白噪声，计算似然函数。在条件最大似然

$0, t \leq 0$, 这样就可以得到$\varepsilon_1 = x_1 - \theta_0, \varepsilon_2 = x_2 - \theta_0 - \theta_1\varepsilon_1$等递推表示，将其代入$\varepsilon_t, t = 1, 2, \ldots, T$的独立联合正态分布密度中就得到了条件似然函数，求其关于σ^2和$\theta_0, \theta_1, \ldots, \theta_q$的最大值点。

在精确最大似然估计中，将$\varepsilon_t, t = 1-q, 2-q, \ldots, -1, 0$也作为未知参数，与其它模型参数一起估计。条件最大似然估

4.5 移动平均模型的预测

因为MA(q)序列在间隔超过q步以后就独立，所以超前多步预测，只能预测到q步，从$q+1$步开始就只能用均值μ预测了。

以MA(1)为例，

$$X_t = \theta_0 + \varepsilon_t + \theta_1\varepsilon_{t-1}.$$

超前一步：

$$\hat{x}_h(1) = E(X_{h+1}|x_1, \ldots, x_h) = \theta_0 + \theta_1\varepsilon_h.$$

这里利用了$E(\varepsilon_{h+1}|x_1, \ldots, x_h) = 0$。$\varepsilon_h$ 是第h个新息，可以作为模型的残差计算，或者通过将MA模型表达为AR模型求

超前两步：

$$\hat{x}_h(2) = E(\theta_0 + \varepsilon_{h+2} + \theta_1\varepsilon_{h+1}|x_1, \ldots, x_h) = \theta_0$$

从两步开始的超前多步预报就变成$EX_t = \theta_0$了。

类似地，对于MA(2)序列，

$$\hat{x}_h(1) = \theta_0 + \theta_1\varepsilon_h + \theta_2\varepsilon_{h-1},\ \hat{x}_h(2) = \theta_0 + \theta_2\varepsilon_h.$$

对$k = 3, 4, \ldots$则有$\hat{x}_h(k) = \theta_0 = EX_t$.

4.6 AR 和 MA 的小结

- □ 对MA(q)模型，ACF对定阶有意义，因为其q后截尾
- □ 对AR(p)模型，PACF对定阶有意义，因为其p后截尾
- □ MA模型的序列不管系数如何总是平稳的，实际上还是因果线性时间序列，当特征根都在单位圆外时是可逆的
- □ AR模型只有当特征根都在单位圆外时才有ϵ_t 与$X_{t-1}, X_{t-2}, \ldots$独立的弱平稳解
- □ 对AR和MA序列，超前多步预测趋于序列的均值、预测均方误差趋于序列的方差