学习笔记：Python for Finance Cookbook, 2nd

Stephen CUI

2023-11-11

目录

Chapter 1

Data Preprocessing

1.1   Changing the frequency of time series data

The formula for realized volatility is as follows:

$$RV = \sqrt{\sum_{i=1}^{T} r_t^2} \tag{1.1}$$

Realized volatility is frequently used for calculating the daily volatility using intraday returns.

1.2   Different ways of imputing missing data

Two of the simplest approaches to imputing missing time series data are:

• Backward filling—fill the missing value with the next known value

• Forward filling—fill the missing value with the previous known value

1.3   Different ways of aggregating trade data

The term **bars** refers to a data representation that contains basic information about the price movements of any financial asset.

冰山订单是大订单被分成较小的限价订单以隐藏实际订单数量。它们被称为"冰山订单",因为可见的订单只是"冰山一角",而大量的限价订单正在等待,准备下单。

Ideally, they would want to have a bar representation in which each bar contains the same amount of information. Some of the alternatives they are using include:

- **Tick bars**—named after the fact that transactions/trades in financial markets are often referred to as ticks. For this kind of aggregation, we sample an OHLCV bar every time a predefined number of transactions occurs.

- **Volume bars**—we sample a bar every time a predefined volume (measured in any unit, for example, shares, coins, etc.) is exchanged.

- **Dollar bars**—we sample a bar every time a predefined dollar amount is exchanged. Naturally, we can use any other currency of choice.

"Tick bars" 在金融领域通常指的是一种按照市场价格变动（ticks）来形成的交易条形图或时间序列。这种图表以市场价格的变动为基础，而不是固定的时间间隔。当市场价格变动达到一定的数量（例如，每次价格变动一个tick），就形成一个新的交易条。

这种方法的优势在于，它可以更好地反映市场的活动和波动，而不受时间因素的影响。相比于固定时间间隔的图表，tick bars 可以更好地捕捉市场的高活跃性时段，提供更精确的信息，尤其对于高频交易者和对市场活动敏感的策略来说更为有用。

Each of these forms of aggregations has its strengths and weaknesses that we should be aware of.

Tick bars offer a better way of tracking the actual activity in the market, together with the volatility. However, a potential issue arises out of the fact that one trade can contain any number of units of a certain asset. So, a buy order of a single share is treated equally to an order of 10,000 shares.

Volume bars are an attempt at overcoming this problem. However, they come with an issue of their own. They do not correctly reflect situations in which asset prices change significantly or when stock splits happen. This makes them unreliable for comparison between periods affected by such situations.

That is where the third type of bar comes into play—the dollar bars. It is often considered the most robust way of aggregating price data. Firstly, the dollar bars help bridge the gap with price volatility, which is especially important for highly volatile markets such as cryptocurrencies. Then, sampling by dollars is helpful to preserve the consistency of information. The second reason is that dollar bars are resistant to the outstanding amount of the security, so they are not affected by actions such as stock splits, corporate buybacks, issuance of new shares, and so on.

Chapter 2

Exploring Financial Time Series Data

In this chapter, we will cover the following recipes:

- Outlier detection using rolling statistics

- Outlier detection with the Hampel filter

- Detecting changepoints in time series

- Detecting trends in time series

- Detecting patterns in a time series using the Hurst exponent

- Investigating stylized facts of asset returns

2.1   Outlier detection using rolling statistics

2.2   Outlier detection with the Hampel filter

We will cover one more algorithm used for outlier detection in time series—the Hampel filter. Its goal is to identify and potentially replace outliers in a given series. It uses a centered sliding window of size $2x$ (given x observations before/after) to go over the entire series.

For each of the sliding windows, the algorithm calculates the median and the median absolute deviation (a form of a standard deviation).

> For the median absolute deviation to be a consistent estimator for the standard deviation, we
> have to multiply it by a constant scaling factor k, which is dependent on the distribution. For
> Gaussian, it is approximately 1.4826.

2.3 Detecting changepoints in time series

A **changepoint** can be defined as a point in time when the probability distribution of a process or
time series changes, for example, when there is a change to the mean in the series.

2.4 检验资产收益的程式化事实

2.4.1 事实 1：Non-Gaussian distribution of returns

2.4.2 事实 2：波动性聚类

波动性聚类是一种模式，其中价格的大变化往往会伴随着大的变化（波动性较高的时
期），而价格的小变化之后往往会出现小的变化（波动性较低的时期）。

2.4.3 事实 3：收益率缺乏自相关性

2.4.4 事实 4：平方/绝对回报的自相关性较小且递减

While we expect no autocorrelation in the return series, it was empirically proven that we can
observe small and slowly decaying autocorrelation (also referred to as persistence) in simple nonlinear
functions of the returns, such as absolute or squared returns. This observation is connected to the
phenomenon we have already investigated, that is, volatility clustering.

2.4.5 Fact 5: Leverage effect

The leverage effect refers to the fact that most measures of an asset's volatility are negatively
correlated with its returns.

Chapter 3

Technical Analysis and Building Interactive Dashboards

3.1 Calculating the most popular technical indicators

Bollinger bands are a statistical method, used for deriving information about the prices and volatility of a certain asset over time. To obtain the Bollinger bands, we need to calculate the moving average and standard deviation of the time series (prices), using a specified window (typically, 20 days). Then, we set the upper/lower bands at K times (typically, 2) the moving standard deviation above/below the moving average. The interpretation of the bands is quite simple: the bands widen with an increase in volatility and contract with a decrease in volatility.

The **relative strength index (RSI)** is an indicator that uses the closing prices of an asset to identify oversold/overbought conditions. Most commonly, the RSI is calculated using a 14-day period and is measured on a scale from 0 to 100 (it is an oscillator). Traders usually buy an asset when it is oversold (if the RSI is below 30) and sell when it is overbought (if the RSI is above 70). More extreme high/low levels, such as 80–20, are used less frequently and, at the same time, imply stronger momentum.

The last considered indicator is the **moving average convergence divergence (MACD)**. It is a momentum indicator showing the relationship between two exponential moving averages (EMA) of a given asset's price, most commonly 26- and 12-day ones. The MACD line is the difference between the fast (short period) and slow (long period) EMAs. Lastly, we calculate the MACD signal line as a 9-day EMA of the MACD line. Traders can use the crossover of the lines as a trading signal. For example, it can be considered a buy signal when the MACD line crosses the signal line from below.

5

Chapter 4

Time Series Analysis and Forecasting

4.1 Time series decomposition

The components of time series can be divided into two types: systematic and non-systematic. The systematic ones are characterized by consistency and the fact that they can be described and modeled. By contrast, the non-systematic ones cannot be modeled directly. The following are the systematic components:

- **Level**—the mean value in the series.

- **Trend**—an estimate of the trend, that is, the change in value between successive time points at any given moment. It can be associated with the slope (increasing/decreasing) of the series. In other words, it is the general direction of the time series over a long period of time.

- **Seasonality**—deviations from the mean caused by repeating short-term cycles (with fixed and known periods).

The following is the non-systematic component:

- **Noise**—the random variation in the series. It consists of all the fluctuations that are observed after removing other components from the time series.

The classical approach to time series decomposition is usually carried out using one of two types of models: additive and multiplicative.

An **additive model** can be described by the following characteristics:

- Model's form—$y(t) = level + trend + seasonality + noise$

- Linear model—changes over time are consistent in size

- The trend is linear (straight line)

- Linear seasonality with the same frequency (width) and amplitude (height) of cycles over time

  A **multiplicative model** can be described by the following characteristics:

- Model's form—$y(t) = level * trend * seasonality * noise$

- Non-linear model—changes over time are not consistent in size, for example, exponential

- A curved, non-linear trend

- Non-linear seasonality with increasing/decreasing frequency and amplitude of cycles over time

Hodrick-Prescott 滤波器——虽然这种方法并不是真正的季节性分解方法，但它是一种数据平滑技术，用于消除与经济周期相关的短期波动。通过消除这些，我们可以揭示长期趋势。HP 滤波器常用于宏观经济学。您可以在 `statsmodels` 的 `hpfilter` 函数中找到它的实现。

4.2 Testing for stationarity in time series

时间序列分析中最重要的概念之一是平稳性（stationarity）。简单地说，平稳时间序列是一个属性不依赖于观察该序列的时间的序列。换句话说，平稳性意味着某个时间序列的数据生成过程（data-generating process, DGP）的统计特性不随时间变化。

更正式地说，平稳性有多种定义，其中一些定义比其他定义更严格。对于实际用例，我们可以使用一种称为弱平稳性（或协方差平稳性）的方法。对于要归类为（协方差）平稳的时间序列，它必须满足以下三个条件：

- 系列的平均值必须恒定

- 系列的方差必须是有限且恒定的

- 相同距离的周期之间的协方差必须恒定

ADF 和 KPSS 检验的一个潜在缺点是它们不允许结构中断的可能性，即数据生成过程的平均值或其他参数的突然变化。 Zivot-Andrews 测试允许该系列中出现单一结构断裂的可能性，但其发生时间未知。

Chapter 5

回测交易策略