Lecture 12

High Frequency Data Analysis - An Introduction

In our analysis of data, specifically financial data, we considered daily or monthly or quarterly data on the prices and the associated returns of a financial asset. Hence, we could also analyse the volatility with respect to the time frequency (daily or monthly or quarterly) at which the given data is collected. For example, we considered the daily closing prices as our data for the analysis. Recall that, a time series is a collection of observations made at *equally spaced* time points. So, in our above example of daily closing prices, it is assumed that these observations are made at a made regularly at the completion of every 24-hour period. But, in reality, this is **not** true. An exchange opens in the morning at around 9 am and closes by around 4 pm. The closing price of an asset is nothing but the *last traded price* on that day, which happens any time up to 4 pm. So, the information about its volatility is not for the whole 24-hour period. Also, it can often be seen that yesterday's closing price is not today's opening price. Hence, the daily data does not give any information on the *overnight volatility* in the prices.

While studying the price dynamics, one of the most important characteristics of interest to big institutional (or individual) investors is the *intraday price dynamics* (such as price, return and volatility during the day). The daily data don't have *any* information about this. So, for studying this intraday dynamics *high frequency data*, or in general the market microstructure (the field of scientific study of price behaviour, transaction and timing cost, and investor behaviour etc.) is used.

High frequency data in finance typically denote observations taken at a finer time scale (like - every hour, every minute, or even every second) These data become available primarily because of advances in information technology and the trend of moving toward electronic trading. The ultimate high frequency data in finance are the transaction-by-transaction data (**transactions data** or real time data) in security markets. Here, time is often measured in seconds. However, high frequency data have some unique characteristics (nonsynchronous trading, bid?ask spread, and others) that do not appear in lower frequencies such as daily, monthly, quarterly and so on. In what follows, we will that a time series approach to high frequency data (with nonsynchronous trading or bid-ask-spread) induce an *erroneous* negative autocorrelation for the returns.

Note: Transactions data are not for *free* unlike the daily data; they are available at a cost. For example, visit
https://archives.nseindia.com/content/press/Download_Real_Time_Tariff_Domestic_01042020.pdf

Nonsynchronous Trading

We should first understand that our usual financial data (on prices) are the observations on the stocks (assets) when they are *traded*. As mentioned earlier, the closing price of a stock (on a given day) is its *last traded price* on the day. In reality, stock trading does not occur in a synchronous manner; different stocks have different trading frequencies, and for a single stock, the trading intensity varies from time to time. In fact, one of the typical features is that the trading frequency of a stock is, in general, high during the opening and the closing periods, and is less during the midday on a trading day at the exchange. Below we present an excerpt of transactions of the stock of the pharmaceutical company *Merck* on April 5, 2005 at NYSE.

symbol	date	time	price	size
MRK	20050405	9:41:37	32.69	100
MRK	20050405	9:41:42	32.68	100
MRK	20050405	9:41:43	32.69	300
MRK	20050405	9:41:44	32.68	1000
MRK	20050405	9:41:48	32.69	2900
MRK	20050405	9:41:48	32.68	200
MRK	20050405	9:41:48	32.68	200
MRK	20050405	9:41:51	32.68	4200
MRK	20050405	9:41:52	32.69	1000
MRK	20050405	9:41:53	32.68	300
MRK	20050405	9:41:57	32.69	200
MRK	20050405	9:42:03	32.67	2500
MRK	20050405	9:42:04	32.69	100
MRK	20050405	9:42:05	32.69	300
MRK	20050405	9:42:15	32.68	3500
MRK	20050405	9:42:17	32.69	800
MRK	20050405	9:42:17	32.68	500
MRK	20050405	9:42:17	32.68	300
MRK	20050405	9:42:17	32.68	100
MRK	20050405	9:42:20	32.69	6400
MRK	20050405	9:42:21	32.69	200
MRK	20050405	9:42:23	32.69	3000
MRK	20050405	9:42:27	32.70	8300
MRK	20050405	9:42:29	32.70	5000

Now, we provide one of the simplest models for nonsynchronous trading data developed by Lo A. and MacKinlay, A.C. (cf. *An econometric analysis of nonsynchronous trading*, Journal of Econometrics (1990); Volume 45: pp 181 - 212).

Suppose that we observe the price of an asset at every given unit of time t. Given t, the asset may or may not be traded at t. It must be understood here that at a given t if the asset is *not* traded, then its price at t is assumed to be the one at the last trade which took place before t. Whereas, at the in-between nontrading time points, the price might or might not have changed - a fact which is not observable. Hence, if $\{P_t\}$ is the *true* price series, and $\{P_t^O\}$ is our "observed" price series, then $P_t^O = P_t$ if and only if there is trading at time t. If there is no trading at t, then $P_t^O = P_s^O$, where s is the 'largest' time before t at which the last trading took place. So, the observed price remains constant at all time points between s and t (both inclusive) as there was no trading at these time points. But, the 'true' price might have assumed different values at all these time points.

By definition, the 'return at time t' is given $Y_t = \log(\frac{P_t}{P_{t-1}})$ where P_t is the price at t. Given the above reasoning, this true return series $\{Y_t\}$ is not observable and the observable return series is given by $\{Y_t^O\}$ where

$$Y_t^O = \log(\frac{P_t^O}{P_{t-1}^O}).$$

For example, for $t = 1, 2, 3, 4, 5$, assume that trading takes place at 1, 2 and 5 and no trading at 3 and 4. Then

$$Y_2^O = \log(\frac{P_2^O}{P_1^O}) = \log(\frac{P_2}{P_1}) = Y_2,$$

$$Y_3^O = \log(\frac{P_3^O}{P_2^O}) = \log(\frac{P_2^O}{P_2^O}) = \log(\frac{P_2}{P_2}) = 0,$$

$$Y_4^O = \log(\frac{P_4^O}{P_3^O}) = \log(\frac{P_2^O}{P_2^O}) = \log(\frac{P_2}{P_2}) = 0,$$

and

$$Y_5^O = \log(\frac{P_5^O}{P_2^O}) = \log(\frac{P_5}{P_2}) = \log(\frac{P_5}{P_4}\frac{P_4}{P_3}\frac{P_3}{P_2}) = Y_3 + Y_4 + Y_5.$$

Let us assume that the return series $\{Y_t\}$ is a sequence of i.i.d. random variables with mean μ and variance σ^2. Further, suppose that the event $E =$ Trading takes place at t is such that $P(E) = \pi$ (free of t) and that E is independent of the sequence $\{Y_t\}$. Then, we can have the following representation of the observed returns $\{Y_t^O\}$.

$$
Y_t^O = \begin{cases} 0 & \text{if no trading at } t \\ Y_t & \text{if trading at } t \text{ and } t-1 \\ Y_{t-1} + Y_t & \text{if trading at } t \text{ and } t-2 \text{ but not at } t-1 \\ \\ \vdots & \vdots \\ Y_{t-(m-1)} + ... + Y_{t-1} + Y_t & \text{if trading at } t \text{ and } t-m \text{ but not in between} \\ \\ \vdots & \vdots \end{cases}
$$

That is,

$$
Y_t^O = \begin{cases} 0 & \text{with probability } 1-\pi \\ Y_t & \text{with probability } \pi^2 \\ Y_{t-1} + Y_t & \text{with probability } \pi^2(1-\pi) \\ \\ \vdots & \vdots \\ Y_{t-(m-1)} + ... + Y_{t-1} + Y_t & \text{with probability } \pi^2(1-\pi)^{m-1} \\ \\ \vdots & \vdots \end{cases}
$$

We want to show that the autocovariance function of the observed return series is negative. That is, we want to show that $\lambda(t, t-k) = Cov\ [Y_t^O, Y_{t-k}^O] < 0,\ \ \forall\ k \geq 1$. We will just show that $\lambda(t, t-1) < 0$. Towards this, we need to find the expression for the product $Y_t^O Y_{t-1}^O$. Then by finding $E[Y_t^O]$ and $E\ [Y_t^O Y_{t-1}^O]$, we will compute $\lambda(t, t-1)$. Now, using the same reasoning as above, we see that

$$
Y_t^O Y_{t-1}^O = \begin{cases} 0 & \text{with probability } (1-\pi)(1+\pi) \\ Y_t Y_{t-1} & \text{with probability } \pi^3 \\ Y_t(Y_{t-1} + Y_{t-2}) & \text{with probability } \pi^3(1-\pi) \\ \\ \vdots & \vdots \\ Y_t(Y_{t-1} + ... + Y_{t-m}) & \text{with probability } \pi^3(1-\pi)^{m-1} \\ \\ \vdots & \vdots \end{cases}
$$

From the above representations, we first see that

$$\begin{aligned}
E[Y_t^O] &= \pi^2 E[Y_t] + \pi^2 \sum_{m=2}^{\infty} (1-\pi)^{m-1} E\left[Y_t + ... + Y_{t-(m-1)}\right] \\
&= \mu\pi^2 \sum_{m=1}^{\infty} m(1-\pi)^{m-1} \\
&= \mu\pi^2 \frac{1}{\pi^2} \\
&= \mu, \qquad \forall\, t.
\end{aligned}$$

Similarly,

$$\begin{aligned}
E\left[Y_t^O Y_{t-1}^O\right] &= \pi^3 E\left[Y_t Y_{t-1}\right] + \pi^3 \sum_{m=2}^{\infty} (1-\pi)^{m-1} E\left[Y_t(Y_{t-1} + ... + Y_{t-m})\right] \\
&= \mu^2\pi^3 \sum_{m=1}^{\infty} m(1-\pi)^{m-1} \\
&= \mu^2\pi^3 \frac{1}{\pi^2} \\
&= \mu^2\pi, \qquad \forall\, t.
\end{aligned}$$

Therefore, we have that

$$\begin{aligned}
Cov\left[Y_t^O, Y_{t-1}^O\right] &= E\left[Y_t^O Y_{t-1}^O\right] - E[Y_t^O]E[Y_{t-1}^O] \\
&= \mu^2\pi - \mu^2 \\
&= -\mu^2(1-\pi) \\
&< 0, \qquad \text{as} \quad 0 < \pi < 1,
\end{aligned}$$

as desired.

Remarks:

1. Indeed one can show that

$$Cov\left[Y_t^O, Y_{t-k}^O\right] = -\mu^2(1-\pi)^k < 0, \quad \text{for each} \ \ k = 1, 2, 3, ...$$

2. The extreme cases when $\pi = 0$ or 1 are not considered here, as $\pi = 0$ refers to the situation when the asset is **never** traded and, when $\pi = 1$ refers to the other extreme situation that, whatever be the time unit t, trading happens **at each** t.

Bid-Ask Spread (BAS)

In many stock exchanges, market makers (commonly known as 'stock brokers' or simply 'brokers') play an important role in facilitating trades. They provide market liquidity by standing ready to buy or sell whenever the public wishes to sell or buy. By market liquidity, we mean the ability to buy or sell significant quantities of a security quickly, anonymously, and with little price impact. In return for providing liquidity, market makers are granted monopoly rights by the exchange to post different prices for purchases and sales of a security. They buy at the bid price P_b and sell at a higher ask price P_a. (For the public, P_b is the sale price and P_a is the purchase price.) The difference $P_a?P_b$ is called the **bid-ask spread** (BAS), which is the primary source of compensation for market makers. Typically, the bid-ask spread is small - namely, very few fundamental units of the currency (like, 1 or 2 cents or 5 or 10 paise etc.).

Below, a snapshot of the 5 rows of the *Microsoft* bid-ask quotes data is given:

Positions	Ex	MMID	Symbol	Bid	BidSize	Mode	Ask	AskSize	Seq
5/1/1997 8:17:24	T		MSFT	121.500	11	12	121.625	11	0
5/1/1997 9:00:44	T		MSFT	121.750	10	12	121.625	11	0
5/1/1997 9:07:27	T		MSFT	121.750	10	12	121.625	10	0
5/1/1997 9:16:30	T		MSFT	121.875	10	12	121.625	10	0
5/1/1997 9:20:29	T		MSFT	121.875	10	12	121.625	3	0

The existence of a bid-ask spread (BAS), although small in magnitude, has several important consequences in time series properties of asset returns. We will show that the presence of BAS introduces *Bid-Ask Bounce* effect (which means that BAS introduces a *negative lag-1* autocorrelation) in an asset return series.

A *frictionless market* is a theoretical trading environment where there are no costs or restraints associated with transactions (buying/selling). In reality, there are costs associated with transactions (which reflect in the price of the asset) and hence the real market is with friction.

Let P_t^* be the *true or fundamental* price in a frictionless market (hence P_t^* is not observable) and let P_t be the observed *market* price of an asset at time t. Then, Roll (1984) considered the following simple model for the observed market price P_t:

$$P_t = P_t^* + I_t(S/2), \qquad (*)$$

where $S = P_a - P_b$, the BAS (free of t) and $\{I_t\}$ is a sequence of i.i.d. two-valued random variables with equal probability defined by

$$I_t = \begin{cases} -1 & \text{with probability } 1/2 \\ +1 & \text{with probability } 1/2. \end{cases}$$

Here, the random variable I_t can be interpreted as an order-type indicator at time t, with $+1$ signifying a buyer-initiated transaction and ?1 as a seller-initiated transaction. Therefore, Roll's price model given in (*) above becomes

$$P_t = P_t^* + \begin{cases} -S/2 & \text{with probability } 1/2 \\ +S/2 & \text{with probability } 1/2. \end{cases}$$

Now, we define the observed *return* at time t as simply the change in the observed price; that is,

$$Return = \Delta P_t = P_t - P_{t-1}.$$

Further, it is assumed that the true price is a constant with respect to time; that is, $P_t^* \equiv P$, for all t. Then, it is easy to see that $\Delta P_t = P_t - P_{t-1} = (I_t - I_{t-1})S/2$, is a random variable such that $E[\Delta P_t] = 0$. Also, it is easy to see that

$$V[\Delta P_t] = S^2/2, \ \ \forall\, t$$

$$\lambda(t, t-1) = Cov\, [\Delta P_t, \Delta P_{t-1}] = -S^2/4, \ \ \forall\, t$$

$$\lambda(t, t-k) = Cov\, [\Delta P_t, \Delta P_{t-k}] = 0, \ \ \forall\, t, \ \ \forall\ \ k \geq 2.$$

Hence the autocorrelation function of the return series can be seen to be

$$\rho(k) = \begin{cases} -1/2 & \text{for } k = 1 \\ 0 & \text{for } k \geq 2. \end{cases}$$

The BAS thus introduces a negative lag-1 autocorrelation in the series of observed returns. This is referred to as the bid-ask bounce in the finance literature. Intuitively, the bounce can be seen as follows. Assume that the fundamental price P_t^* is equal to $(P_a + P_b)/2$. Then P_t assumes the value P_a or P_b. If the previously observed price is P_a (the higher value), then the current observed price is either unchanged or lower at P_b. Thus, ΔP_t is either 0 or $-S$. However, if the previous observed price is P_b (the lower value), then ΔP_t is either 0 or S . The negative lag-1 correlation in ΔP_t becomes apparent. However, the BAS does not introduce any serial correlation beyond lag 1.