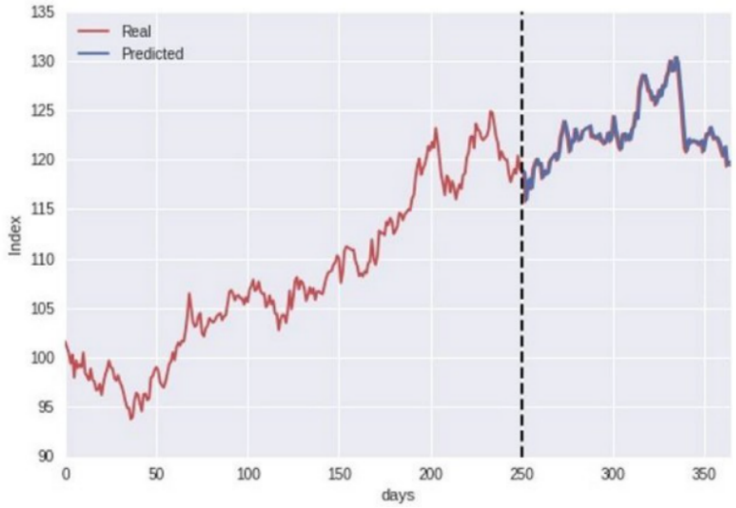
第一章 时间序列分析

最早的时间序列分析可以追溯到 7000 年前的古埃及。古埃及人把尼罗河涨落的情况逐天记录下来，就构成所谓的时间序列。对这个时间序列长期的观察使他们发现尼罗河的涨落非常有规律。由于掌握了尼罗河泛滥的规律，使得古埃及的农业迅速发展，从而创建了埃及灿烂的史前文明。

时间序列是日常生活中最常面对的数据，几乎无处不在。如大气中每一时刻的温度、湿度、大气压强；个人穿戴设备上采集的每分钟的心跳、体温、步行的步数；股票、期货价格、大宗商品交易价格走势；医学检查心电图、血压监测；Windows 的任务管理器中 CPU、内存、磁盘、网络等随着时间而波动的曲线等等，都是时间序列数据。



1.1 时间序列的概念

1. 时间序列：按照时间的顺序把随机事件变化发展的过程记录下来就构成了一个时间序列。

时间序列数据一般包含 3 个部分：时间点、主体（被测量对象）、测量值。在数据科学上，时序数据通常以数据表的形式（如 Excel 表格）记录，可用于查询、分析、预测。

2. 时间序列分析：对时间序列进行观察、研究，揭示其变化发展的规律，预测、控制其未来的走势。

3. 随机序列：按时间顺序排列的一组随机变量 $X_1, X_2, \ldots,$

4. 观察值序列：为了揭示随机序列的性质，需要通过观察值进行推断。观察值序列为随机序列的 n 个有序观察值 $x_1, x_2, \ldots, x_n$。

1.1.1 时间序列的特点

(1) 序列中的数据或数据点的位置依赖于时间，但不一定是时间的严格函数。

(2) 每一时刻上的取值或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。

(3) 前后时刻（不一定是相邻时刻）的数值或数据点的位置有一定的相关性，这种相关性即为系统的动态规律性。

(4) 从整体上看，时间序列往往呈现出某种趋势性或出现周期性变化的现象。

1.2 时间序列的分类及分析方法

1.2.1 时间序列的分类

时间序列的分类如下。

(1) 按研究对象分：一元时间序列和多元时间序列。

(2) 按时间参数分：离散时间序列和连续时间序列。

(3) 按统计特性分：平稳时间序列和非平稳时间序列宽平稳与严平稳。

(4) 按分布规律分：高斯型时间序列和非高斯型时间序列。

1.2.2 时间序列的分析方法

(1) 确定性变化分析：趋势变化分析、周期变化分析、循环变化分析。

(2) 随机性时间序列分析：包括频域和时域分析方法。

早期的频域分析方法借助傅里叶分析从频率的角度揭示时间序列的规律。后来借助了傅里叶变换，用正弦、余弦项之和来逼近某个函数。20 世纪 60 年代，引入极大熵谱估计理论，进入现代谱分析阶段。

时域分析模型的发展脉络如下。

1927 年，由 G.U. Yule 提出的 AR 模型；

1931 年，由 G.T. Walker 提出的 MA 模型、ARMA 模型；

1970 年，由 G.E.P. Box 和 G.M. Jenkins 提出的 ARIMA 模型（Box-Jenkins 模型）；

1980 年，由汤家豪（H. Tang）等提出的门限自回归（TAR）模型；

1982 年，由 Robert F. Engle 等提出的 ARCH 模型；

1985 年，由 Bollerslov 提出的 GARCH 模型；

1987 年，由 C. Granger 提出的协整（co-integration）理论。

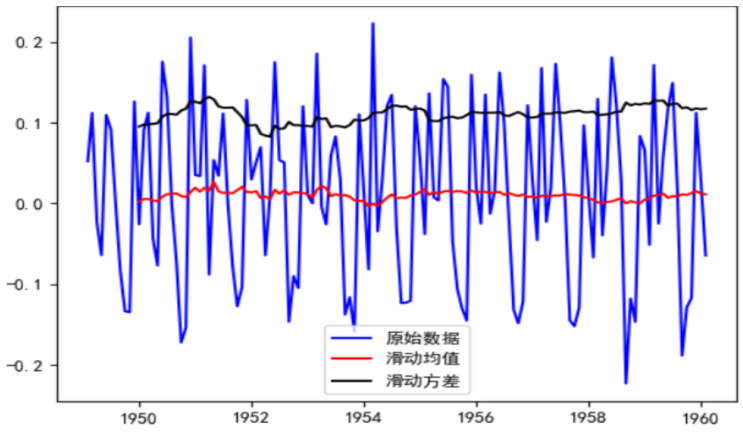
1.2.3 研究工具

SPSS、SAS、R、S-Plus、Eviews、Matlab 等。

1.3 常见的数据特征及数据预处理

常见的数据特征如下。

(1) 平稳序列：粗略地讲，统计特征不随时间的推移基本不发生改变，比如均值、方差在任意时间段基本不变。



(2) 非平稳序列：一般具有一个或多个下列特性：趋势性、跳跃、周期性等。

(3) 差分平稳序列：非平稳序列中，数据由较强的趋势性，经过差分变换后的序列符合平稳序列的特征。

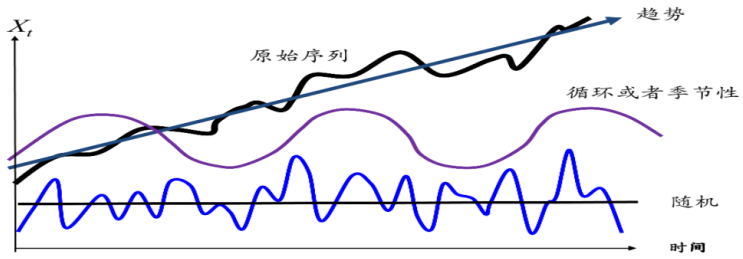
1.3.1 数据预处理

数据的预处理包括：奇异值检测、缺失值填补、数据转换、数据分解。

常见的数据转换如将数据正态化处理的 Box-Cox 变换。

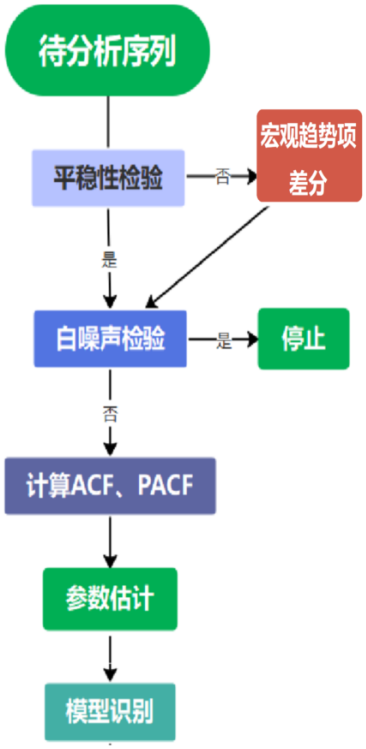
时间序列数据按照构成的特征分为：趋势性、季节性、波动性。

常用的数据分解有：X-13、ETS、STL 等。如 X-13ARIMA-SEATS 是美国人口普查局开发的官方时间序列分解工具，可将时间序列分解为趋势-循环成分（Trend-Cycle）、季节性成分（Seasonal）、不规则成分（Irregular）。



1.4 时间序列分析的流程

1. 宏观趋势项、差分（1 步或多步差分）
将非平稳序列转化为平稳序列。



2. 白噪声序列：纯随机、无规律、无预测价值的序列。
3. 自相关函数（ACF）：反应同一序列在不同时序的取值之间的相关性
4. 偏自相关函数（PACF）：在计算相关性时移除了中间变量的间接影响

1.5 时间序列的分解

常见的非平稳时间序列模型有以下加法分解模型：

$$X_t = u_t + C_t + R_t,$$

u_t：宏观趋势项，可用一些经验方法进行预测，比如回归、拟合、移动平均等；

C_t：周期项。

R_t 是随机变动项，$E(R_t) = 0$，$E(R_t^2) = \sigma^2$。

1.6 ARIMA 模型

ARIMA 模型（Autoregressive Integrated Moving Average model），也称差分整合移动平均自回归模型，适用于非平稳时间序列数据。

原理：将非平稳时间序列经过差分运算后，转换为平稳时间序列，然后进行回归所建立的模型。

ARIMA(p, d, q) 模型包含如下 3 个模块：

1. 差分模块：次数为 d，将非平稳序列转化为平稳序列。
2. 自回归模块（AR(p)），p 为自回归项数，寻找当前值与历史值之间的关系。
3. 移动平均模块（MA(q)）：q 为移动平均项数，分析模型中误差项的累加。

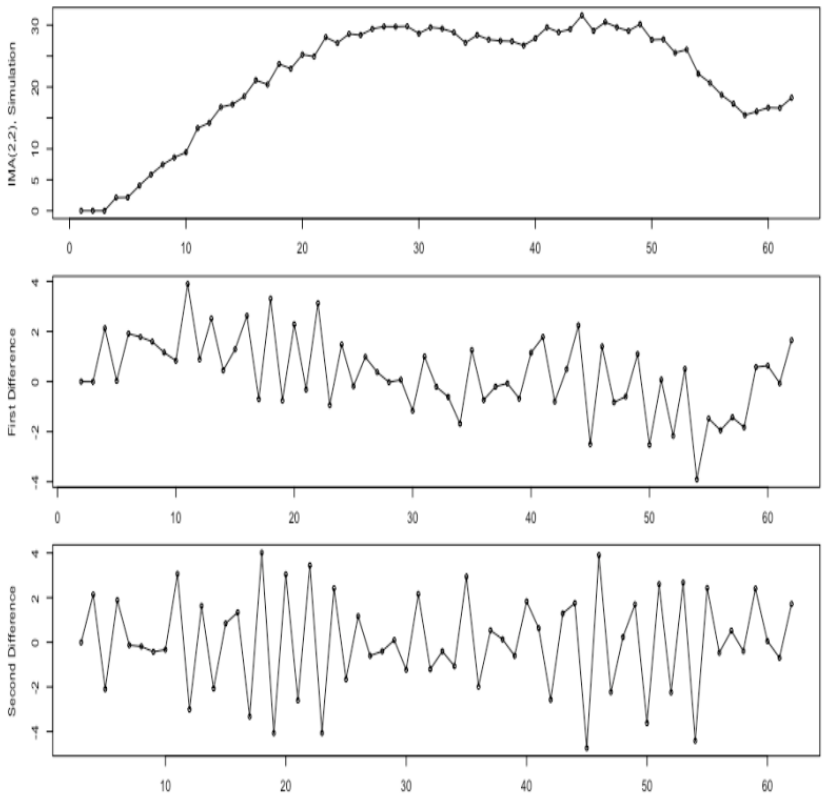
1. 差分模块

一阶差分：

$$\nabla R_t = R_t - R_{t-1}$$

二阶差分：

$$\nabla^2 R_t = \nabla R_t - \nabla R_{t-1}$$



1.6.1 Matlab 命令

newseries = diff(timeseries, n)

n：需要差分的阶数

2. 自回归模块

意义：寻找当前值与历史值之间的关系，p 为自回归项的阶数。

一阶自回归模型（AR(1)）：

$$y_t = \phi_0 + \phi_1 y_{t-1} + \varepsilon_t$$

其中，ϕ_0 和 ϕ_1 为回归系数，ε_t 为方差等于 σ^2 的白噪声。

二阶自回归模型（AR(2)）：

$$y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t.$$

高阶自回归模型（AR(p)）：

$$y_t = \phi_0 + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \varepsilon_t.$$

自回归模型定阶的方法有：偏自相关系数法、信息准则法、模型诊断法。

自回归模型预测：一步向前预测、多步向前预测。

3. 移动平均模块

意义：分析的是下面模型中的误差项的累加，每个预测值都可被认为是一个历史预测误差的加权移动平均值，q 为移动平均项数。

一阶移动平均模型（MA(1)）：

$$y_t = \mu + \varepsilon_t - \theta_1 \varepsilon_{t-1}$$

其中，μ 和 θ_1 为模型的参数，ε_t 为方差等于 σ^2 的白噪声。

二阶移动平均模型（MA(2)）：

$$y_t = \mu + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2}$$

高阶移动平均模型（MA(q)）：

$$y_t = \mu + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q}$$

移动平均模型定阶的方法有：自相关系数法、信息准则法、模型诊断法。

移动平均模型预测：一步向前预测、多步向前预测。

1.6.2 ARIMA 模型的数学表达

ARIMA(p,d,q) 的数学表达式如下：

$$y_t = \mu + \sum_{i=1}^{p} \phi_i y_{t-i} - \sum_{i=1}^{q} \theta_i y_{t-i} + \varepsilon_t$$

其中，μ, ϕ_i, θ_i 为模型的参数，ε_t 为方差等于 σ^2 的白噪声。

自回归移动平均模型定阶的方法有：拓展的自相关系数法、信息准则法、模型诊断法。

自回归移动平均模型预测：一步向前预测、多步向前预测。

1.7 线性时间序列建模

1976 年，Box 和 Jenkins 给出了如下的建模步骤。

Step 1. 对数据进行变换，或者差分处理，以确保数据（近似）满足平稳性假设。
Step 2. 对平稳的数据进行 ARIMA$(p,1,q)$ 建模，给出 p 和 q 的初值。
Step 3. 估计上述 ARIMA$(p,1,q)$ 模型中的参数。
Step 4. 对上述模型进行模型诊断，以判断模型设定是否和数据特征相匹配。
Step 5. 若模型不能通过诊断，则重新选取 p 和 q，并重复步骤 3 和 4，直至通过模型诊断。

说明：

(1) Step 1 中，常用的数据变换包括差分、对数差分等。带有季节性数据，可以通过 X-13 进行季节调整。具有确定性时间趋势的数据，可以将其对时间进行回归去掉趋势性。
(2) Step 2 中，常用的 p 和 q 的取值通常较小，如 0,1,2,3。
(3) Step 3 中，模型参数的估计通常采用极大似然函数估计法。
(4) Step 4 中，模型诊断包括利用 Ljung-Box 检验来检测估计误差是否为白噪声，以及估计误差的分布是否符合在构造极大似然估计时所设定的分布，是否具有异方差等。

对下面的 data 数据进行预测，主函数命令如下。

```
data=[10930,10318,10595,10972,7706,6756,9092,10551,9722,10913,11151,8186,6422 ...
    6337,11649,11652,10310,12043,7937,6476,9662,9570,9981,9331,9449,6773,6304,9355 ...
    10477,10148,10395,11261,8713,7299,10424,10795,11069,11602,11427,9095,7707,10767 ...
    12136,12812,12006,12528,10329,7818,11719,11683,12603,11495,13670,11337,10232 ...
    13261,13230,15535,16837,19598,14823,11622,19391,18177,19994,14723,15694,13248 ...
    9543,12872,13101,15053,12619,13749,10228,9725,14729,12518,14564,15085,14722 ...
    11999,9390,13481,14795,15845,15271,14686,11054,10395];
ddata = diff(data,1);

Y = ddata(:);
max_p = 3;          % AR的最大阶数
max_q = 3;          % MA的最大阶数
criterion = 'AIC'; % 使用AIC准则

[best_p, best_q] = select_arima_pq(Y, max_p, max_q, criterion);

Mdl = arima(best_p, 1, best_q);
EstMdl = estimate(Mdl,Y);
numPeriods = 10;
[YF, YMSE] = forecast(EstMdl, numPeriods, Y)
figure;
plot(Y);
hold on;
plot(length(Y)+1:length(Y)+numPeriods, YF,'r')
```

调用的函数命令如下。

```
function [best_p, best_q] = select_arima_pq(Y, max_p, max_q, criterion)
% 输入:
%   Y - 时间序列数据 (列向量)
%   max_p - p的最大值 (默认: 5)
%   max_q - q的最大值 (默认: 5)
%   criterion - 选择标准: 'AIC'(默认) 或 'BIC'
%
% 输出:
%   best_p - 最佳AR阶数
%   best_q - 最佳MA阶数

% 设置默认值
if nargin < 2 || isempty(max_p), max_p = 5; end
if nargin < 3 || isempty(max_q), max_q = 5; end
if nargin < 4 || isempty(criterion), criterion = 'AIC'; end
Y = Y(:);

% 计算不同(p,q)组合的AIC/BIC
aic_matrix = NaN(max_p+1, max_q+1);
bic_matrix = NaN(max_p+1, max_q+1);

for p = 0:max_p
```

```
    for q = 0:max_q
        try
            Mdl = arima(p, 1, q);      % 创建ARIMA(p,1,q)模型
            [~, ~, logL] = estimate(Mdl, Y, 'Display', 'off'); % 估计模型参数
            numParams = p + q + 1;     % 计算参数个数: AR(p) + MA(q) + 方差 = p + q + 1
            aic = -2*logL + 2*numParams; % 计算AIC
            bic = -2*logL + numParams*log(length(Y)-1); % 计算BIC
            aic_matrix(p+1, q+1) = aic; % 存储结果
            bic_matrix(p+1, q+1) = bic;
        catch ME
            aic_matrix(p+1, q+1) = Inf;
            bic_matrix(p+1, q+1) = Inf;
        end
    end
end

% 选择最佳模型
if strcmpi(criterion, 'BIC')
    criterion_matrix = bic_matrix;
    criterion_name = 'BIC';
else
    criterion_matrix = aic_matrix;
    criterion_name = 'AIC';
end
[~, min_idx] = min(criterion_matrix(:)); % 找到最小AIC/BIC值的位置
[best_p_idx, best_q_idx] = ind2sub(size(criterion_matrix), min_idx);
best_p = best_p_idx - 1;
best_q = best_q_idx - 1;
end
```

1.8 Matlab 命令

1. lbqtest 函数

lbqtest(timeseries)：判断待分析的时间序列 timeseries 是否是白噪声序列。

结果为 0 表示是白噪声序列，结果为 1 表示不是白噪声序列。

2. adftest 函数

adftest(timeseries)：利用 ADF 准则判断时间序列是否平稳。

结果为 0 表示时间序列不平稳，结果为 1 表示时间序列平稳。

3. AIC（赤池信息准则）：

$$AIC = -2LogL + 2k$$

用以衡量统计模型拟合优良性，该值越小越好。理论依据：基于信息论，目标是预测未来观测值的准确性。

4. BIC（贝叶斯信息量准则）：

$$BIC = -2LogL + kln(n)$$

防止模型复杂度过高而出现过拟合问题，该值越小越好。理论依据：基于贝叶斯统计，目标是找到真实模型的后验概率最大的候选模型。

5. Mdl=arima(p,d,q)

使用自回归度 p、差分度 d 和移动平均度 q 创建 ARIMA 模型

6. estimate 函数

[EstMdl, EstParamCov, LogL] = estimate(Mdl,newseries)

对于已知模型 Mdl 和数据样本 newseries 进行参数估计

返回值：估计模型 EstMdl、估计参数的协方差矩阵 EstParamCov、对数似然函数最优值 LogL

1.9 时间序列预测方法

1.9.1 均值预测法

采用历史数据的平均值作为对未来观测的预测。简单均值预测是指采用历史数据的等权重平均值预测：

$$\hat{y}_{T+h|T} = \frac{1}{T}\sum_{t=1}^{T} y_t$$

1.9.2 朴素预测法

采用当前的观测值作为对未来值的预测：

$$\hat{y}_{T+h|T} = y_T$$

1.9.3 滑动平均预测法

采用滑动平均窗口对窗口内的数据采用等权重预测：

$$\hat{y}_{T+h|T} = \frac{1}{k}\sum_{t=T-k+1}^{T} y_t$$

1.9.4 指数平滑法

结合了简单平均法、朴素预测法和滑动平均法，对所有历史观测值进行加权平滑：

$$\hat{y}_{T+1|T} = \alpha y_T + \alpha(1-\alpha)y_{T-1} + \alpha(1-\alpha)^2 y_{T-2} + \cdots$$

1.9.5 模型预测法

基于模型构造预测的方法。

1.10 模型的应用

参见优秀论文 2023-C-02。