

11

Rolling Window

移动窗口

捕捉和分析统计量随时间变化的趋势和模式



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



```

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

```

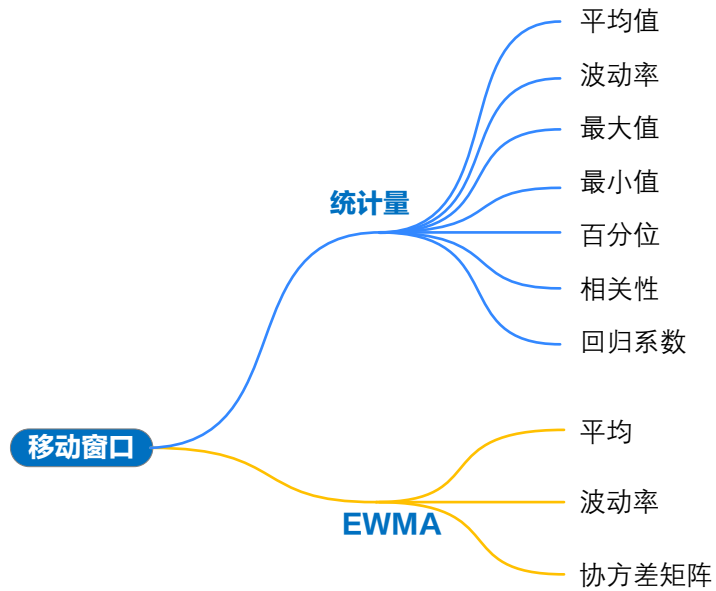
本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

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



11.1 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，如图 1 所示。移动窗口的宽度叫做**回望窗口长度** (lookback window length)。

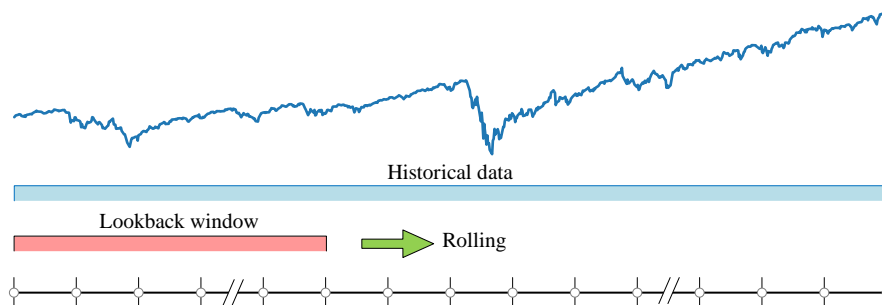


图 1. 移动窗口

如图 2 所示，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。随着移动窗口不断移动，该统计量不断产生；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。

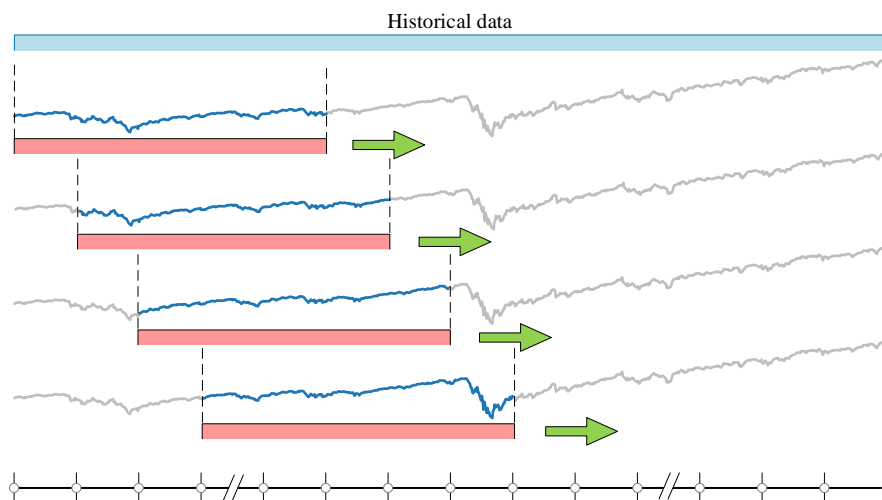


图 2. 移动窗口产生时间序列

最大、最小

如图 3 所示，利用长度为 100 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；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。

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

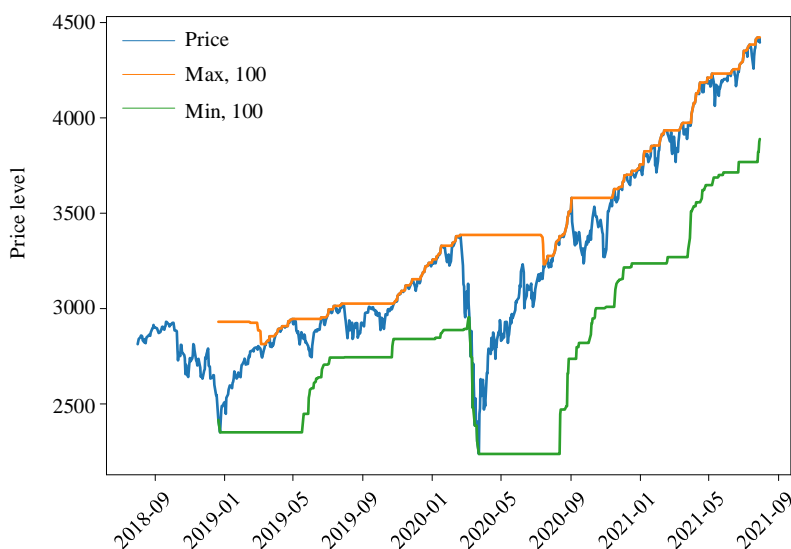


图 3. 移动最大和最小，回望窗口长度为 100

简单移动平均

简单移动平均 (simple moving average, SMA)，用来计算平均数的回望窗口内的每个样本的权重完全一致：

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

移动平均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。

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。

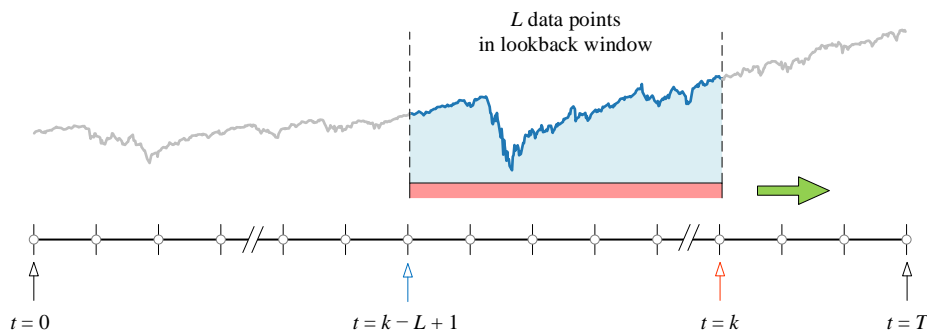


图 4. 回望窗口内数据序号

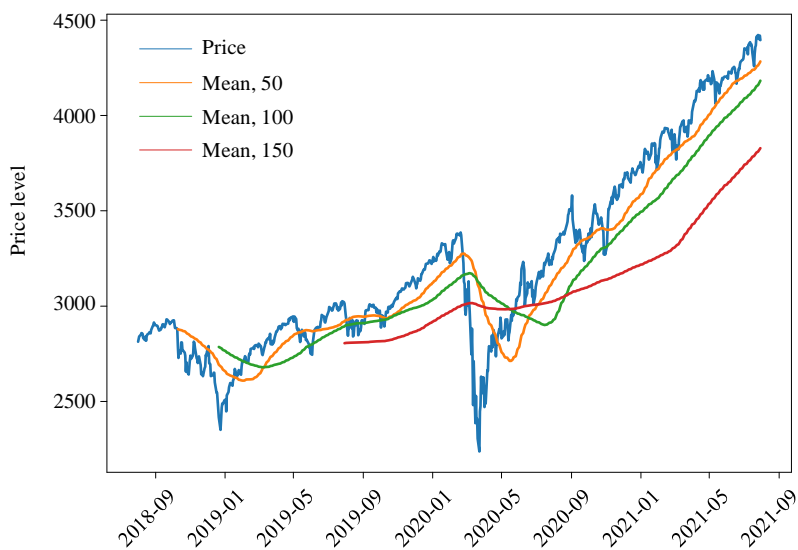


图 5. 移动平均，不同窗口长度

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()` 可以分别计算移动标准差、偏度和峰度。

请大家改变回望窗口长度比较结果。

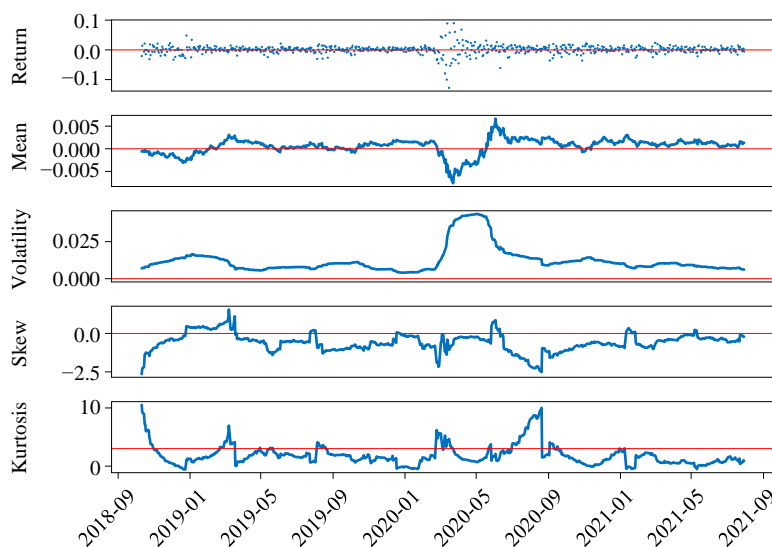


图 6. 日收益率的移动期望、波动率 (标准差)、偏度和峰态

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()` 计算移动百分位值。

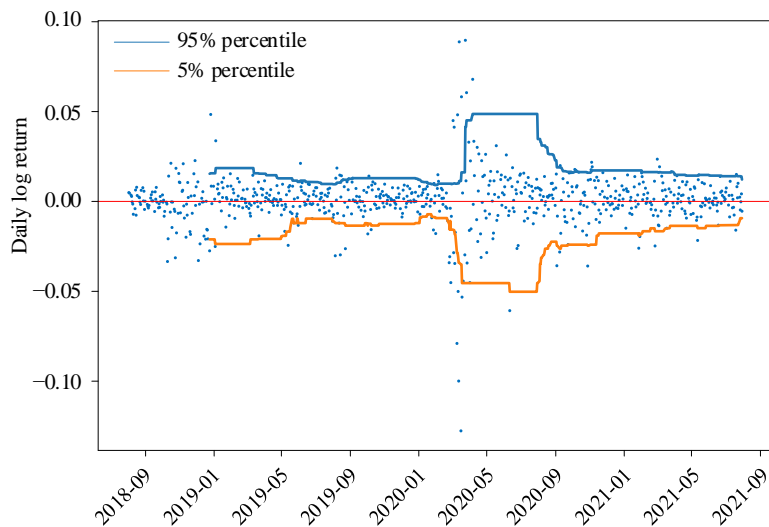


图 7. 移动百分位，95% 和 5%

《编程不难》介绍过，在使用 `pandas.DataFrame.rolling()` 计算移动窗口统计量时，我们还需注意参数 `center`。如图 8 (a) 所示，当设置 `center = False` 时，移动窗口的标签将被设置为窗口索引的右边缘；也就是说，窗口的标签与移动窗口的右边界对齐。这意味着移动窗口中的数据包括右边界，但不包括左边界。

如图 8 (b) 所示，当 `center=True` 时，移动窗口的标签将被设置为窗口索引的中心。也就是说，窗口的标签位于移动窗口的中间。这意味着移动窗口中的数据将包括左右两边的数据，并且标签位于窗口中央。

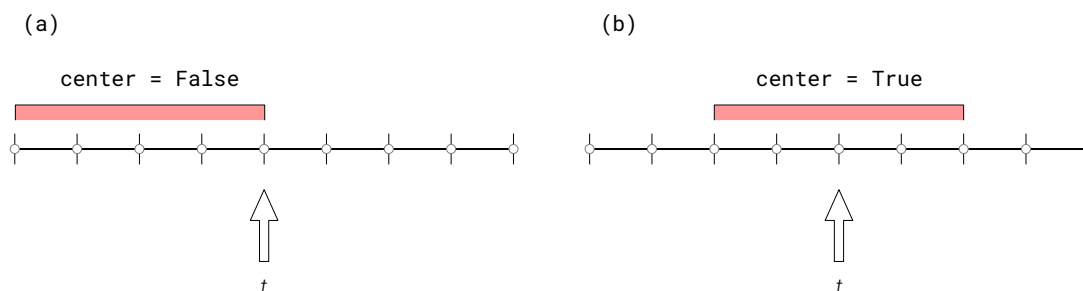


图 8. 移动窗口位置

11.2 移动波动率

回望窗口长度为 L 的条件下，时间序列 x_i 移动波动率/标准差为：

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

其中， μ 为回望窗口内数据 x_i 的平均值。

当 L 足够大，且 μ 几乎为 0 时，(2) 可以简化为：

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

观察 (3)，可以发现相当于对回望窗口内 $(x_i)^2$ 数据，施加完全相同的权重 $1/L$ 。

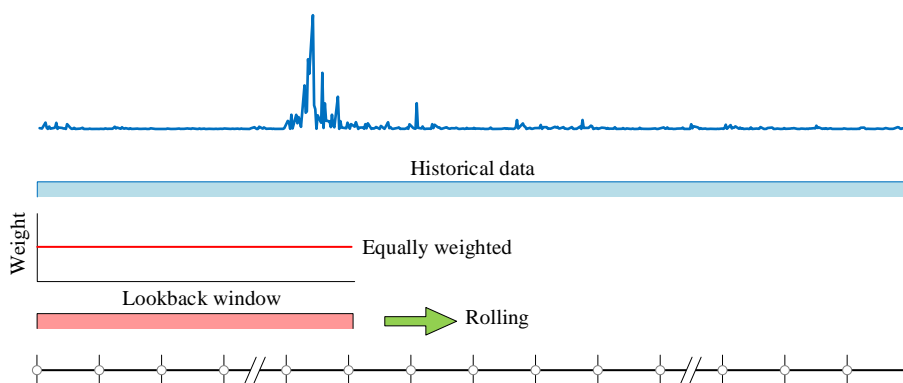


图 9. 移动平均

(3) 常用来计算股票收益率的波动率。图 11 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。

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

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。本章最后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。这是本章后文要介绍的内容。

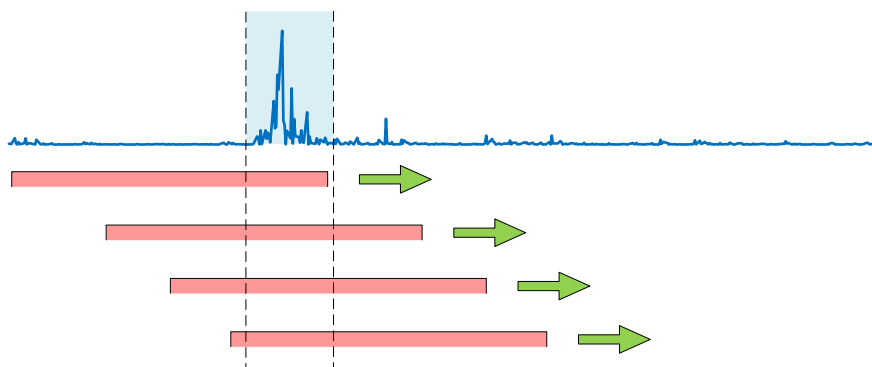


图 10. 尾大不掉的“陈旧”数据

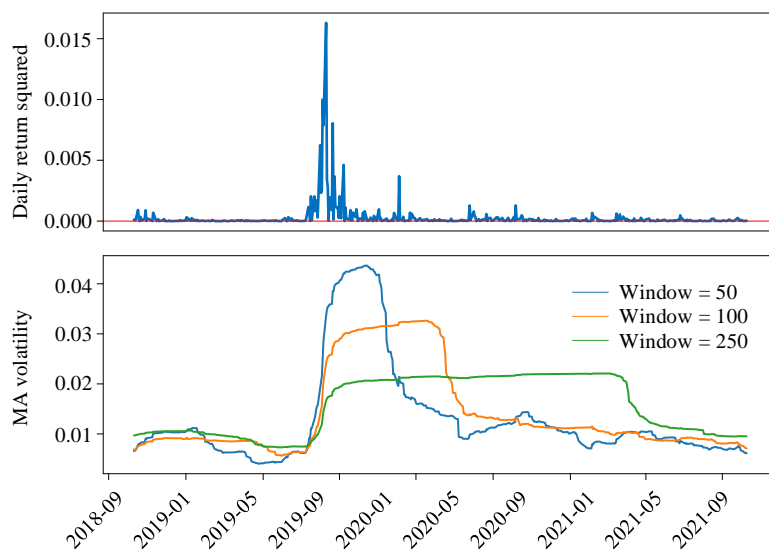
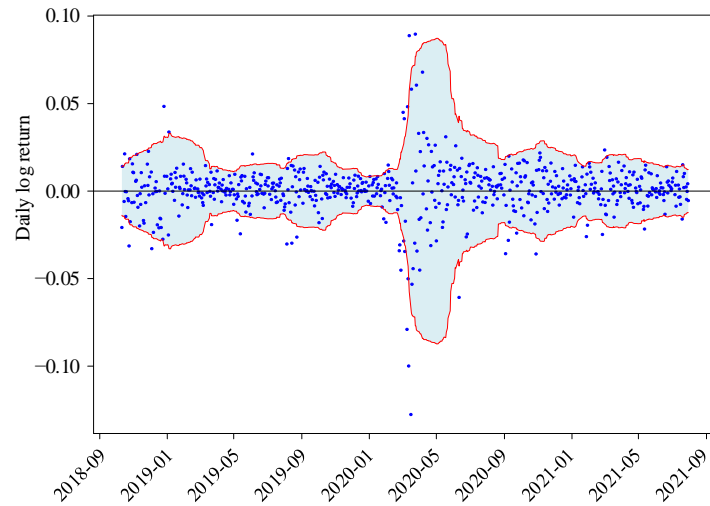
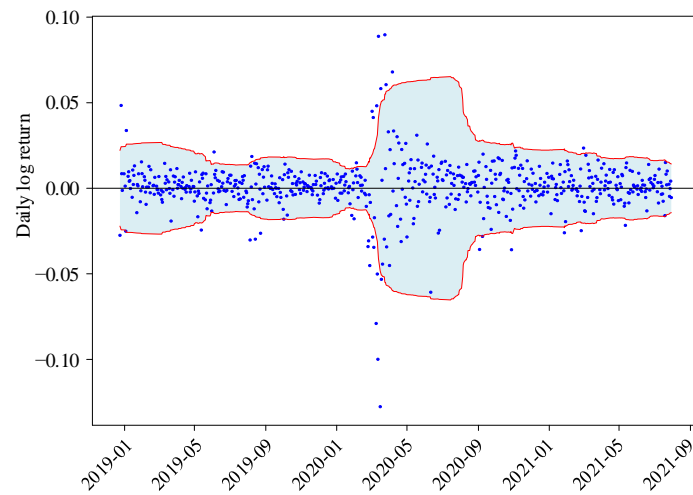
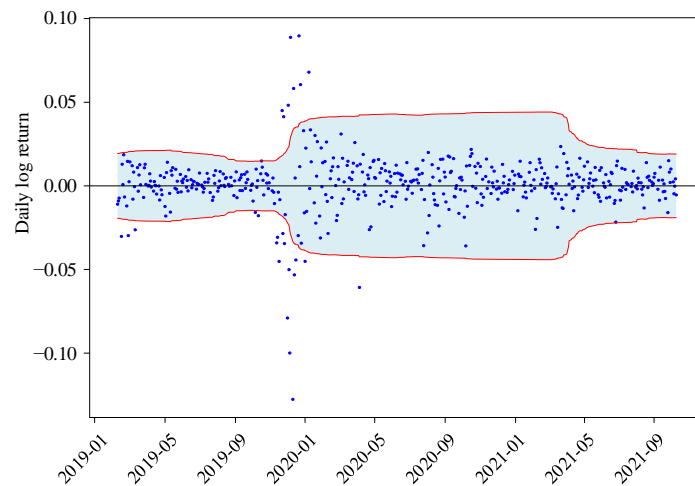


图 11. 移动平均 MA 单日波动率，不同窗口长度

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。 $+2\sigma$ 曲线被称之为 $+2\sigma$ 上轨， -2σ 曲线常被称之为 -2σ 下轨。图 12~图 14 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 12. $\pm 2\sigma$ 移动平均 MA 波动率带宽，窗口长度 50 天图 13. $\pm 2\sigma$ 移动平均 MA 波动率带宽，窗口长度 100 天图 14. $\pm 2\sigma$ 移动平均 MA 波动率带宽，窗口长度 250 天

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_Ch11_01.py 绘制本节主要图像。

11.3 相关性

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 15 所示为移动相关性。

《编程不难》还专门介绍过如何计算并处理成对相关系数，如图 16 所示。请大家回顾学习。

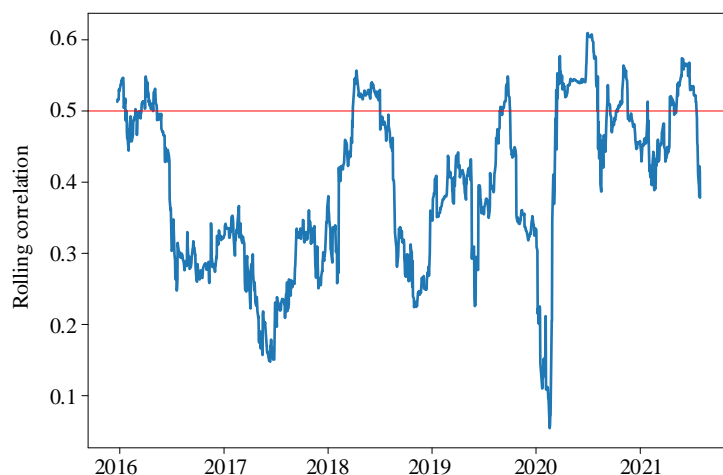


图 15. 移动相关性

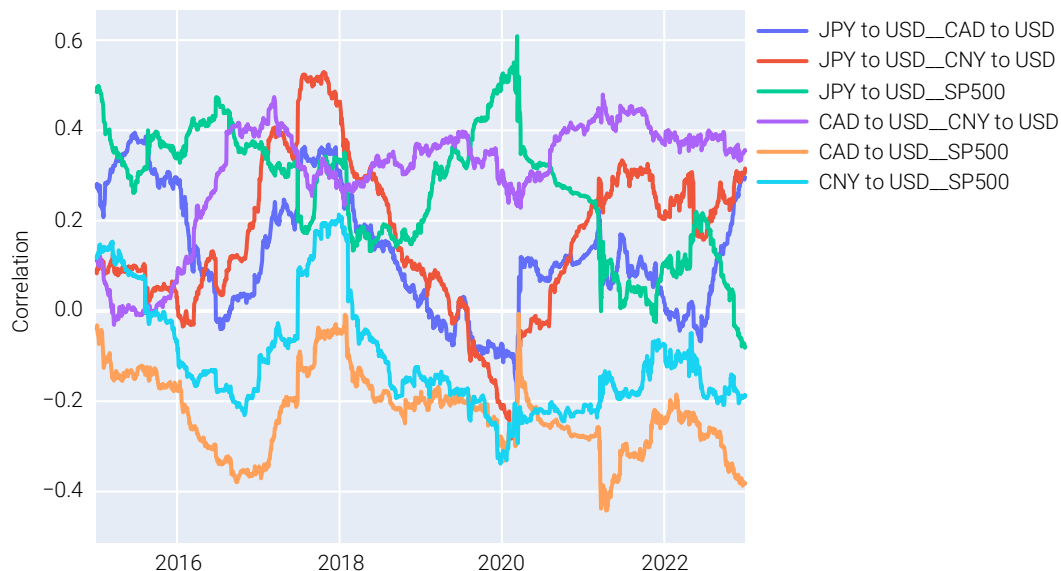
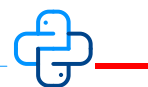


图 16. 成对移动相关性，图片来自《编程不难》



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_Ch11_02.py 绘制图 15。

11.4 回归系数

类似地，回归系数也随着移动窗口数据不断变化。

本节利用 `statsmodels.regression.rolling.RollingOLS()` 计算移动 OLS 线性回归系数。

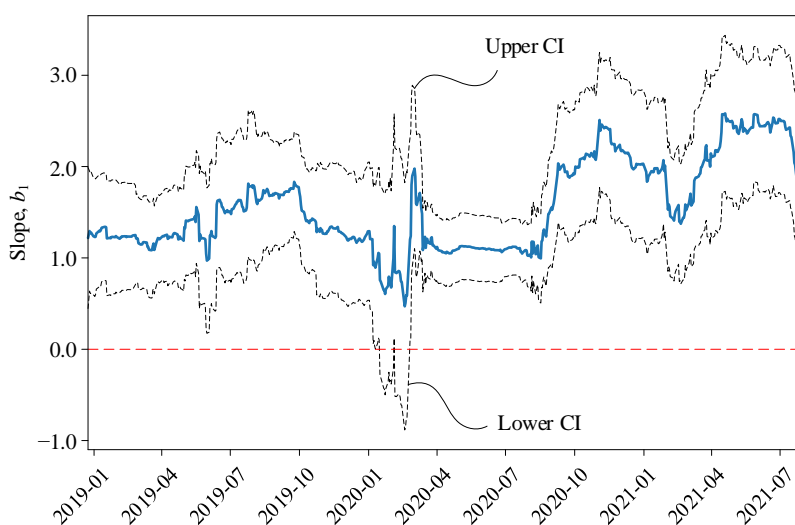


图 17. 回归斜率系数，移动窗口长度 100

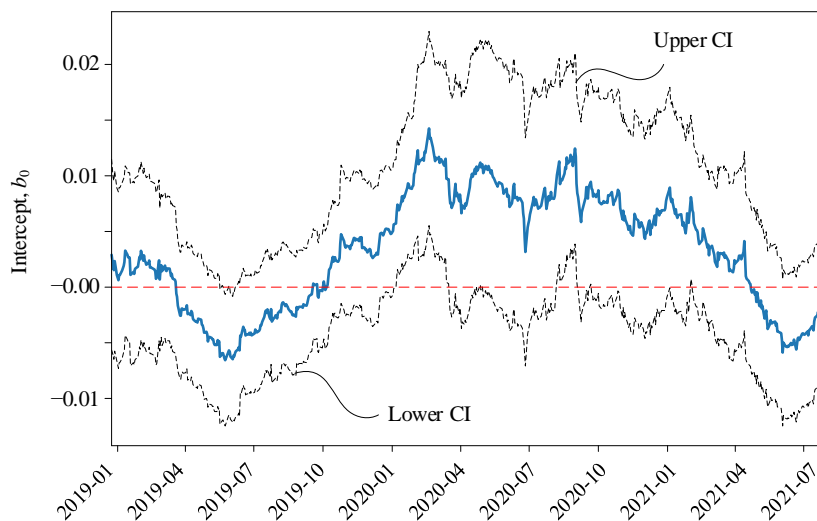


图 18. 回归截距系数，移动窗口长度 100



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_Ch11_03.py 绘制图 17 和图 18。

11.5 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 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。

指数加权移动平均 (exponential moving average, EMA, or exponentially weighted moving average) 可以通过如下公式计算

$$\bar{x}_{\text{EWMA}} = \left(\frac{1-\lambda}{1-\lambda^L} \right) \frac{x_{k-L+1}\lambda^{L-1} + x_{k-L+2}\lambda^{L-2} + \dots + x_{k-2}\lambda^2 + x_{k-1}\lambda^1 + x_k\lambda^0}{L} \quad (4)$$

其中， λ 为**衰减系数** (decay factor)。

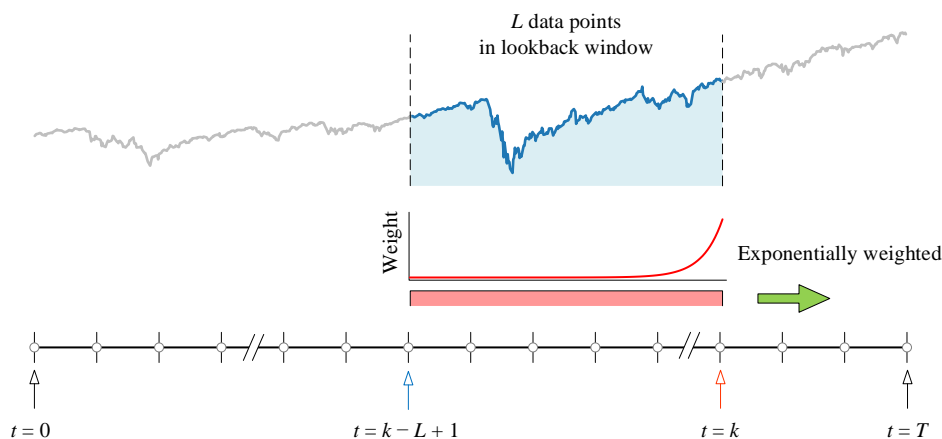


图 19. 回望窗口内数据指数加权移动平均

图 20 所示为 EWMA 权重随衰减系数变化。

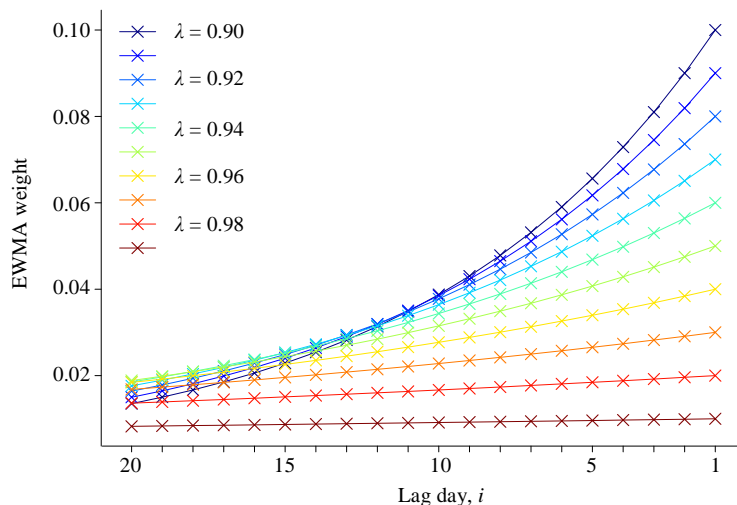


图 20. EWMA 权重随衰减系数变化

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

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

$$\lambda^{HL} = \frac{1}{2} \Leftrightarrow HL = \frac{\ln(1/2)}{\ln(\lambda)} \quad (5)$$

图 21 所示为半衰期 HL 随衰减系数变化。

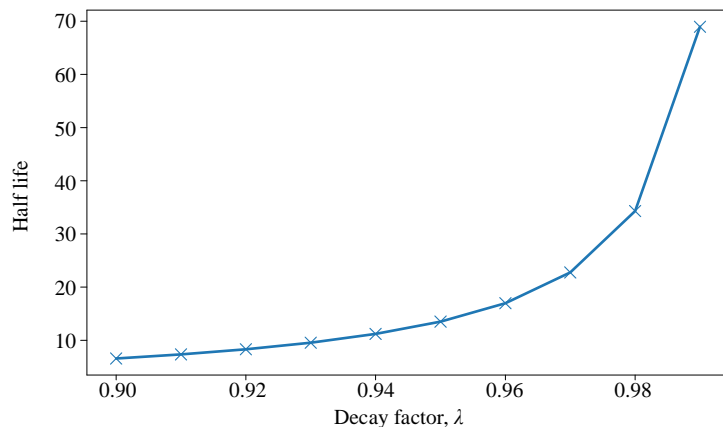


图 21. 半衰期随衰减系数变化

图 22 所示为衰减因子不同条件下，EWMA 平均值变化情况。

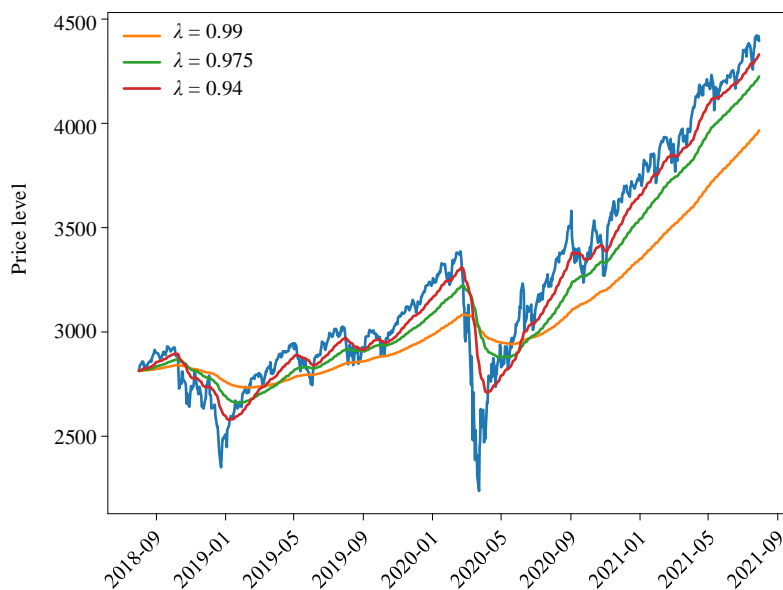


图 22. 指数加权移动平均

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：

$$\lambda = 1 - \alpha \quad (6)$$

可以得到 α 和半衰期 HL 关系：

$$\alpha = 1 - \exp\left(\frac{\ln(0.5)}{HL}\right) \quad (7)$$



Bk6_Ch11_04.py 绘制图 20 和图 21。

11.6 EWMA 波动率

用 EWMA 方法计算波动率时，常使用如下迭代公式：

$$\sigma_n^2 = \lambda \sigma_{n-1}^2 + (1 - \lambda) r_{n-1}^2 \quad (8)$$

其中， λ 为衰减因子； 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。

上式也可以看做是一种“贝叶斯推断”。 σ_{n-1}^2 代表“先验”，权重为 λ ； r_{n-1}^2 代表“新数据”，权重为 $1 - \lambda$ 。

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：

$$\begin{cases} \sigma_n^2 = \lambda \sigma_{n-1}^2 + (1 - \lambda) r_{n-1}^2 \\ \sigma_{n-1}^2 = \lambda \sigma_{n-2}^2 + (1 - \lambda) r_{n-2}^2 \\ \sigma_{n-2}^2 = \lambda \sigma_{n-3}^2 + (1 - \lambda) r_{n-3}^2 \\ \sigma_{n-3}^2 = \lambda \sigma_{n-4}^2 + (1 - \lambda) r_{n-4}^2 \end{cases} \quad (9)$$

将 (9) 几个算式依次迭代，可以得到：

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

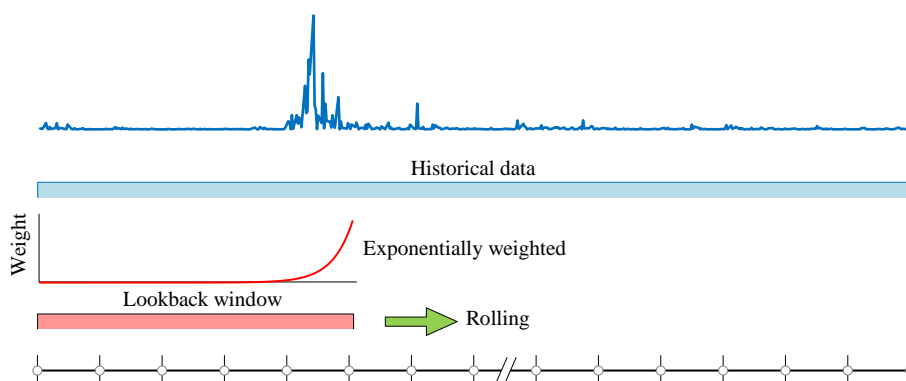


图 23. 指数加权移动平均计算波动率

图 24 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。

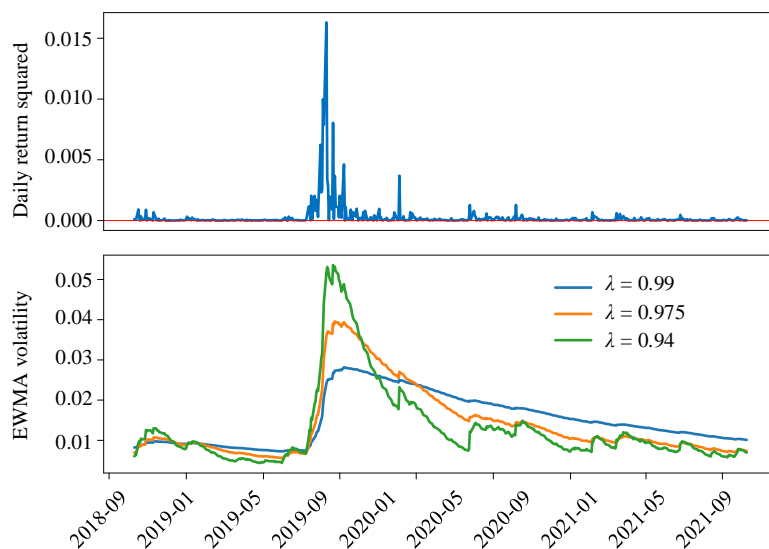


图 24. EWMA 单日波动率，不同衰减因子

图 25~图 27 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。

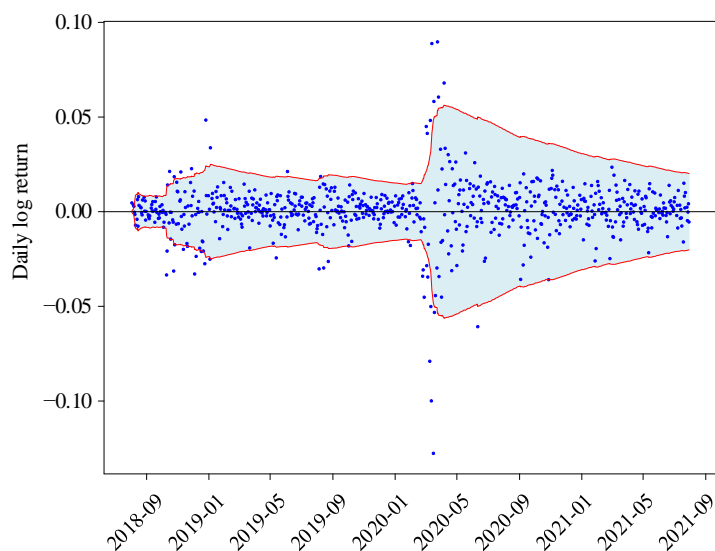
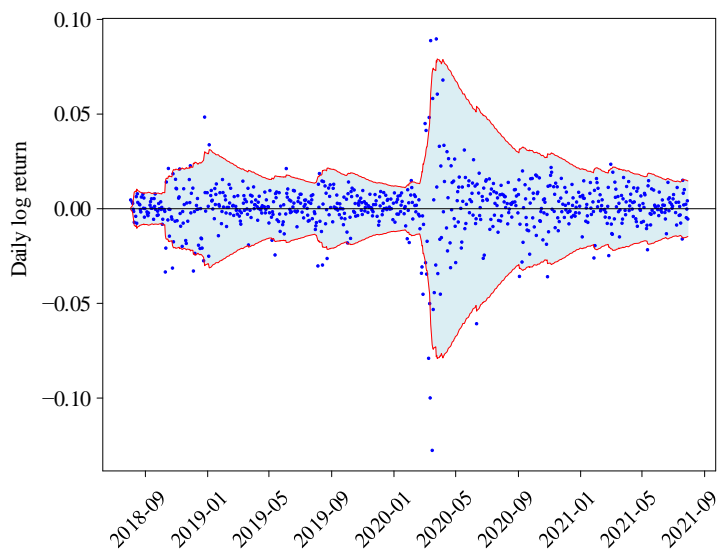
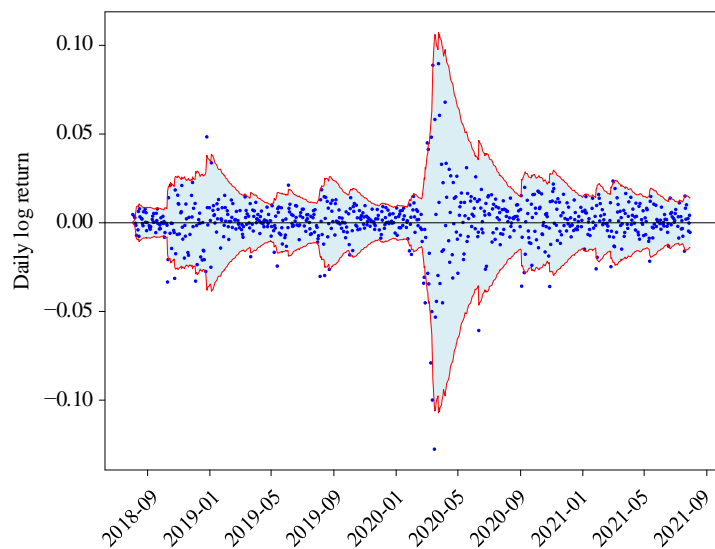


图 25. $\pm 2\sigma$ EWMA 波动率带宽， $\lambda = 0.99$

图 26. $\pm 2\sigma$ EWMA 波动率带宽, $\lambda = 0.975$ 图 27. $\pm 2\sigma$ EWMA 波动率带宽, $\lambda = 0.94$ 

Bk6_Ch11_05.py 绘制本节主要图像。

EWMA 协方差矩阵

既然 EWMA 可以用来计算波动率，这种方法也必然可以计算 EWMA 协方差矩阵。

如果用 r_1, r_2, \dots, r_D 代表 D 个特征，并假设移动窗口内一共有 L 个历史数据点 $r_j(1), r_j(2), \dots, r_j(L)$ 。序号 $i = 1, 2, \dots, L$ 代表时间点， $r_j(L)$ 代表 r_j 最新数据点。

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

为了计算 EWMA 协方差矩阵，我们首先构造矩阵 \mathbf{R}

$$\mathbf{R} = \sqrt{\frac{1-\lambda}{1-\lambda^L}} \begin{bmatrix} r_1(L) & r_2(L) & \cdots & r_D(L) \\ \lambda^{\frac{1}{2}} r_1(L-1) & \lambda^{\frac{1}{2}} r_2(L-1) & \cdots & \lambda^{\frac{1}{2}} r_D(L-1) \\ \vdots & \vdots & \ddots & \vdots \\ \lambda^{\frac{L-1}{2}} r_1(1) & \lambda^{\frac{L-1}{2}} r_2(1) & \cdots & \lambda^{\frac{L-1}{2}} r_D(1) \end{bmatrix} \quad (11)$$

其中， λ 的取值范围为 $0 < \lambda < 1$ 。假设 $r_j(i)$ 已经去均值。

EWMA 协方差矩阵便可以通过下式计算得到

$$\boldsymbol{\Sigma} = \mathbf{R}^T \mathbf{R} \quad (12)$$

其中，

$$\text{cov}(r_i, r_j) = \boldsymbol{\Sigma}_{i,j} = (\mathbf{R}^T \mathbf{R})_{i,j} = \frac{1-\lambda}{1-\lambda^L} \sum_{k=0}^{L-1} \lambda^k r_i(L-k) r_j(L-k) \quad (13)$$

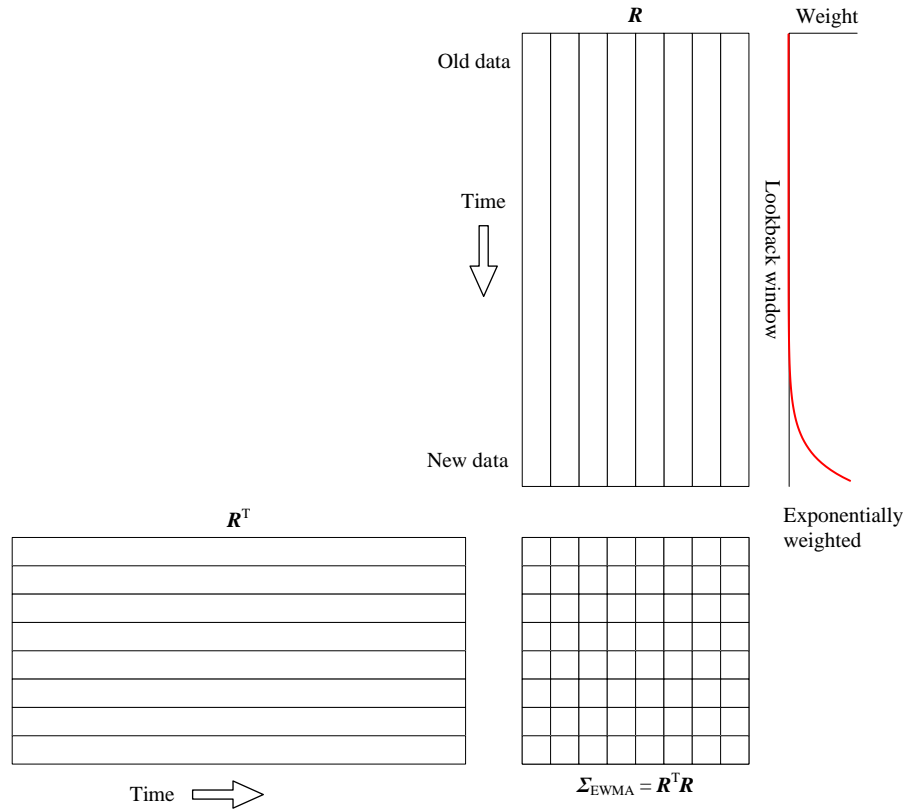


图 28. 计算 EWMA 协方差矩阵原理

特别地，当 λ 趋向于 1 时，

$$\lim_{\lambda \rightarrow 1} \frac{1-\lambda}{1-\lambda^L} = \frac{1}{L} \quad (14)$$

这便是一般的协方差矩阵中用到的等权重。

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

如图 29 所示，随着移动窗口不断移动，我们可以在每个时间点估计得到一个 EWMA 协方差矩阵；这也意味着，EWMA 协方差矩阵随时间变化。

这也很好理解，协方差矩阵的对角线元素为方差，非对角线元素为协方差。如果盯着图 29 中协方差矩阵某个位置元素看，这意味着我们看到的是方差或协方差随时间变化。同样的方法也适用于相关系数矩阵。图 30 所示为 4 个不同日期的 EWMA 相关性系数矩阵。

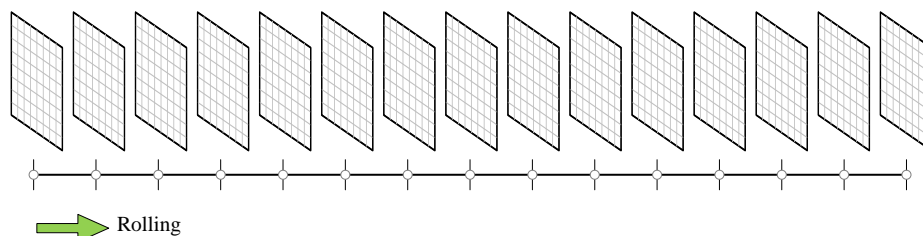


图 29. EWMA 协方差矩阵

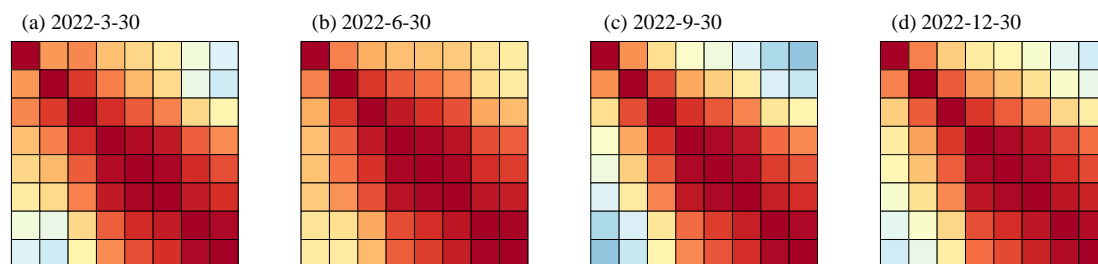


图 30. 4 个日期的 EWMA 相关性系数矩阵，衰减因子为 0.97

在时间序列分析中，移动窗口是一种常见的技术，用于计算某种统计量或指标的移动值。

一般来说，移动窗口是在时间序列上滑动的固定大小的窗口，用于计算各种统计量或指标，如平均值、最大值、最小值等等。通过在时间序列上滑动窗口，可以观察到数据在不同时间点的变化趋势。

移动波动率是在时间序列中使用移动窗口计算的波动率。它通常用于衡量时间序列中波动的变化，并可以帮助识别波动的趋势。

移动相关性是通过在两个时间序列上使用移动窗口计算相关系数，以观察它们之间的变化关系。这有助于识别时间序列之间的动态关系。

在移动窗口内使用回归分析来计算回归系数，以观察自变量和因变量之间的关系如何随时间演变。这对于捕捉变化关系的趋势非常有用。

指数加权移动平均 EWMA 是一种移动平均的方法，对不同时间点的数据赋予不同的权重。较近期的数据点被赋予更高的权重，而较远期的数据点则权重降低。这有助于更敏感地捕捉数据的短期变化。