7 Rolling Window
移动窗口

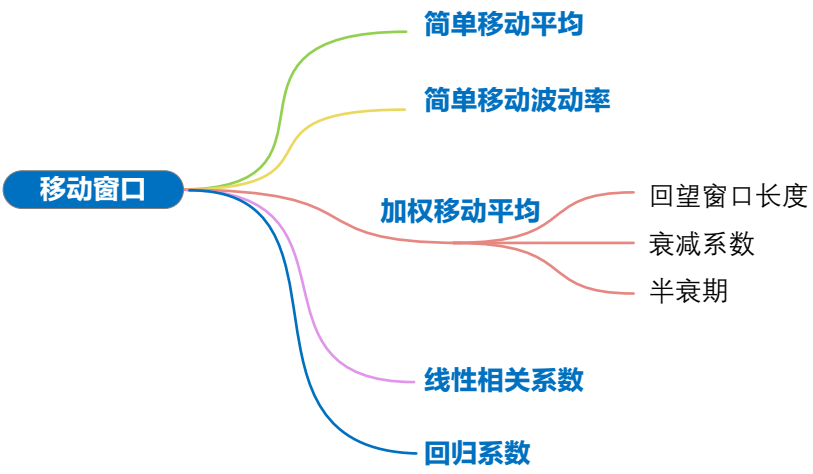
移动窗口展示数据之间动态关系

没有一种语言比数学更普遍、更简单、更没有错误、更不晦涩……更容易表达所有自然事物的不变关系。它用同一种语言解释所有现象，仿佛要证明宇宙计划的统一性和简单性，并使主导所有自然原因的不变秩序更加明显。

There cannot be a language more universal and more simple, more free from errors and obscurities...more worthy to express the invariable relations of all natural things than mathematics. It interprets all phenomena by the same language, as if to attest the unity and simplicity of the plan of the universe, and to make still more evident that unchangeable order which presides over all natural causes.

—— 约瑟夫·傅里叶 (Joseph Fourier) | 法国数学家、物理学家 | 1768 ~ 1830

- `df.ewm().mean()` 计算数据帧 `df` EWMA 平均值
- `df.ewm().std()` 计算数据帧 `df` EWMA 标准差/波动率
- `df.rolling().corr()` 计算数据帧 `df` 的移动相关性
- `df.rolling().kurt()` 计算数据帧 `df` 滚动峰度
- `df.rolling().max()` 计算数据帧 `df` 滚动最大值
- `df.rolling().mean()` 计算数据帧 `df` 滚动均值
- `df.rolling().min()` 计算数据帧 `df` 滚动最小值
- `df.rolling().quantile()` 计算数据帧 `df` 滚动百分位值
- `df.rolling().skew()` 计算数据帧 `df` 滚动偏度
- `df.rolling().std()` 计算数据帧 `df` MA 平均值
- `statsmodels.regression.rolling.RollingOLS()` 计算移动 OLS 线性回归系数



7.1 移动窗口

移动窗口 (rolling window, moving window) 是一种重要的时间序列统计计算方法。移动窗口按照一定规律沿着历史数据移动，每一个位置都产生一个统计量，比如最大值、最小值、平均值、加权平均值、标准差等等。移动窗口方法可以消除时间序列中的随机噪声，减少数据波动，更好地反映数据的趋势和周期性。

随着移动窗口不断滚动，特定统计量不断产生；因此，通过移动窗口得到的数据是序列数据，也就是时间序列。移动窗口的宽度叫做**回望窗口长度** (lookback window length)。

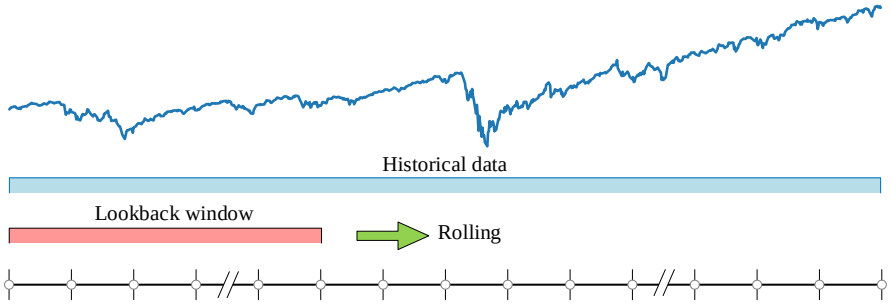


图 1. 移动窗口

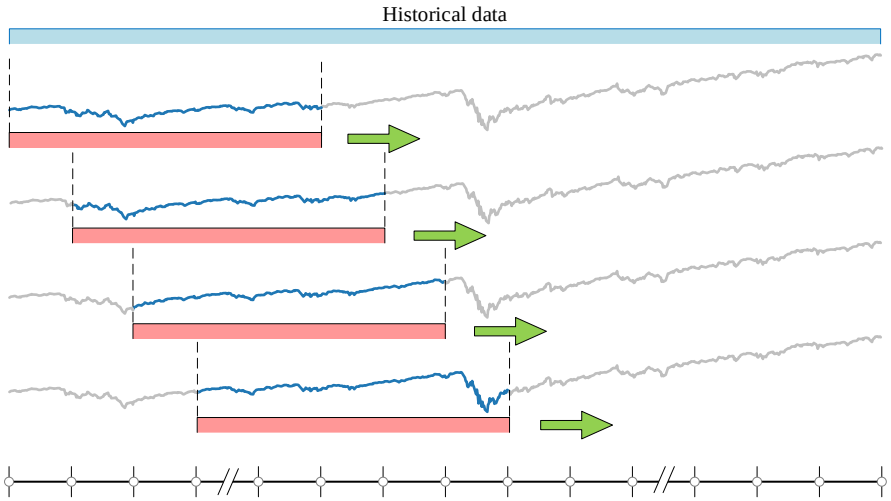


图 2. 移动窗口不断移动产生新的时间序列

最大、最小

如图 3 所示，利用长度为 100 营业日的回望窗口，我们可以得到移动最大值 (橙色) 和移动最小值 (绿色) 曲线。随着移动窗口移动到每一个位置，便利用回望窗口内的数据产生一个最大值和最小值。当移动窗口最左端和历史数据的最左端对齐时，产生第一个数据；而这个数据位于移动

窗口的最右端。因此，移动窗口数据长度比历史数据长度短。对于某个数据帧数据 df，移动最大值和最小值时间序列可以利用 df.rolling().max() 和 df.rolling().min()两个函数计算得到。

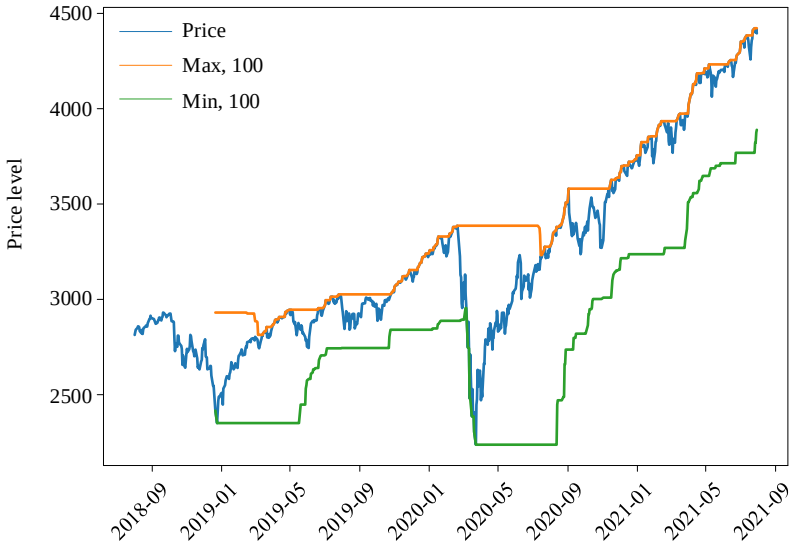


图 3. 移动最大和最小，回望窗口长度为 100

简单移动平均

简单移动平均数 (simple moving average, SMA)，是时间序列分析中常用的一种方法，用于平滑时间序列数据。SMA 的计算方法是将某一时间段内的数据求平均值，然后移动到下一个时间段内，继续计算平均值，如此重复直到计算完整个时间序列。SMA 具体运算如下：

$$\begin{aligned}\bar{x}_{\text{SMA\_}k} &= \frac{x_{k-L+1} + x_{k-L+2} + \ldots + x_{k-2} + x_{k-1} + x_k}{L} \\ &= \frac{x_{(k-L)+1} + x_{(k-L)+2} + \ldots + x_{k-2} + x_{k-1} + x_k}{L} \\ &= \frac{1}{L}\sum_{i=1}^{L} x_{(k-L)+i}\end{aligned} \tag{1}$$

SMA 有助于消除短期波动带来的数据噪音，突出长期趋势。移动平均相当于一个滤波器；回望窗口长度影响着统计量数据平滑度。SMA 的计算过程中，每个数据点的权重相等，因此对于较短的时间段，SMA 能够更好地反映数据的短期趋势和波动性，但对于长期趋势和周期性较弱的数据，则可能不太准确。

图 5 比较回望窗口分别为 50、100 和 150 三种情况的移动平均值。可以发现，回望窗口越长，得到的统计量时间序列看起来越平滑。

对于数据帧数据 df，移动平均可以用 df.rolling().mean() 计算得到。对于采样频率为营业日的数据，常见的移动窗口回望长度可以是 5 天 (一周)、10 天 (两周)、20 天 (一个月)、60 天 (一个季度)、125/126 天 (半年) 或 250/252 天 (一年) 等等。

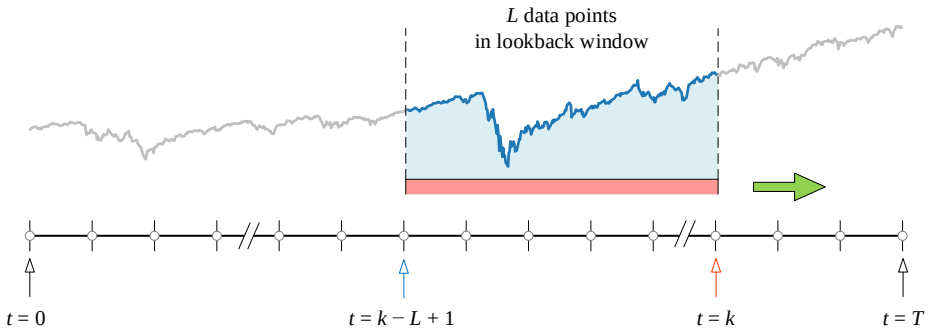


图 4. 回望窗口内数据序号

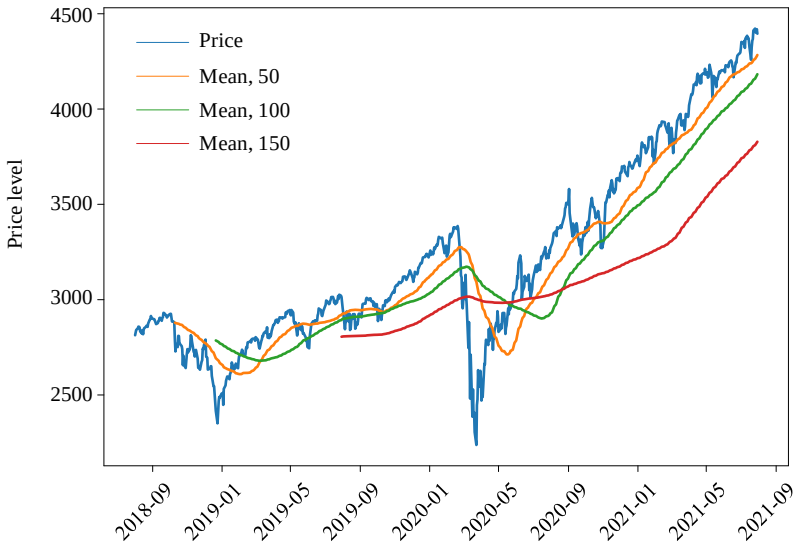


图 5. 移动平均，不同窗口长度

其他统计量

此外，移动窗口还可以帮助我们理解数据统计特点的动态特征。图 6 所示为日收益率的移动期望、波动率、偏度和峰态。**波动率** (volatility) 就是标准差。可以发现数据的统计特征随着时间移动不断改变。

对于数据帧数据 df，df.rolling().std()、df.rolling().skew() 和 df.rolling().kurt() 可以分别计算滚动标准差、偏度和峰度。

请大家改变回望窗口长度比较结果。

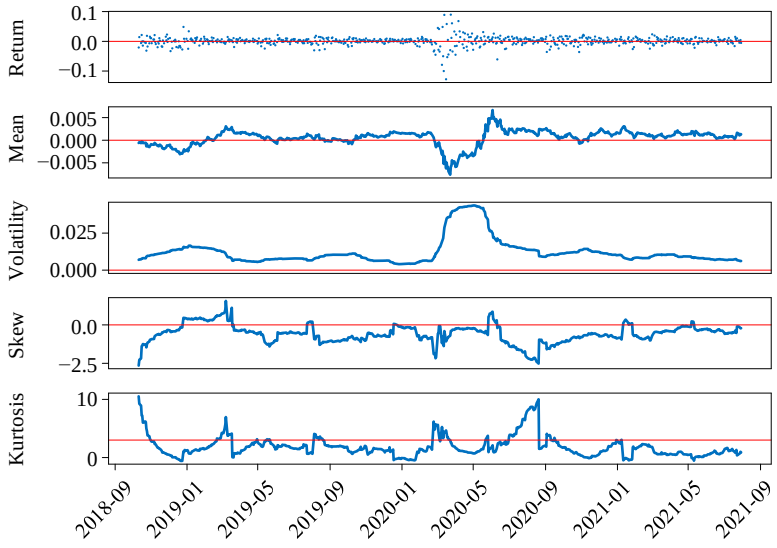


图 6. 日收益率的移动期望、波动率 (标准差)、偏度和峰态

类似地，图 7 所示为日收益率的 95%和 5%移动百分位变化。对于数据帧数据 df，df.rolling().quantile() 计算滚动百分位值。

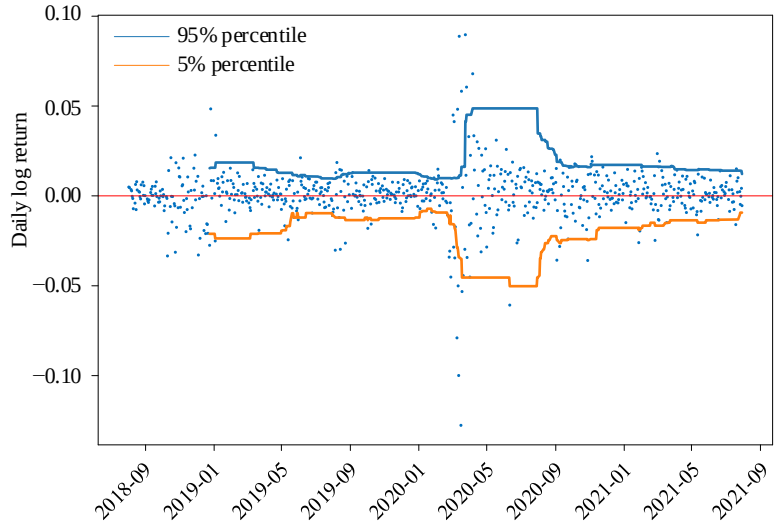


图 7. 移动百分位，95%和 5%

7.2 移动波动率

回望窗口长度为 L 的条件下，时间序列移动波动率为：

$$\sigma_{\text{daily\_}k} = \sqrt{\frac{1}{L-1}}\sqrt{\sum_{i=1}^{L}\left(x_{(k-L)+i}-\mu\right)^2} \tag{2}$$

其中，μ 为回望窗口内数据 x_i 的平均值。时间序列波动率的大小可以反映时间序列数据的风险程度，即数据变化的不确定性程度。通常情况下，波动率越大，数据变化的不确定性越高，风险也就越大。在金融市场分析中，时间序列波动率被广泛应用于风险管理和投资决策。例如，股票的波动率可以帮助投资者评估其风险水平，从而做出更明智的投资决策。

当 L 足够大，且 μ 几乎为 0 时，(2) 可以简化为：

$$\sigma_{\text{daily}} = \sqrt{\frac{\sum_{i=1}^{L}\left(x_{(k-L)+i}\right)^2}{L}} \tag{3}$$

(3) 相当于对回望窗口内 $(x_i)^2$ 数据，施加完全相同的权重 $1/L$；因此，这种波动率也被叫做**移动平均波动率** (moving average volatility)。

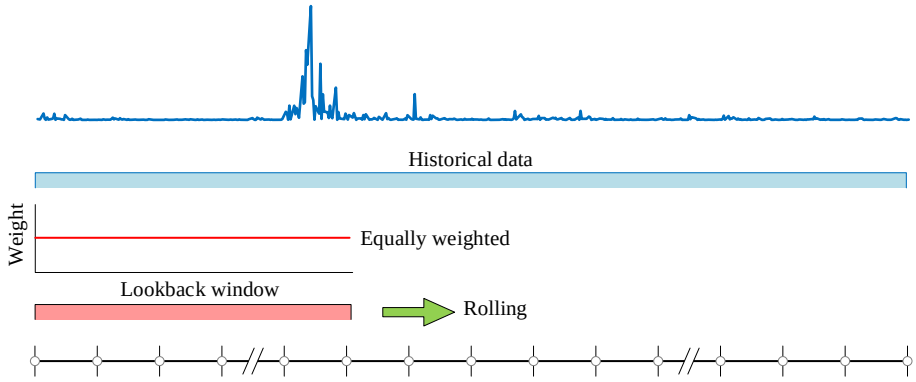


图 8. 移动平均

(3) 常用来计算股票收益率的波动率。图 10 所示为不同窗口长度条件下得到的移动平均波动率。可以发现，窗口长度越长数据越平缓，但是对数据变化响应越缓慢。

白话说，回望窗口长度越长，窗口内相对更具影响力的“陈旧”数据越尾大不掉，代谢的周期越长。下一节介绍的指数加权移动平均 EWMA，便很好地解决这一问题；哪怕回望窗口越长，EWMA 计算得到的波动率也能更快地跟踪数据变化规律。

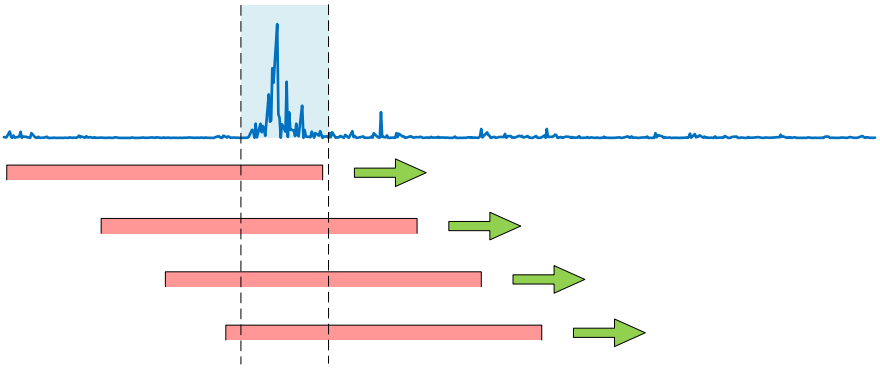


图 9. 尾大不掉的“陈旧”数据

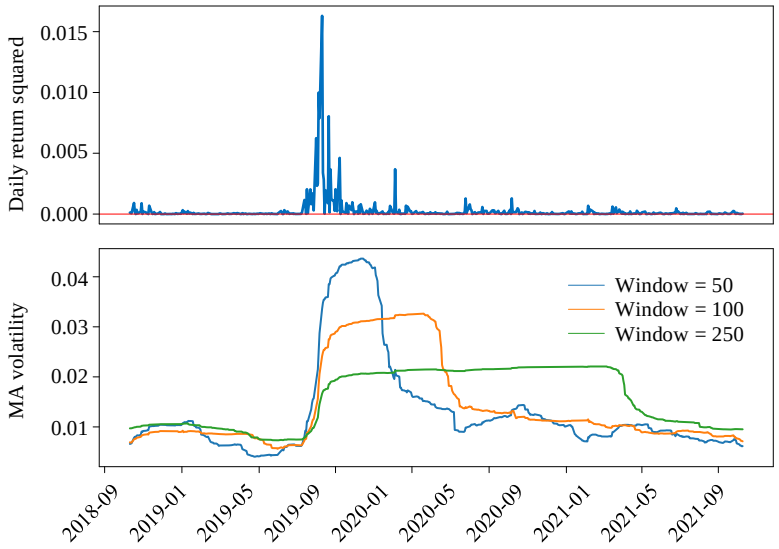


图 10. 移动平均 MA 单日波动率，不同窗口长度

此外，$\pm 2\sigma$ 波动率带常用来检测时间数据中可能存在的异常值。+2σ 曲线被称之为+2σ 上轨，−2σ 曲线常被称之为−2σ 下轨。图 11~图 13 分别展示窗口长度为 50 天、100 天和 250 天的$\pm 2\sigma$ 移动平均 MA 波动率带宽。

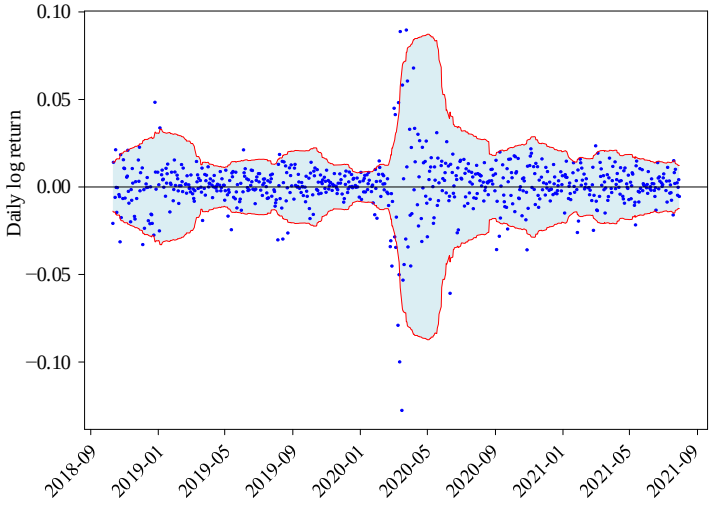


图 11. ±2σ 移动平均 MA 波动率带宽，窗口长度 50 天

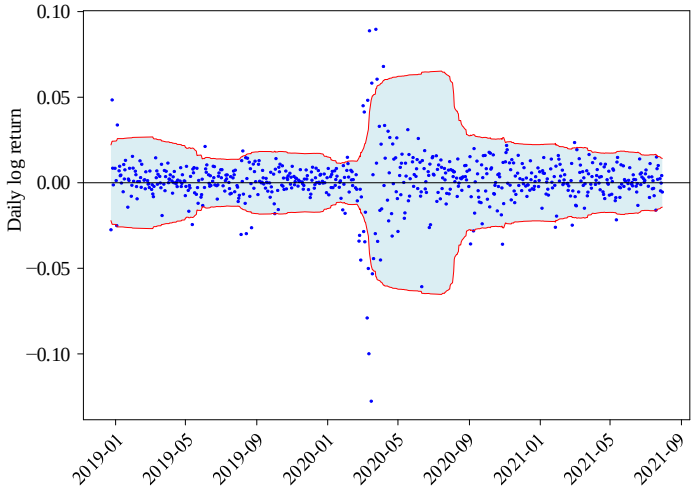


图 12. ±2σ 移动平均 MA 波动率带宽，窗口长度 100 天

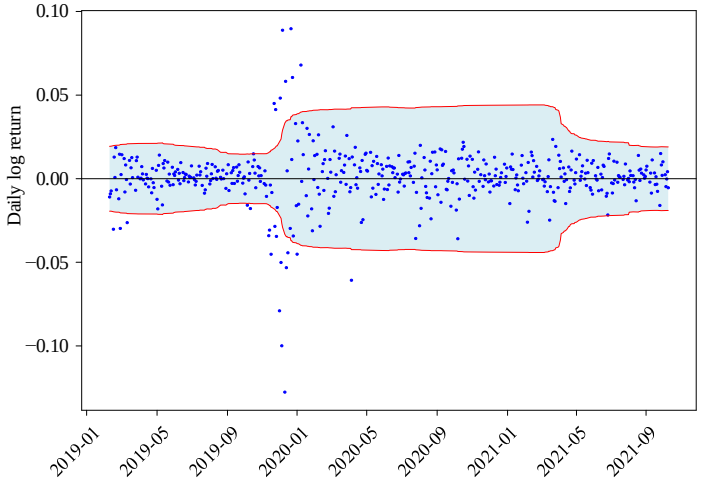


图 13. ±2σ 移动平均 MA 波动率带宽，窗口长度 250 天

时间平方根法则

时间平方根法则 (square root of time) 可以将日波动率转化为年化波动率：

$$\sigma_{annual} = \sqrt{250} \cdot \sigma_{daily} \tag{4}$$

式中的 250 代表假设一年有 250 营业日。图 14 所示为不同窗口长度条件下，移动平均 MV 年化波动率随时间变化情况。

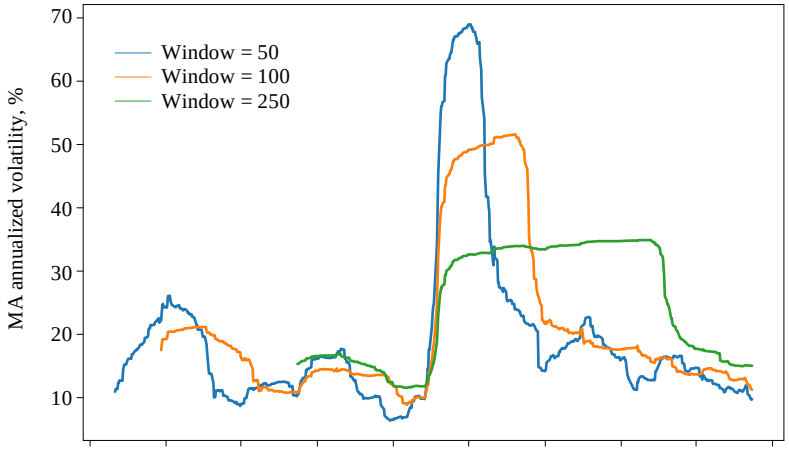


图 14. 移动平均 MV 年化波动率百分数，不同窗口长度

Bk6_Ch07_01.py 绘制上一节和本节主要图像。

7.3 指数加权移动平均

指数加权移动平均 (exponentially-weighted moving average, EWMA) 可以用来计算平均值、标准差、方差、协方差和相关性等等。EWMA 是对前文的简单移动平均的改进。EWMA 方法的特点是，对窗口内越近期的数据给予更高权重，越陈旧数据越低权重。权重的衰减过程为指数衰减。这种方法可以在平滑数据的同时保留较新数据的影响。

指数加权移动平均数 (exponential moving average, EMA, or exponentially weighted moving average) 定义为：

$$\bar{x}_{\text{EWMA\_}k} = \left(\frac{1-\lambda}{1-\lambda^{L}}\right)\frac{x_{k-L+1}\lambda^{L-1} + x_{k-L+2}\lambda^{L-2} + \ldots + x_{k-2}\lambda^{2} + x_{k-1}\lambda^{1} + x_{k}\lambda^{0}}{L} \tag{5}$$

其中，λ 为**衰减系数** (decay factor)。

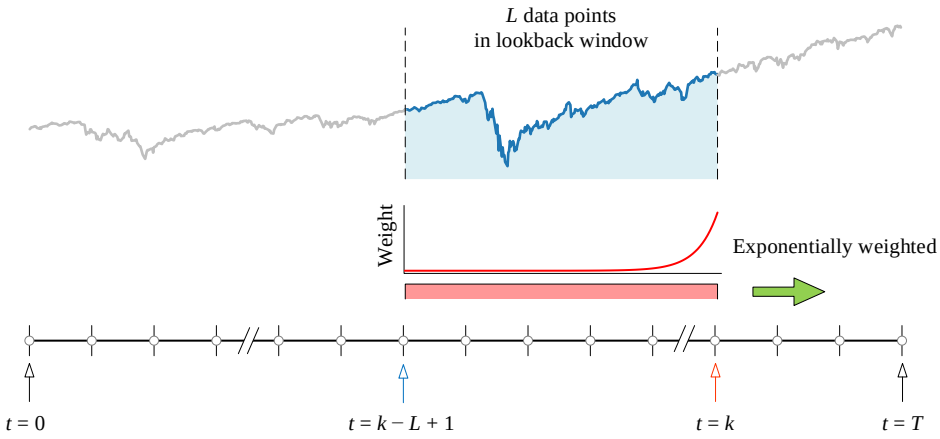


图 15. 回望窗口内数据指数加权移动平均

图 16 所示为 EWMA 权重随衰减系数变化。

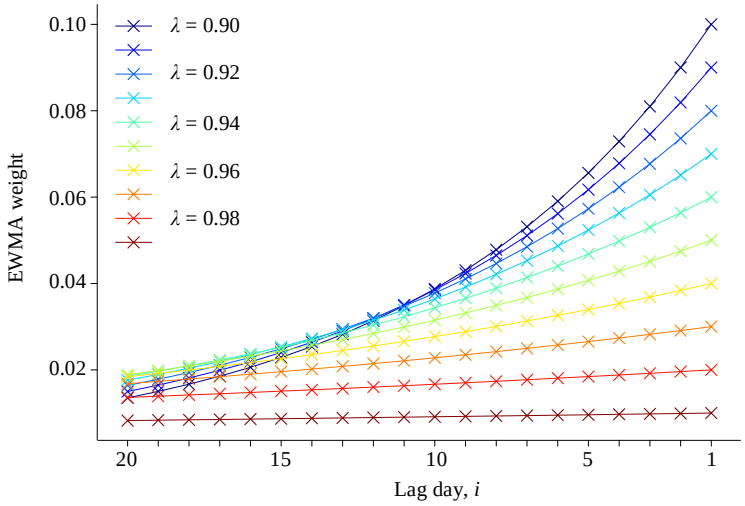


图 16. EWMA 权重随衰减系数变化

EWMA 中**半衰期** (half life, HL) 指的是权重衰减一半的时间，具体定义如下：

$$\lambda^{HL} = \frac{1}{2} \Leftrightarrow HL = \frac{\ln(1/2)}{\ln(\lambda)} \tag{6}$$

图 17 所示为半衰期 HL 随衰减系数 λ 变化。

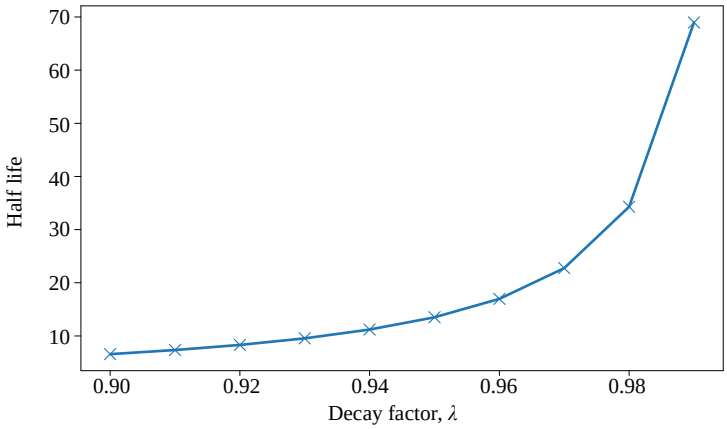


图 17. 半衰期随衰减系数变化

Bk6_Ch07_02.py 绘制图 16 和图 17。

图 18 所示为衰减因子不同条件下，EWMA 平均值变化情况。对比三条曲线，不难发现衰减系数 λ 越小 (比如红线)，EWMA 平均值更贴近真实趋势 (蓝线)，但是平滑度降低。

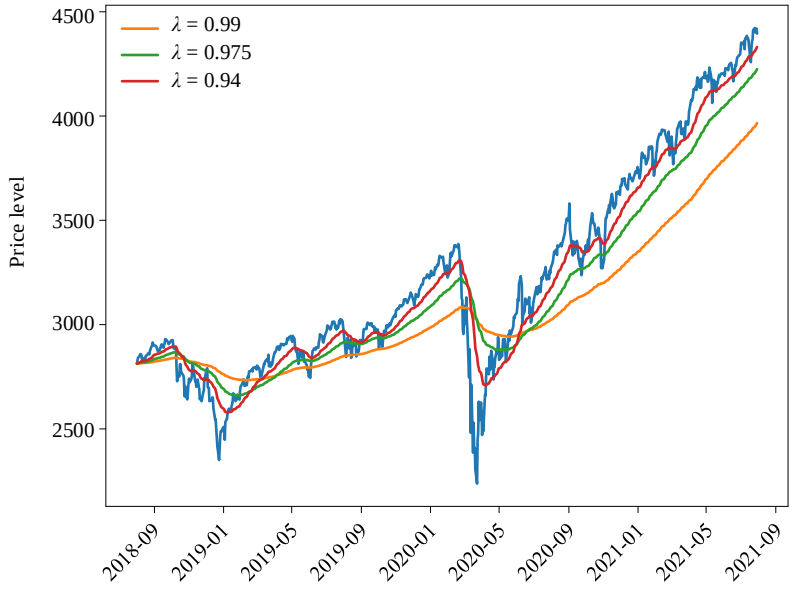


图 18. 指数加权移动平均

给定数据帧数据 df，df.ewm().mean() 可以用来计算指数加权移动平均。这个函数可以还是使用平滑系数 α。衰减因子 λ 与平滑系数 α 有关系如下：

$$\lambda = 1 - \alpha \tag{7}$$

容易得到 α 和半衰期 HL 关系：

$$\alpha = 1 - \exp\left(\frac{\ln(0.5)}{HL}\right) \tag{8}$$

7.4 EWMA 波动率

用 EWMA 方法计算波动率时，常使用如下迭代公式：

$$\sigma_n^2 = \lambda\sigma_{n-1}^2 + (1-\lambda)r_{n-1}^2 \tag{9}$$

其中，λ 为**衰减因子** (decay factor)；σ_n 是当前时刻的波动率；σ_{n-1} 是上一时刻的波动率；r_{n-1} 是上一时刻的回报率。

如下所示，列出四个时间点 n、$n-1$、$n-2$ 和 $n-3$ 的 EWMA 波动率计算式：

$$\begin{cases} \sigma_n^2 = \lambda\sigma_{n-1}^2 + (1-\lambda)r_{n-1}^2 \\ \sigma_{n-1}^2 = \lambda\sigma_{n-2}^2 + (1-\lambda)r_{n-2}^2 \\ \sigma_{n-2}^2 = \lambda\sigma_{n-3}^2 + (1-\lambda)r_{n-3}^2 \\ \sigma_{n-3}^2 = \lambda\sigma_{n-4}^2 + (1-\lambda)r_{n-4}^2 \end{cases} \tag{10}$$

将 (10) 几个算式依次迭代，可以得到：

$$\sigma_n^2 = (1-\lambda)\left(r_{n-1}^2 + \lambda r_{n-2}^2 + \lambda^2 r_{n-3}^2 + \lambda^3 r_{n-4}^2\right) + \lambda^4\sigma_{n-4}^2 \tag{11}$$

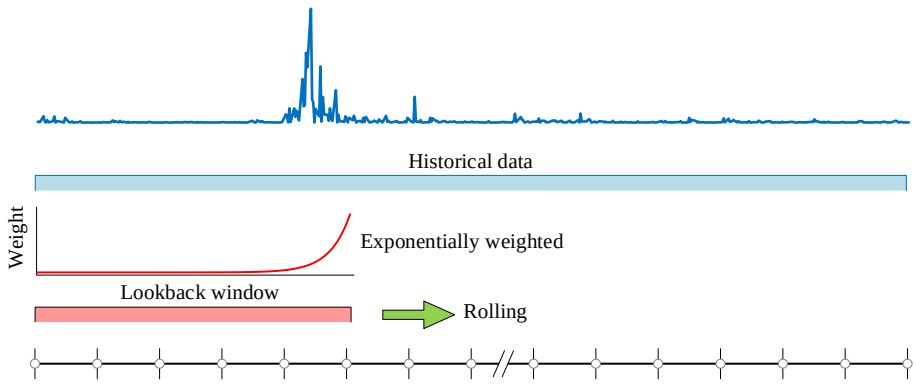


图 19. 指数加权移动平均计算波动率

图 20 所示为不同衰减因子条件下 EWMA 单日波动率。相比 MA 方法，EWMA 可以更快跟踪数据变化。衰减因子越小，跟踪速度越快。

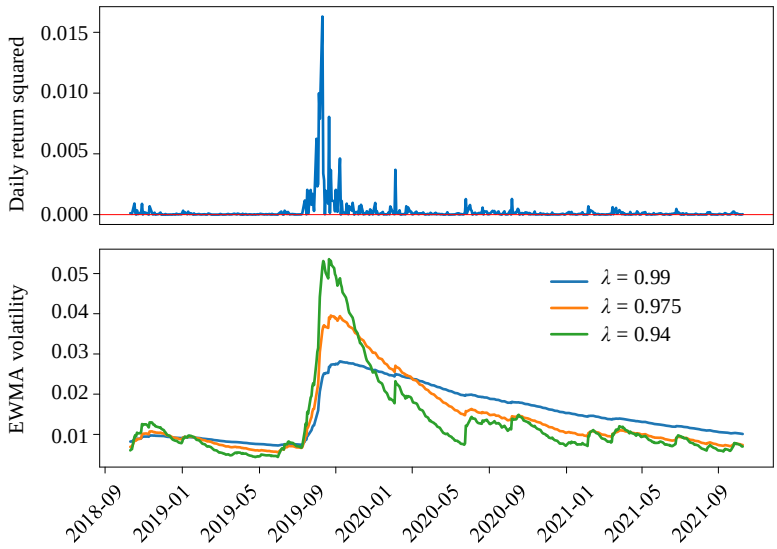


图 20. EWMA 单日波动率，不同衰减因子

图 21~图 23 分别展示衰减因子为 0.99、0.975 和 0.94 的$\pm 2\sigma$移动平均 MA 波动率带宽。

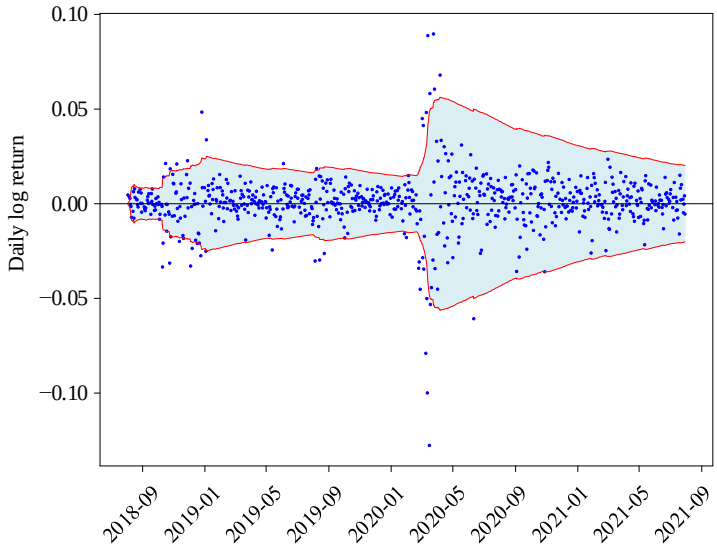


图 21. $\pm 2\sigma$ EWMA 波动率带宽，$\lambda = 0.99$

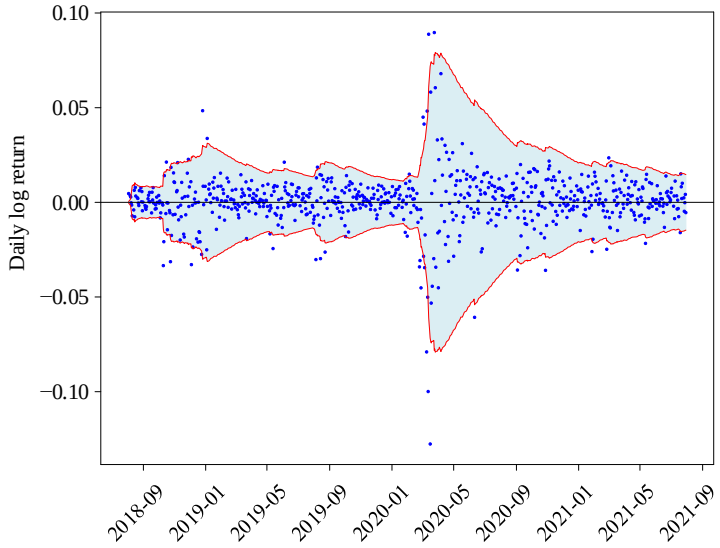


图 22. $\pm 2\sigma$ EWMA 波动率带宽，$\lambda = 0.975$

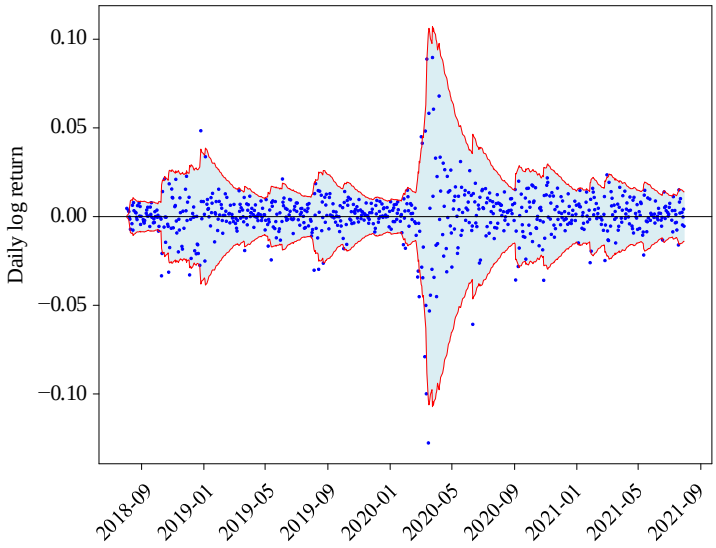


图 23. $\pm 2\sigma$ EWMA 波动率带宽，$\lambda = 0.94$

时间平方根法则将 EWMA 日波动率得到年化波动率。图 24 比较六个年化波动率。

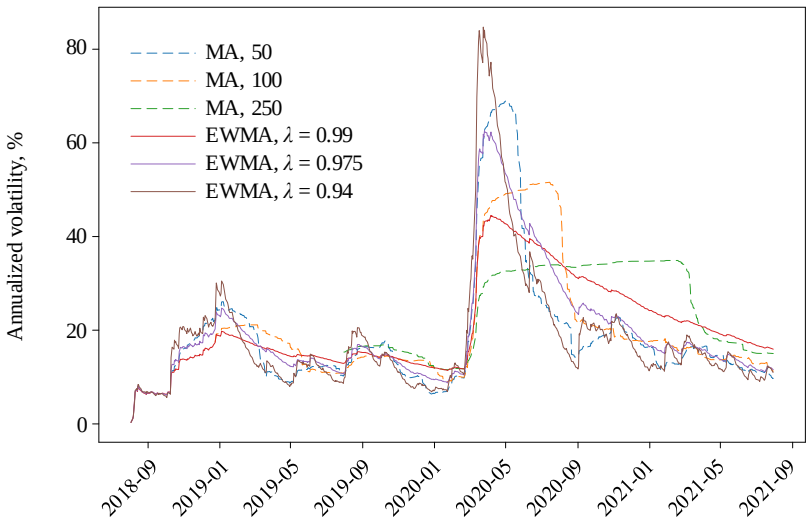


图 24. 比较 6 个年化波动率

7.5 相关性系数

除了平均值、波动率等，相关性系数也随着时间不断变化。df.rolling().corr() 可以计算数据帧 df 的移动相关性。图 25 所示为移动相关性系数。在处理数据时，但凡发现移动相关性系数发生剧烈波动时，都需要大家格外小心。因为移动相关性系数的陡然增大、降低，都可能是由为数不多的几个数据点造成的。而这几个数据点有可能是离群值，值得我们深入探究。

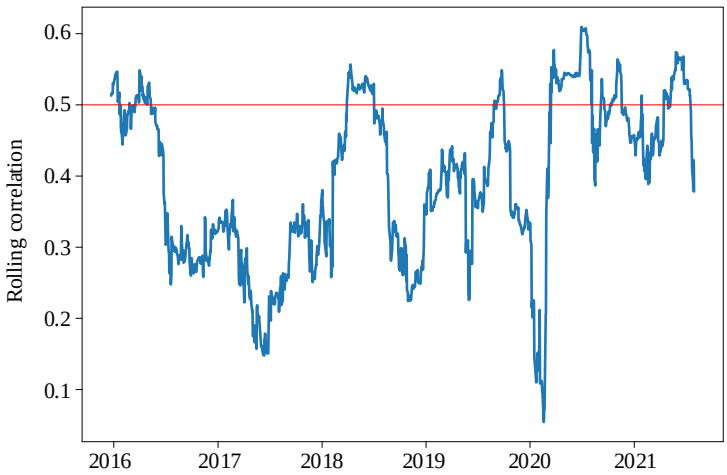


图 25. 移动相关性

Bk6_Ch07_03.py 绘制图 25。

7.6 回归系数

类似地，回归系数也随着移动窗口数据不断变化。

图 26 和图 27 用 statsmodels.regression.rolling.RollingOLS() 计算移动 OLS 线性回归系数。

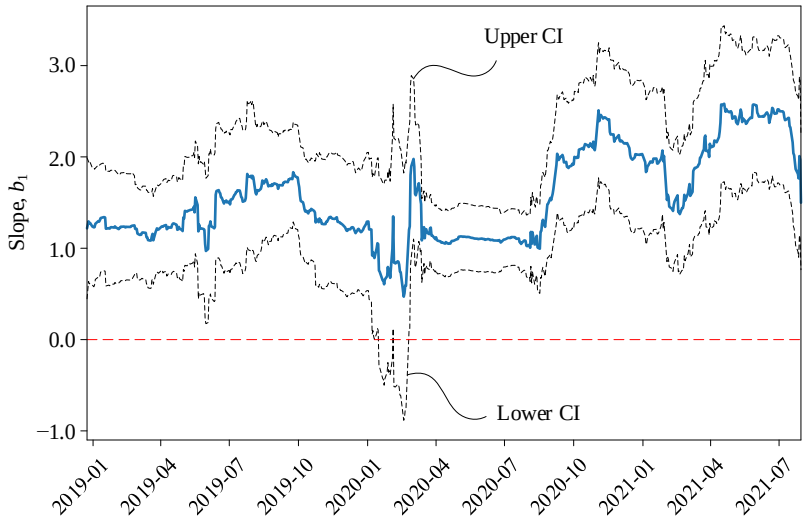


图 26. 回归斜率系数，移动窗口长度 100

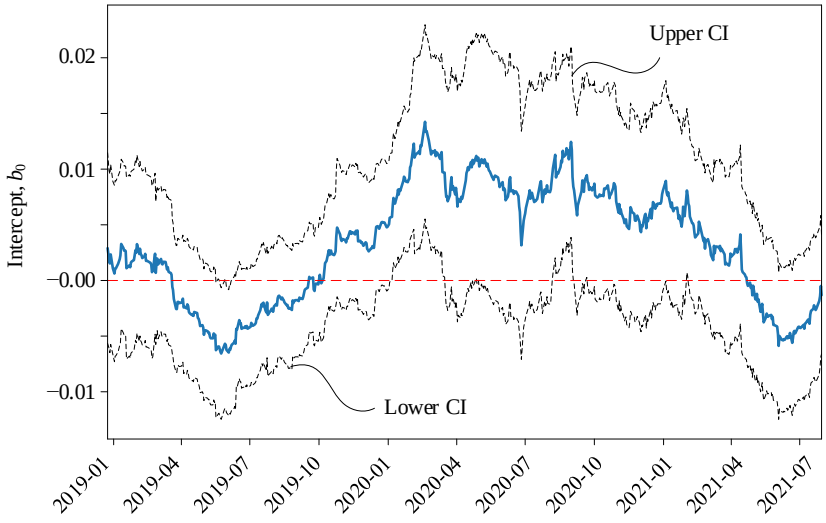
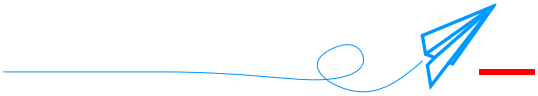


图 27. 回归截距系数，移动窗口长度 100

Bk6_Ch07_04.py 绘制图 26 和图 27。



总结来说，时间序列分析中，移动窗口是一种常用的技术，用于对时间序列数据进行平滑处理和预测分析。通过在时间序列上滑动固定大小的窗口，计算每个窗口中数据点的平均值或加权平均值来平滑数据。简单移动平均法 SMA 是最基本的移动窗口方法，它将窗口内的数据点简单平均处理，对于时间序列的短期波动有较好的平滑效果。移动波动率是指在移动窗口内计算的标准差或方差，它通常用于评估时间序列的波动性。指数加权移动平均法 EWMA 是一种加权移动平均方法，它通过指数函数来计算每个数据点的权重，使得较近期的数据点的权重更大，从而更好地捕捉跟踪到时间序列变化趋势。此外，相关性系数、线性回归系数也都随时间 (移动窗口变化)。