Project 2

High-Frequency Financial Econometrics

Guilherme Salomé

September 7, 2018

The purpose of this project is to understand the integrated variance estimators, realized variance and bipower variance, and to understand why we use high-frequency data (sampled every 5 min.) but not data at the highest frequency available (tick data).

This project is Due on September 13th by 11:59 PM. You must push your local repository with the `report.pdf` file back to GitHub before the deadline.

All plots in the report must be self-contained. Self-contained means that a reader who only sees your figure (image and caption, but not the surrounding text) can understand what you are plotting. This translates to all plots having axis titles, correct units on the axis, and a caption that summarizes what is plotted.

This project makes use of stock data. Refer to this page to get access to the data (requires Duke login). You must complete the exercises below for both of your stocks using 5 min. data, unless stated otherwise.

The data files follow the .csv format and contain the prices of different assets. The name of the file represents the ticker symbol for a given stock. For example, the file `AAPL.csv` contains data for Apple's stock. Each file has 3 columns (no headers): date, time and price. The first column of a file contains the date of a given price in the `YYYYMMDD` format. For example, a date of `20070103` means January 3rd of 2007. The second column contains the time of a given price in the `HHMM` or `HHMMSS` format. For example, a time of `935` means that the price in the 3rd column was recorded at 9:35 am. If the value is `93500`, it means the price was recorded at 9:35:00 am (this is only for 5 seconds data). The last column contains the price in dollars of the stock at the given date and time.

Exercise 1

The purpose of this exercise is to import the data into Matlab, convert the first two columns into the datenum format, and verify whether the data has any issues.

A.

Create a function `loadStockData` that reads stock data in the `.csv` format, and outputs two vectors. The first vector should contain the date and time of each price observation in the datenum format. The second vector should contain the stock prices converted to log-prices.

There are many ways to read data in Matlab, choose the appropriate one. Make sure your function works for a smaller subset of the data, and then test it for an entire file.

Your function should not take more than 2 seconds to load the data and create the vectors of dates and prices.

B.

Our theory assumes our data is sampled at regular intervals of size Δ_n. Fix $T = 1$, then we observe:

$$X_0, X_{\Delta_n}, X_{2\Delta_n}, \ldots, X_{n\Delta_n}$$

This means we observe $n + 1$ log-prices per day, for each of the T days. To facilitate the notation, denote $N \equiv n + 1$.

What is N and what is T for your data? Are N and T the same for both of your stocks?

C.

Answer the following questions:

- What are the stock market hours?
- Given the market hours and that the data is sampled every 5 minutes, what value of N were you expecting?
- Does it differ from the value of N you computed in the previous question?
- What could be the reason for the difference?

D.

If there are $N \equiv n + 1$ price observations per day, we can compute n returns:

$$\Delta_i^n X \equiv X_{i\Delta_n} - X_{(i-1)\Delta_n} \text{ for } i = 1, 2, \ldots, n$$

These n returns can be computed for each day $t = 1, 2, \ldots, T$.

Note that we never take differences across days, because doing so generates the overnight jump. In high frequency finance we routinely exclude the overnight move, which is spread over a 17.5 hour period and is governed by different dynamics than the within day returns.

Create a function `getReturns` that takes log-prices, N and T, and computes the log-returns within days, for each day of the sample.

Notice that the dates for prices are different than that for returns (there is one less return per day than there are prices). The `getReturns` function should also take the dates for prices and return the correct dates to be used for plotting returns.

E.

Whenever you have a new data set you should inspect it for errors. Finding missing or weird values in a data set is common, after all any process of recording data is subject to errors.

Plot the stock prices and the geometric returns. Geometric returns are simply:

$$r_i^n \equiv 100 \times |\Delta_i^n X| \text{ for } i = 1, 2, \ldots, n$$

Do you see any outliers?

F.

Answer the following questions:

- What are stock splits?
- Why do they occur?
- At what time do they take place?
- How can you identify a stock split in your data?
- How would you correct for a stock split?
- Are there any stock splits in your data?
- Does Google Finance or Yahoo Finance correct for stock splits?
- Do stock splits affect within day returns?

Exercise 2

A.

Create a function `getRV` to compute the realized variance for the 5-min returns (RV) day-by-day.

Notice that the realized variance is computed for each day, so in order to plot you will need to adjust the dates once more. However, given the dates for the return series this can be easily done in one line.

Plot the RV series (annualized, see below) and interpret.

The stock market trades stocks using dollars, not "log-dollars". For this reason, when we plot prices we want to plot the prices in dollars (e^X). A similar situation occurs with the variance. When investors discuss expectations regarding variance the unit that is used is: volatility per year (annualized volatility). However, when we compute RV_t we are computing the realized variance for a day, so the units are in variance per day. When plotting variance measures (either RV or BV, or other estimators) you will want to convert the units from variance per day to volatility per year. Additionally, it is common to also put the volatility per year in percentage terms. To do so, apply the following transformation to RV_t:

$$100\sqrt{RV \times 252}$$

for each $t = 1, 2, \ldots, T$.

B.

Repeat the previous question but for the Bipower Variance.

C.

Plot the realized variance and the bipower variance on the same figure. Make the bipower variance plot line transparent so you can compare both. Discuss the moves of RV, BV and their differences.

D.

Read the Introduction, Section 1 and Conclusion of Huang and Tauchen (2005). Compute the daily time series of the relative contribution of jumps:

$$C_t \equiv \frac{\max\{RV_t - BV_t, 0\}}{RV_t} \text{ for } t = 1, 2, \ldots, T$$

What percent of the total RV is accounted for by the jump variation (on average)? Is the value close to the ones found by Huang and Tauchen (2005)?

Exercise 3

It is natural to question why we do not drill down to the ultra-high frequency data and use all available information. The reason is that at the ultra-high frequency the data become totally contaminated by trading friction noise, sometimes called market microstructure noise. The purpose of this exercise is to illustrate the effects of the noise empirically.

A.

Read this short article about the realized variance and the volatility signature plot.

How do you interpret the realized variance? What is the volatility signature plot? What can we learn from the volatility signature plot?

B.

Download the Ultra High-Frequency data (sampling every 5 seconds) assigned to you.

Load the data in Matlab and obtain the dates and log-prices. What is N and what is T?

C.

To create the volatility signature plot you will need to compute the within day returns for different frequencies.

$$RV_{t,J} \equiv \sum_{i=1}^{\lfloor n/J \rfloor} \left(X_{iJ\Delta_n} - X_{(i-1)J\Delta_n} \right)^2 \text{ for } J = 1, 2, \ldots, J_{max}$$

for days $t = 1, 2, \ldots, T$.

Notice that when $J = 1$, we are computing the usual RV using all samples of the day. When $J = 2$ we compute the RV as if the data was sampled every 10 seconds. For $J = 12$ we compute the RV as if the data was sampled every minute, and so on.

Let $J_{max} = 120$, that is, stop the volatility signature at the 10-min. frequency.

Compute $RV_{t,J}$ for $J = 1, 2, \ldots, J_{max}$ for the 1st day of your data. Plot the annualized volatility values against the sampling frequency in minutes.

D.

The theory predicts that (at higher frequencies)

$$|RV_{t,J_1} - RV_{t,J_2}| \approx 0$$

for values of J_1 and J_2 small enough. Do your plots suggest that the above is true for ultra-high sampling frequencies? If not, why not? (We will cover the reasons in far more detail later in the course.)

E.

If we sample at a frequency where the volatility signature function is reasonably flat, then we can be assured that the market microstructure noise is not very important. In other words, at such a frequency, the data are dominated by signal instead of noise. Do the volatility signature plot indicate that the volatility signature function is reasonably flat for sampling intervals corresponding to about 3 to 8 minutes?