

10

Time Series

时间数据

具有时间戳的数据序列



我们能看到有限长的未来，但是面对无限多的问题。

We can only see a short distance ahead, but we can see plenty there that needs to be done.

—— 艾伦·图灵 (Alan Turing) | 英国计算机科学家、数学家，人工智能之父 | 1912 ~ 1954



- ◀ `statsmodels.api.tsa.seasonal_decompose()` 季节性调整
- ◀ `numpy.random.uniform()` 生成满足均匀分布的随机数
- ◀ `df.ffill()` 向前填充缺失值
- ◀ `df.bfill()` 向后填充缺失值
- ◀ `df.interpolate()` 插值法填充缺失值
- ◀ `seaborn.boxplot()` 绘制箱型图
- ◀ `seaborn.lineplot()` 绘制线图

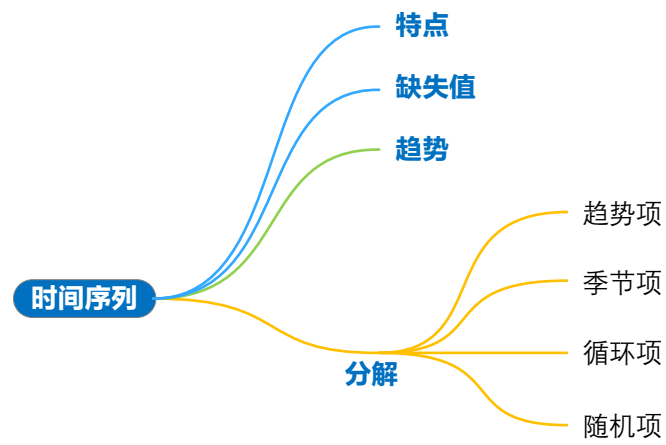
本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com



10.1 时间序列数据

时间序列是一种特殊的数据类型，它是某一特征在不同时间点上顺序观察值得到的序列。时间戳(timestamp)可以精确到年份、月份、日期，甚至是小时、分、秒。

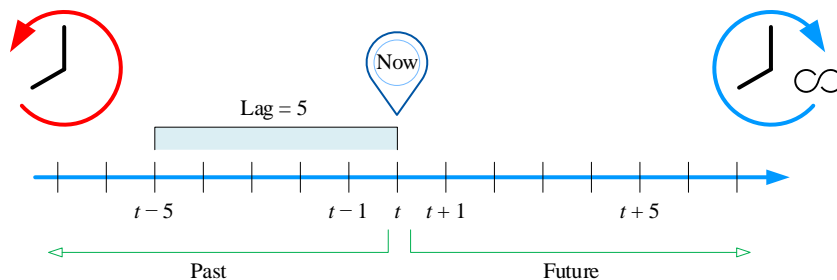


图 1. 时间轴

图 2 所示为 2020 年度中 9 支股票的每个营业日股价数据。图 2 中数据共有 253 行，每行代表一个日期及当日股价水平；共有 10 列，第 1 列为时间戳，其余 9 列每列为股价数据。除去时间戳一列和表头，图 2 可以看成是一个矩阵。

| Date | TSLA | TSM | COST | NVDA | FB | AMZN | AAPL | NFLX | GOOGL |
|-------------|--------|--------|--------|--------|--------|---------|--------|--------|---------|
| 2-Jan-2020 | 86.05 | 58.26 | 281.10 | 239.51 | 209.78 | 1898.01 | 74.33 | 329.81 | 1368.68 |
| 3-Jan-2020 | 88.60 | 56.34 | 281.33 | 235.68 | 208.67 | 1874.97 | 73.61 | 325.90 | 1361.52 |
| 6-Jan-2020 | 90.31 | 55.69 | 281.41 | 236.67 | 212.60 | 1902.88 | 74.20 | 335.83 | 1397.81 |
| 7-Jan-2020 | 93.81 | 56.60 | 280.97 | 239.53 | 213.06 | 1906.86 | 73.85 | 330.75 | 1395.11 |
| 8-Jan-2020 | 98.43 | 57.01 | 284.19 | 239.98 | 215.22 | 1891.97 | 75.04 | 339.26 | 1405.04 |
| 9-Jan-2020 | 96.27 | 57.48 | 288.75 | 242.62 | 218.30 | 1901.05 | 76.63 | 335.66 | 1419.79 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 21-Dec-2020 | 649.86 | 104.44 | 364.25 | 533.29 | 272.79 | 3206.18 | 128.04 | 528.91 | 1734.56 |
| 22-Dec-2020 | 640.34 | 103.55 | 361.32 | 531.13 | 267.09 | 3206.52 | 131.68 | 527.33 | 1720.22 |
| 23-Dec-2020 | 645.98 | 103.37 | 361.18 | 520.37 | 268.11 | 3185.27 | 130.76 | 514.48 | 1728.23 |
| 24-Dec-2020 | 661.77 | 105.57 | 363.86 | 519.75 | 267.40 | 3172.69 | 131.77 | 513.97 | 1734.16 |
| 28-Dec-2020 | 663.69 | 105.75 | 370.33 | 516.00 | 277.00 | 3283.96 | 136.49 | 519.12 | 1773.96 |
| 29-Dec-2020 | 665.99 | 105.16 | 371.99 | 517.73 | 276.78 | 3322.00 | 134.67 | 530.87 | 1757.76 |
| 30-Dec-2020 | 694.78 | 108.49 | 373.71 | 525.83 | 271.87 | 3285.85 | 133.52 | 524.59 | 1736.25 |
| 31-Dec-2020 | 705.67 | 108.63 | 376.04 | 522.20 | 273.16 | 3256.93 | 132.49 | 540.73 | 1752.64 |

图 2. 股票收盘股价数据

图 3 利用线图可视化股票收盘股价走势。图 3 (b) 右图初始股价归一化处理。

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

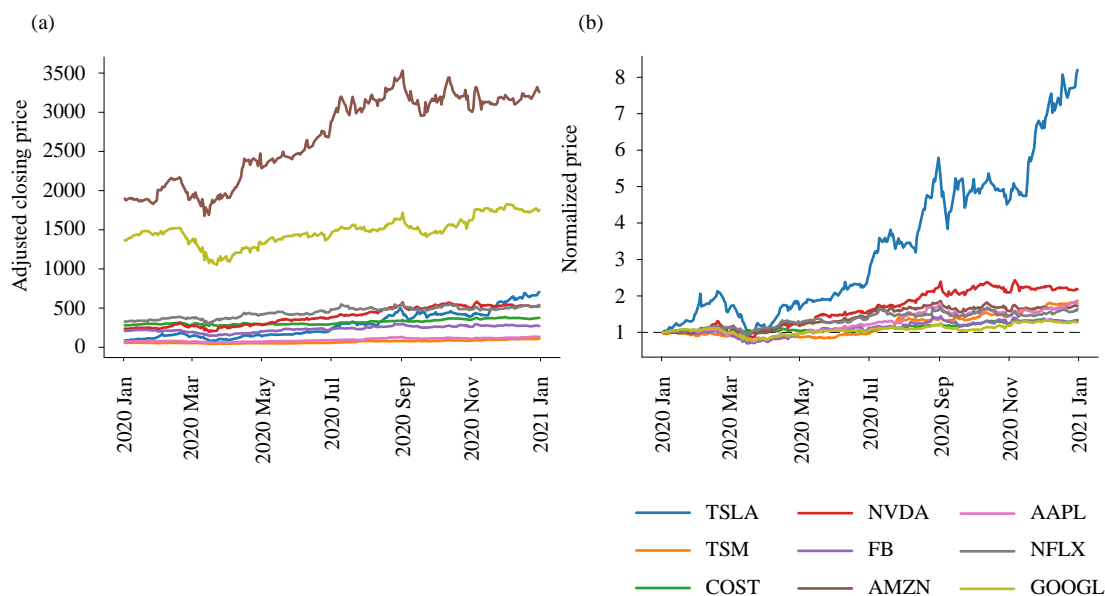


图 3. 股票收盘股价走势，和初始值归一化，时间序列数据

本书后续会用到收益率 (return) 这个概念。我们先介绍损益 (Profit and Loss, PnL) 这个概念。如图 4 所示，只考虑收盘价 S 在 t 时刻和 $t-1$ 时刻 (工作日) 的变动，通过如下公式计算出 t 时刻的日损益：

$$\text{PnL}_t = S_t - S_{t-1} \quad (1)$$

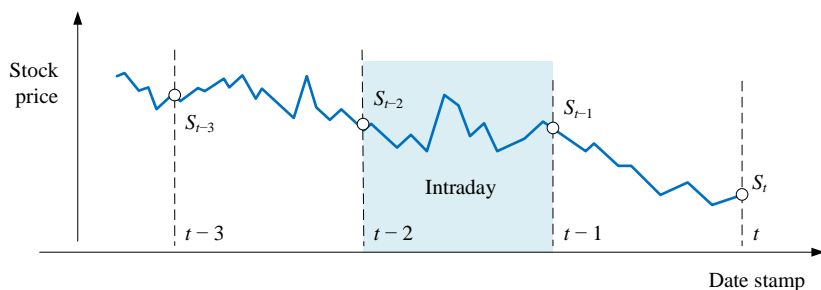


图 4. 某股票的价格变动

在不考虑分红 (dividend) 的条件下，单日简单回报率 (daily simple return) 可以这样计算：

$$r_t = \frac{S_t - S_{t-1}}{S_{t-1}} \quad (2)$$

金融分析还经常使用日对数回报率 (daily log return)：

$$r_t = \ln\left(\frac{S_t}{S_{t-1}}\right) \quad (3)$$

本书后续经常使用日对数收益率。

图 5 所示为一只股票在不同年份的日收益率分布，利用高斯分布估计样本分布多数情况下似乎是个不错的选择。图 6 所示为利用 KDE 估算得到概率密度。大家可以发现数据的统计量 (均值、方差、均方差、偏度、峰度) 随着时间变化。

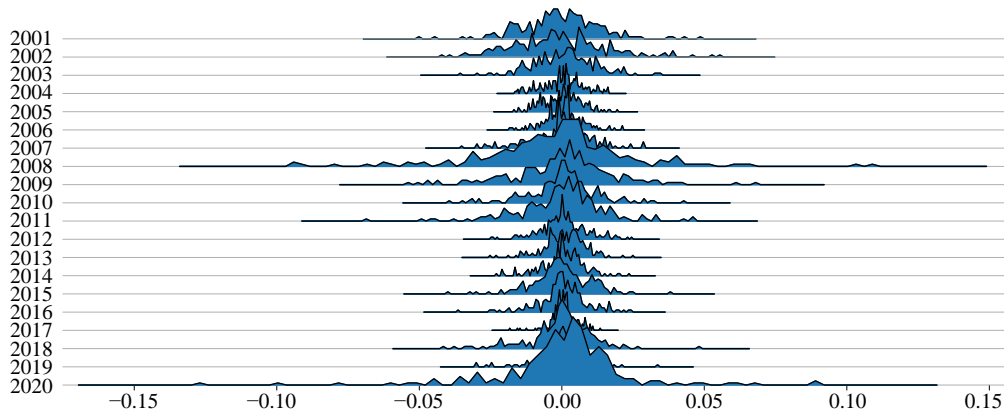


图 5. 收益率数据山脊图，按年分类

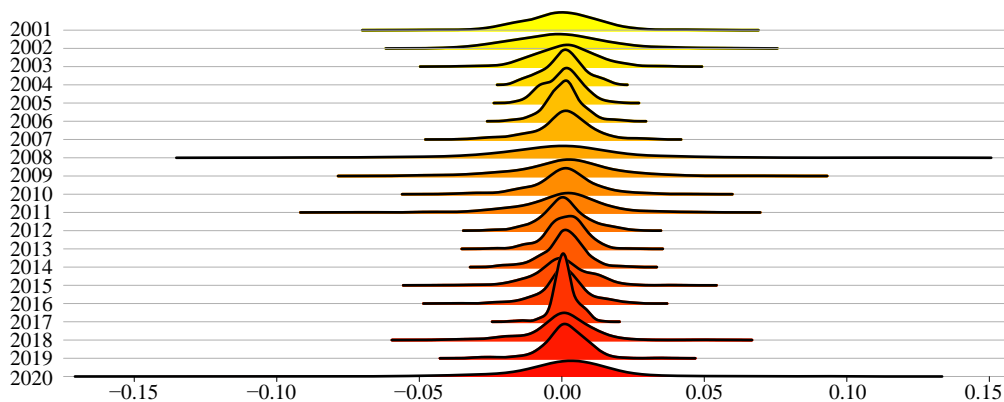


图 6. 收益率数据 KDE 山脊图，按年分类

鸢尾花数据，我们可以打乱数据的先后排列。但是时间序列是一个顺序序列，数据的先后顺序一般情况是不允许打乱的。有些情况，我们可以不考虑数据点的时间，比如图 7 所示回归分析。

本书第 10、11 章将介绍线性回归模型。

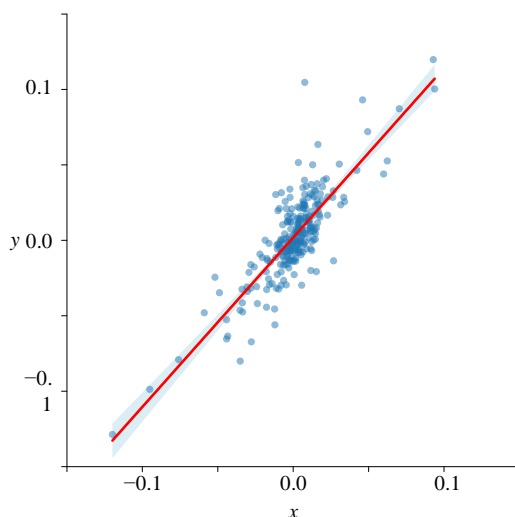


图 7. 线性 OLS 回归分析和散点图

10.2 处理时间序列缺失值

时间数据序列在分析建模之前，也需要注意数据中的缺失值和异常值处理。本书前文有专门章节介绍如何处理缺失值和异常值。本节从时间序列角度加以补充缺失值处理。

前文强调，时间序列数据是顺序观察的数据；因此在处理缺失值时，有其特殊性。比如，时间序列数据可以采用均值、众数、中位数、插值等一般方法，也可以采用如向前、向后这种方法。

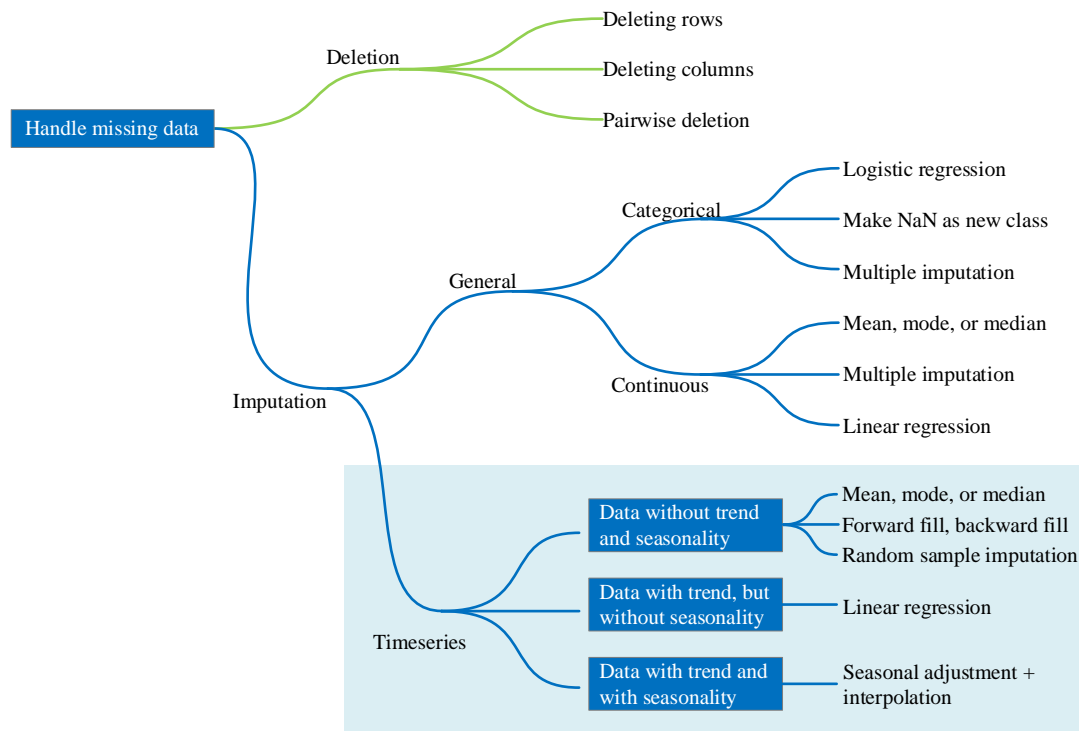


图 8. 处理缺失值

图 9 ~ 图 11 所示为三种不同处理时间序列缺失值的基本方法。

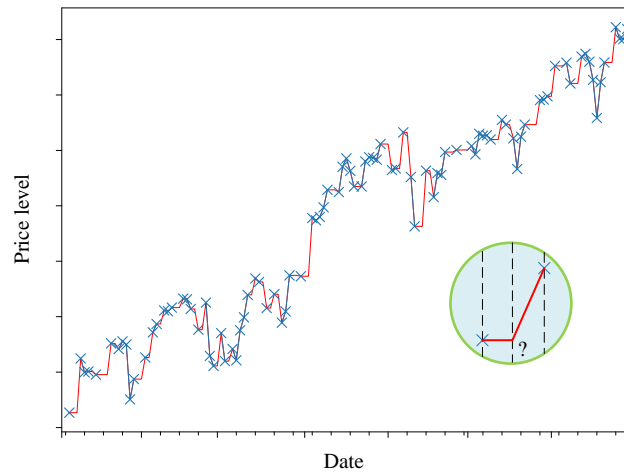


图 9. 向前插值填充缺失值

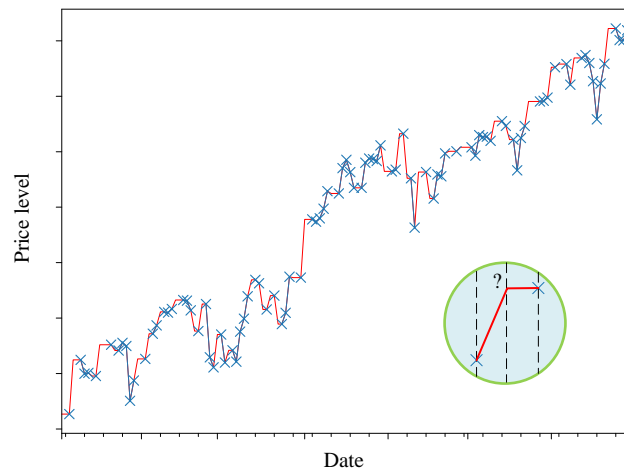


图 10. 向后插值填充缺失值

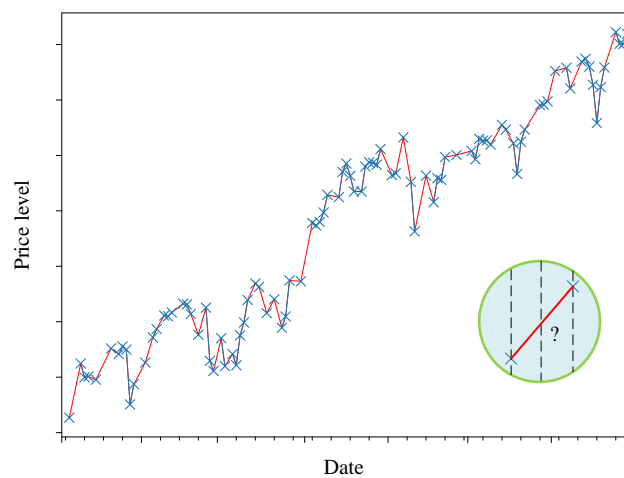


图 11. 线性插值填充缺失值

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com



Bk6_Ch10_01.ipynb 绘制图 9 ~ 图 11。

10.3 从时间数据中发现趋势

本节利用美国失业率数据介绍如何从时间数据中发现趋势。图 12 所示为失业率的原始数据。数据从 1950 年开始到 2021 年，每月有一个数据点。

观察图 12 这幅图，虽然存在“噪音”，我们已经能够大致看到失业率的按照年份的大致走势。下一章会介绍移动平均的方法来消除“噪音”。

观察图 12 的局部图中，我们还发现不同年份中一年内失业率存在某种特定的“模式”。也就是说，图中的“噪音”可能存在重要的价值！

图 13 所示为按月同比规律。与历史同时期比较，例如 2005 年 7 月份与 2004 年 7 月份相比称其为同比。相比图 12，图 13 更容易发现失业率变化规律。

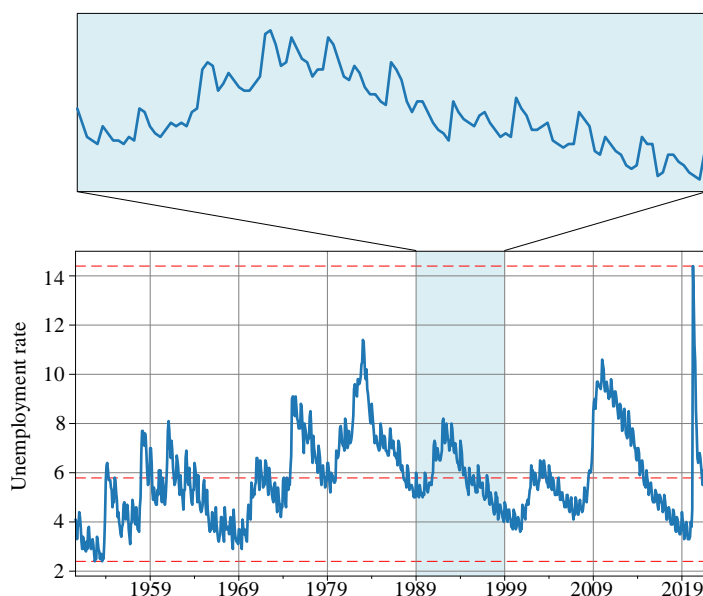


图 12. 原始失业率数据和局部放大图

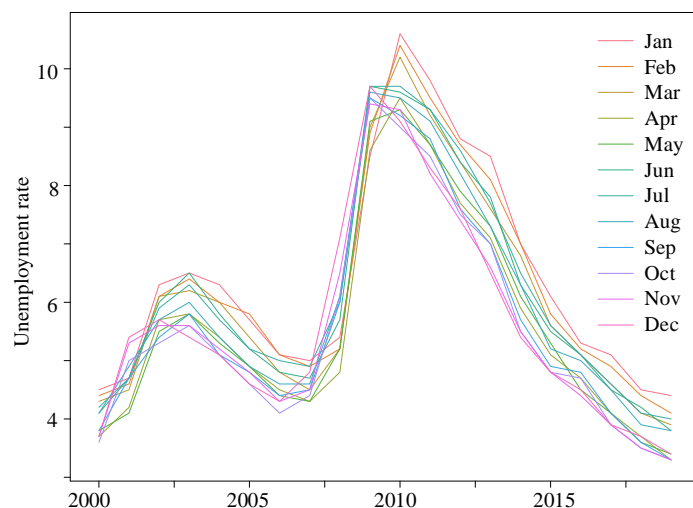


图 13. 失业率，按月同比

图 14 所示为年内环比数据。与上一统计段比较，例如 2005 年 7 月份与 2005 年 6 月份相比较称其为环比。我们似乎发现失业率存在某种年度周期规律。一年之内春天的失业率往往较低，这似乎和春天农业生产用工有关。而每一年的一月份的失业率显著提高，这可能和圣诞节、新年节庆之后用工下降有关。

为了进一步看到失业率随年度变化，我们可以用箱型图对年内失业率数据加以归纳，如图 15 所示。箱型图的均值代表年度失业率的平均水平。箱型图的四分位间距 IQR 告诉我们年度失业率的变化幅度。显然，失业率在 2020 年出现“前所未闻”的大起大落。

图 16 所示为月份失业率箱型图。比较月份失业率的平均值变化，一月份的平均失业率确实陡然升高，这也印证了之前的猜测。下一节，我们就介绍如何将不同的成分从原始时间数据中分离出来。

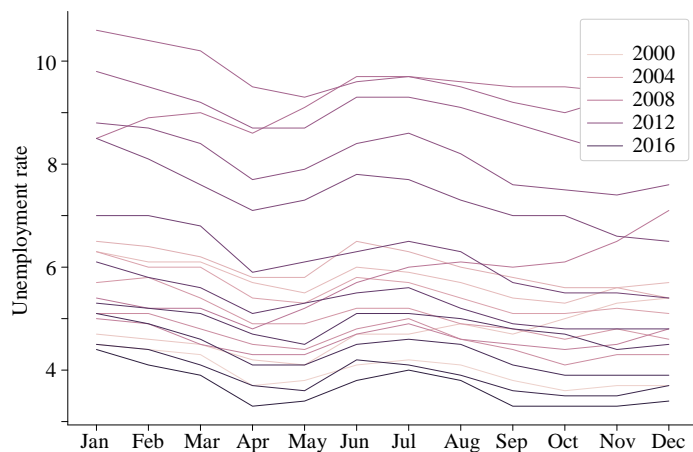


图 14. 失业率，年内环比

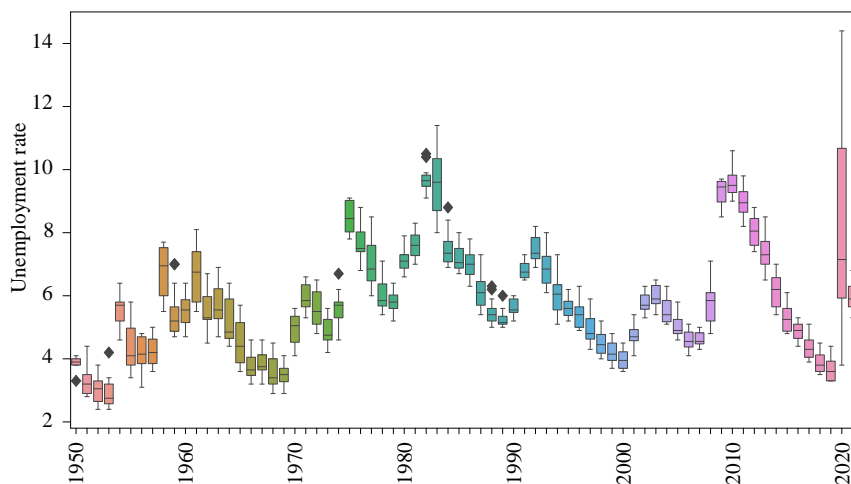


图 15. 年度失业率数据箱型图

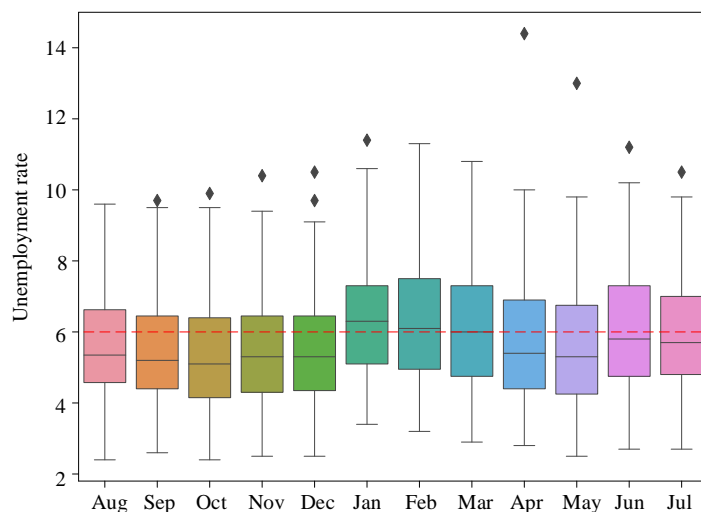


图 16. 月份失业率数据箱型图



Bk6_Ch10_02.ipynb 绘制本节图像。

10.4 时间序列分解

时间序列有如图 17 所示的几种主要的组成部分。具体定义如下：

- ▶ **趋势项** (trend component) $T(t)$ ，表征时间序列中确定性的非季节性长期总体趋势，通常呈现出线性或非线性的持续上升或者持续下降。当一个时间序列数据长期增长或者长期下降时，表示该序列有趋势。在某些场合，趋势代表着“转换方向”。例如从增长的趋势转换为下降趋势。
- ▶ **季节项** (seasonal component) $S(t)$ ，表征时间序列中确定性的周期季节性成分，是在连续时间内（例如连续几年内）在相同时间段（例如月或季度）重复性的系统变化。当时间序列中的数据受到季节性因

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

素（例如一年的时间或者一周的时间）的影响时，表示该序列具有季节性。季节性总是一个已知并且固定的频率。

- ◀ **循环项** (long-run cycle component) $C(t)$ 。循环项代表是相对周期更长（例如几年或者十几年）的重复性变化，但一般没有固定的平均周期，往往与大型经济体的经济周期息息相关。有时由于时间跨度较短，循环项很难体现出来，这时可能就被当作趋势项来分析了。当时间序列数据存在不固定频率的上升和下降时，表示该序列有周期性。这些波动经常由经济活动引起，并且与“商业周期”有关。周期波动通常至少持续两年。
- ◀ **随机项** (stochastic component) $I(t)$ ，表征时间序列中随机的不规则成分，体现出一定的自相关性以及持续时间内无法预测的周期。该成分可以是噪声，但不一定是。往往认为随机项包含有与业务自身密切相关的信息。

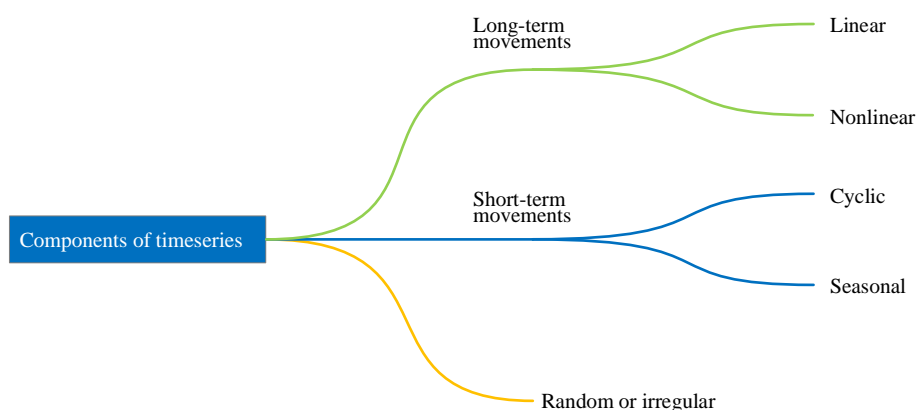


图 17. 时间序列成分

许多时间序列同时包含趋势、季节性以及周期性。基于以上的主要成分，一个时间序列可以有以下几种组合模型。

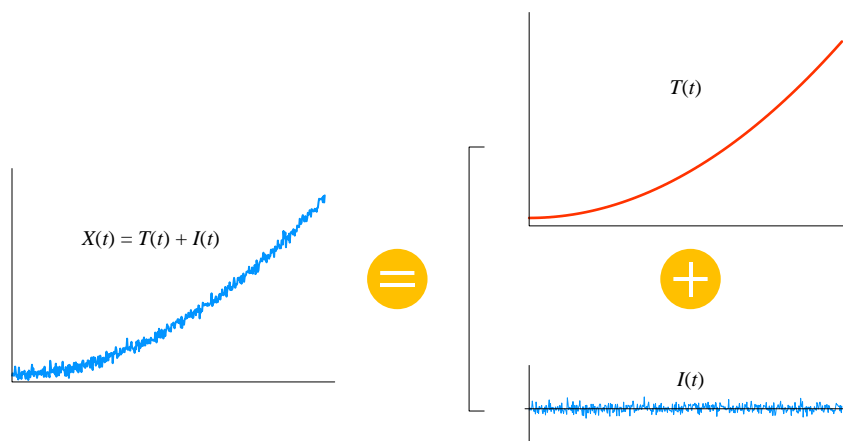
加法模型

加法模型 (additive model)，各个成分直接相加得到：

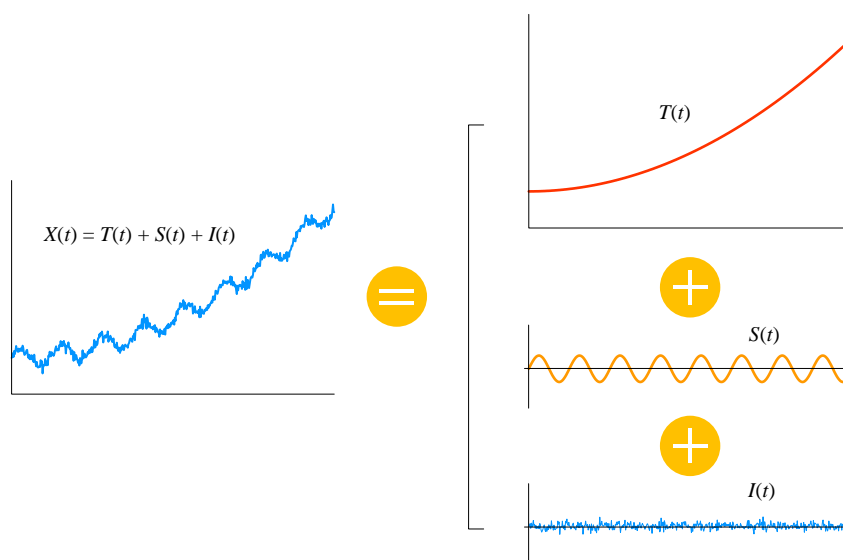
$$X(t) = T(t) + S(t) + C(t) + I(t) \quad (4)$$

这可能是最常用的时间序列分解方式。如果一个时间序列仅仅由趋势项 $T(t)$ 和随机项 $I(t)$ 构成：

$$X(t) = T(t) + I(t) \quad (5)$$

图 18. 累加分解，原始数据 $X(t)$ 被分解为趋势成分 $T(t)$ 和噪音成分 $I(t)$

标普 500 指数长期来看随时间增长，按照经济周期涨跌，短期来看指数每天波动不止。长期趋势成分 (trend component) $T(t)$ 就可以描述这种时间序列的长期行为，而不规则成分 (irregular component) $I(t)$ 描述的就是噪音成分，或者说是随机运动成分。

图 19. 累加分解，原始数据 $X(t)$ 被分解为趋势成分 $T(t)$ 、季节成分 $S(t)$ 和噪音成分 $I(t)$

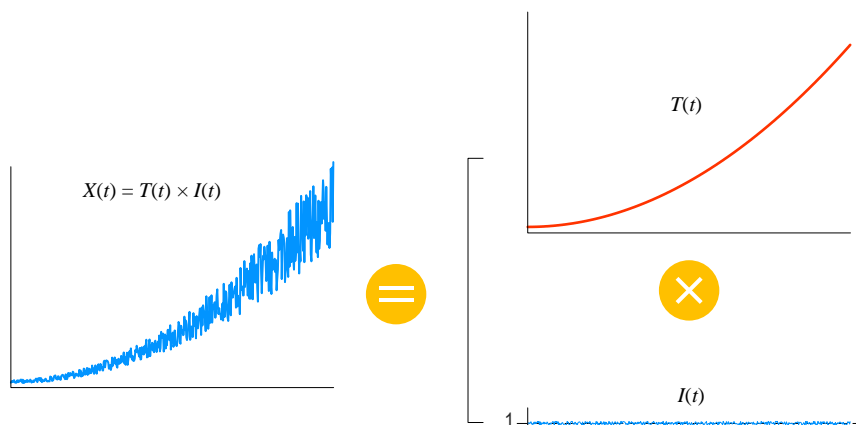
乘法模型

乘法模型 (multiplicative model)，各个成分直接相乘得到：

$$X(t) = T(t) \cdot S(t) \cdot C(t) \cdot I(t) \quad (6)$$

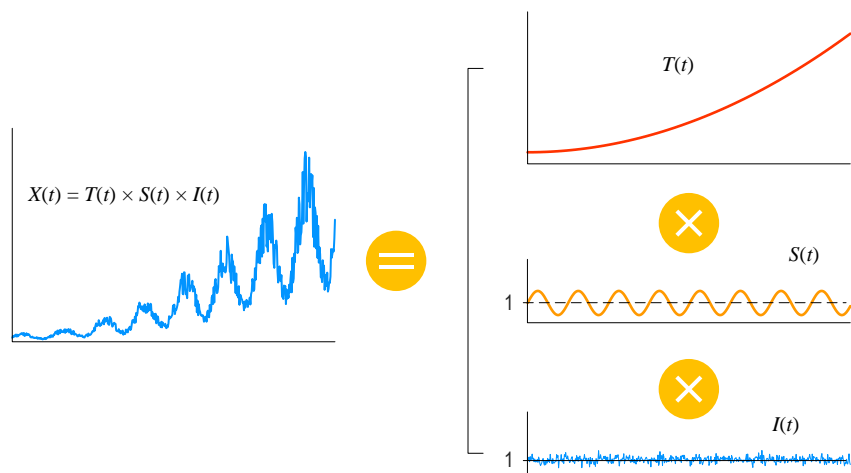
如果只考虑趋势项 $T(t)$ 和随机项 $I(t)$ ：

$$X(t) = T(t) \cdot I(t) \quad (7)$$

图 20. 累乘分解，原始数据 $X(t)$ 被分解为趋势成分 $T(t)$ 和噪音成分 $I(t)$

考虑季节成分的乘法模型：

$$X(t) = T(t) \cdot S(t) \cdot I(t) \quad (8)$$

图 21. 累乘分解，原始数据 $X(t)$ 被分解为趋势成分 $T(t)$ 和噪音成分 $I(t)$

当然，时间序列还可以存在其他分解模型。比如对数加法模型 (log-additive model)，时间序列取对数后由各个成分相加得到：

$$\ln X(t) = T(t) + S(t) + C(t) + I(t) \quad (9)$$

相当于对 $X(t)$ 进行对数转换。对于更复杂的时间序列分解模型，本书不做介绍。

10.5 季节调整

本节利用 `scipy.stats.tsa.seasonal_decompose()` 函数完成本章前文失业率数据的季节性调整。这个函数同时支持加法模型，`seasonal_decompose(series, model='additive')`，和乘法模型，`seasonal_decompose(series, model='multiplicative')`。本节采用的是默认的增加模型。

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

图 22 所示为失业率数据的分解。图 22 (a) 为原始数据，图 22 (b) 为趋势成分，图 22 (c) 为季节成分，图 22 (d) 为噪音成分。注意，图 22 四副子图的纵轴尺度完全不同。图 23、图 24、图 25 三幅图分别展示这四种成分。

`scipy.stats.tsa.seasonal_decompose()` 函数采用比较简单卷积方法进行季节调整，对于更复杂的季节性调整，建议大家了解 X11，本书不做展开。

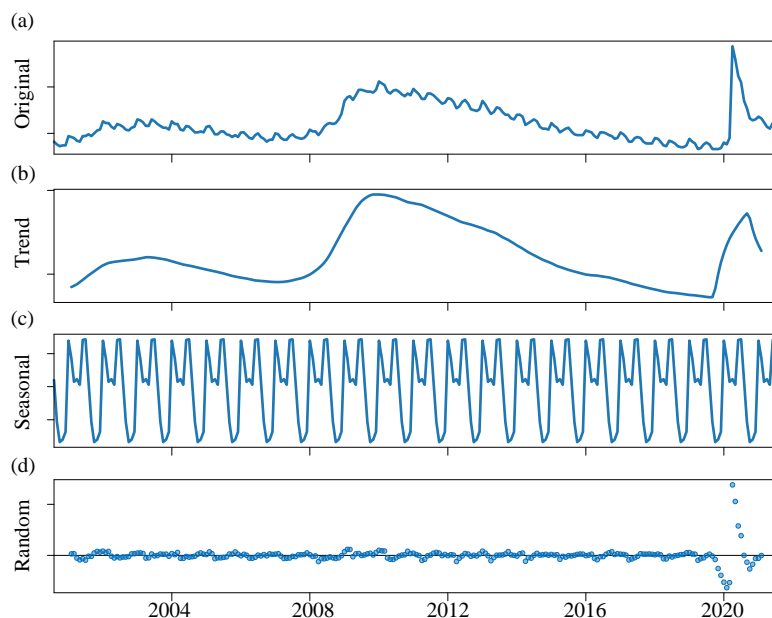


图 22. 失业率数据的分解

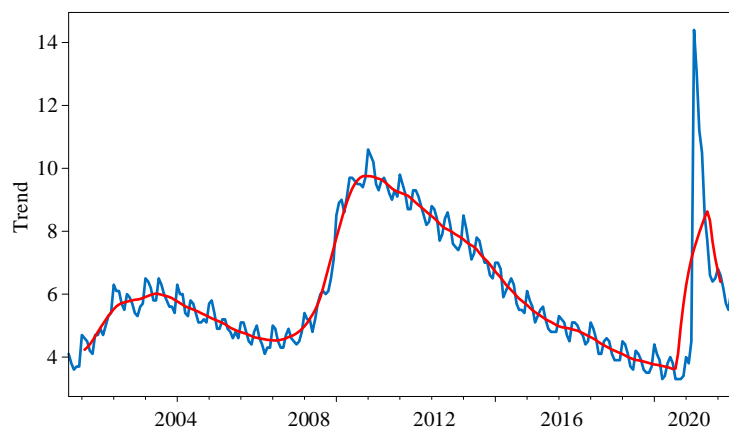


图 23. 比较原始数据和趋势成分

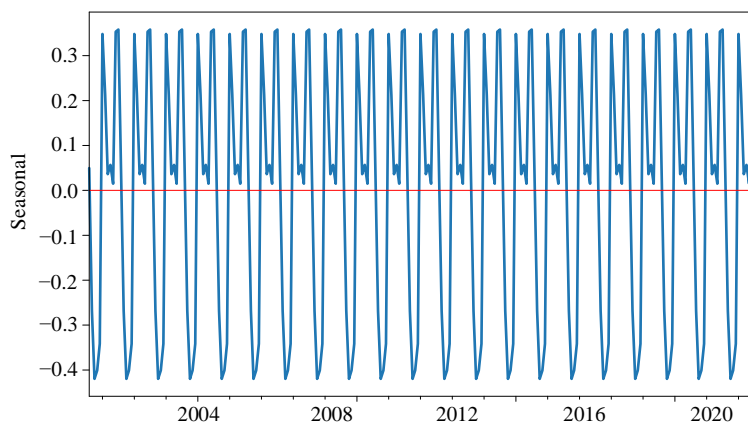


图 24. 季节成分

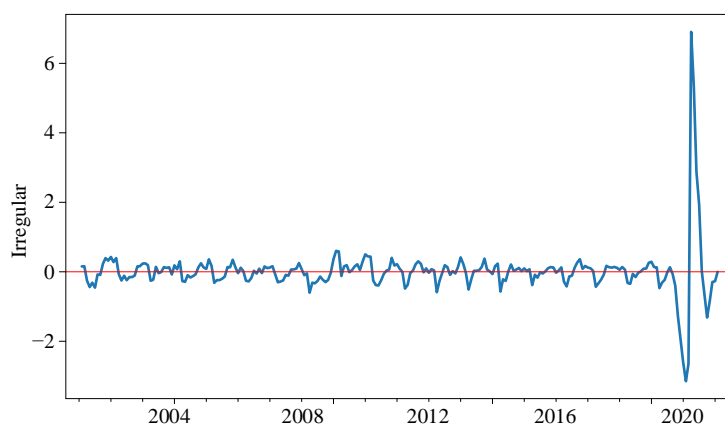


图 25. 噪音成分



Bk6_Ch06_03.ipynb 绘制本节图像。



时间序列是按照时间顺序排列的数据点序列，通常用于描述随时间变化的现象，如股价、气温、销售额等。时间序列分析可以揭示数据的趋势、季节性和其他周期性模式。

时间序列数据中常常会出现缺失值，可能是由于数据采集过程中的错误、设备故障或其他原因导致的。处理缺失值的方法包括插值或者直接删除包含缺失值的时间点。

时间序列的趋势成分是指数据在长期内呈现的整体上升或下降的变化趋势。趋势成分反映了数据的长期演变趋势，可以是线性的、非线性的、逐渐增长或减少的。

季节调整是为了消除时间序列中由于季节性变化引起的周期性模式。季节性通常是指在一年内某个固定时间范围内重复出现的模式，例如节假日、季节性销售高峰等。季节调整有助于更好地识别和理解时间序列中的趋势成分。处理时间序列数据时，常用的方法包括时间序列分解，将时间序列分解为趋势、季节性和残差等成分。这有助于更好地理解数据的结构，从而进行更准确的分析和预测。

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger: <https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com