# Estimating the Integrated Variance

## 1 Instructions

Project 2 is due on September 13th by 10:00 pm. This is a hard deadline, so no exceptions. You must push your local repository back to GitHub before the deadline. Your repository must contain:

- The Matlab code you used to complete the project;
- A script named `main.m` file that generates all required plots;
- A `report.pdf` file with your answers to the project questions. The report must also contain an Appendix with the code used to solve the project;
- 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.

You can obtain the repository for this project by clicking **on this link**.

## Questions

The purpose of this project is to implement the realized variance and bipower variance estimators for the integrated variance. You will also learn why we use high-frequency data (sampled every 5 min.) but not data at the highest frequency available (tick data). This project makes use of stock data. Refer to the Data page for instructions on how to download the data and which files to download (requires Duke login). You must complete all exercises for both of your stocks using the data at the 5-minutes sampling frequency, unless stated otherwise.

### Exercise 1

The purpose of this exercise is to get comfortable with the data set. You will import the data into Matlab, verify whether the data has any issues and obtain an appropriate representation of the sampling times of the data (using the datenum format).

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 used for data sampled at higher frequencies). The last column contains the price in dollars of the stock at the given date and time.

**A.**

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

There are many ways to read data in Matlab, choose the appropriate one. Your function should not take more than 2 seconds (approximately) to load the data and create the vectors of dates and prices. If it is taking too long try to simplify your approach.

Hint 1: Read the help file of `datenum` and understand how to use `DateNumber = datenum(Y,M,D,H,MN,S)`. How can you recover `Y`, `M`, `D`, `H`, `MN` and `S` from your data?

Hint 2: Consider the Matlab function `reshape` when building the outputs of the `load_stock` matrix. Can you store the prices in a vector or is it better to store in a matrix? Why? What extra information does a matrix provide you with?

**B.**

Our theory assumes our data is sampled at regular intervals of size $\Delta_n$.

$$X_0, X_{\Delta_n}, X_{2\Delta_n}, \ldots, X_{n\Delta_n}, X_{n\Delta_n+1}, \cdots, X_{2n\Delta_n}, \cdots, X_{Tn\Delta_n}$$

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, define $N \equiv n + 1$.

What is $N$ and what is $T$ for your data? Are $N$ and $T$ the same for both of your stocks? How can you find $N$ and $T$ by code?

**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 compute returns across different days. Doing so would generate an 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 that 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). You might want your function to also 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 returns. Returns should be plotted in percentage:

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

Do you see any outliers in your data?

**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

The purpose of this exercise is to implement estimators for the integrated variance.

**A.**

Create a function `realized_var` to compute the realized variance (RV) for the 5-min returns day-by-day. That is, for each day $t = 1, 2, \cdots, T$, you will use that day's returns to compute the realized variance estimator, which estimates that day's integrated variance. Notice that the realized variance is computed for each day, so in order to plot the estimates you will need to adjust the dates once again.

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 a stock's variance, the unit that is assumed is: volatility per year (annualized volatility). However, when we compute $\text{RV}_t$ we are computing the realized variance for a day, so the units are in variance per day. When plotting measures of variance you will want to convert the units from variance per day to volatility per year (standard deviation per year). Additionally, it is common to also put the volatility per year in percentage terms. To do so, apply the following transformation to $\text{RV}_t$:

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

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

Plot the realized variance estimates and interpret.

**B.**

Repeat the previous question but for the Bipower Variance. That is, create a function `bipower_var` that computes the bipower variance (BV) estimator for the 5-min returns day-by-day. Plot it and interpret the estimates.

**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. Contrast the integrated variance estimates of RV and BV. What are the similarities and 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:

$$C_t \equiv \frac{\max\left\{\text{RV}_t - \text{BV}_t, 0\right\}}{\text{RV}_t} \text{ for } t = 1, 2, \ldots, T$$

What percent of the total realized variance 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. Why stop at 5-minutes when we could go down to tick data? The reason is that as the sampling frequency increases, the data starts becoming affected 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 data sampled at the 5-seconds frequency (only the stocks that were assigned to you). Modify the function `load_stock` to also handle data sampled at the 5-seconds frequency. Load the data in Matlab and obtain the appropriate dates and log-prices. What is $N$ and $T$ in this case?

**C.**

To create the volatility signature plot you will need to compute the within day returns for different frequencies. Then, use these returns to compute the realized variance estimator. We can do both things in one step:

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

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

Notice that when $J = 1$, we are computing the usual RV using all samples of the day. That is, we are using the data sampled every 5 seconds. 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 $\text{RV}_{t,J}$ for $J = 1, 2, \ldots, J_{max}$ for the 1st day of your data. Plot the volatility values (use the correct units) against the sampling frequency in minutes. Interpret the plot.

**D.**

The signature plot based on a single day is quite noisy. We can obtain a less noisy volatility signature plot by averaging the results across several days. That is, compute:

$$\text{RV}_J = \frac{1}{T} \sum_{t=1}^{T} \text{RV}_{t,J} \text{ for } J = 1, 2, \cdots, J_{max}$$

Compute this average volatility signature and plot it against the sampling frequency in minutes. Interpret the plot.

**E.**

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.)

**F.**

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?

**G. (Optional, PhD required)**

A natural question is whether the microstructure noise affecting the high-frequency prices changed over time. That is, are estimates using ultra-high-frequency data in 2007 more or less affected by microstructure noise than estimates using ultra-high-frequency data in 2016? Download all 5-seconds data files for the IBM stock (from 2007 to 2017). Create the average volatility signature plot for every year. You might need to center the values (subtract the mean) for the signature of each year, so that you can plot the signature for all years in the same figure (otherwise you may run into scaling issues). Interpret the plot. Do the volatility signature plot indicate that the volatility signature function is reasonably flat for sampling intervals corresponding to about 3 to 8 minutes? Do the signature plots change throughout the years?

## Exercise 4 (Optional for all)

The purpose of this exercise is to find the asymptotic distribution of the realized variance in a simpler context, but being rigorous on the use of the theorems.

Consider the case where the log-prices follow a Gaussian diffusion:

$$dX_t = \sigma_t dW_t$$

In our lectures we used $\sqrt{c_t}$ above instead of $\sigma_t$, but they are equivalent, just define $\sigma_t \equiv \sqrt{c_t}$. We will tackle the case where $\sigma : [0,1] \mapsto \mathbb{R}_+$ is a continuous and deterministic function (this is one step above the constant coefficients model, but one step below the model with stochastic variance). Also assume that $\sigma^2$ is Lipschitz.

Let $\{Z_i\}_{i\in\mathbb{N}}$ be a sequence of i.i.d. standard Normal random variables. Then, we can write the returns as:

$$r_{n,i} = \sqrt{\Delta_n}\sigma_{i/n}Z_i$$

for $i = 1, 2, \cdots, n$.

Show that:

$$\sqrt{n}\left(\sum_{i=1}^{n} r_{n,i}^2 - \int_0^1 \sigma_s^2 ds\right) \xrightarrow{d} \mathcal{N}(0, \Sigma)$$

You will also need to find what is $\Sigma$ above.

**The Central Limit Theorem**

The starting point for this problem is finding the appropriate central limit theorem (CLT). Consider a simplified version of the Lindeberg's CLT:

**Theorem 1** *Consider an array* $\{X_{n,i}\}$ *that takes values on* $\mathbb{R}$ *and is independent with zero mean. Suppose that* $\sum_{i=1}^{n} \mathbb{E}\left[X_{n,i}^2\right] \to \Sigma$ *as* $n \to \infty$*. Suppose that* $\forall \varepsilon > 0$*:*

$$\lim_{n\to\infty} \sum_{i=1}^{n} \mathbb{E}\left[\|X_{n,i}\|^2 \mathbb{1}_{\{\|X_{n,i}\|>\varepsilon\}}\right] = 0$$

*This is known as Lindberg's condition.*

*Then:* $\sum_{i=1}^{n} X_{n,i} \xrightarrow{d} \mathcal{N}(0, \Sigma)$.

The theorem suggests we should try to find an appropriate $X_{n,i}$, since it would lead to the convergence result as well as the value of $\Sigma$.

### A. Split the Problem

Show that $\sum_{i=1}^{n} r_{n,i}^2 - \int_0^1 \sigma_s^2 ds$ can be written as a sum of two terms:

$$\text{First term} : \sum_{i=1}^{n} \frac{1}{n} \sigma_{i/n}^2 (Z_i^2 - 1)$$
$$\text{Second term} : \sum_{i=1}^{n} \frac{1}{n} \sigma_{i/n}^2 - \int_0^1 \sigma_s^2 ds$$

### B. Convergence of the Second Term

If we let $n \to \infty$, what will the second term converge to?

- B.1. Show that:

$$\left| \sum_{i=1}^{n} \frac{1}{n} \sigma_{i/n}^2 - \int_0^1 \sigma_s^2 ds \right| \leq \sum_{i=1}^{n} \int_{(i-1)\frac{1}{n}}^{i\frac{1}{n}} \left| \sigma_{i/n}^2 - \sigma_s^2 \right| ds$$

- B.2. Let $K$ be the Lipschitz constant for $\sigma^2$, then show that:

$$\left| \sum_{i=1}^{n} \frac{1}{n} \sigma_{i/n}^2 - \int_0^1 \sigma_s^2 ds \right| \leq \frac{K}{n}$$

- B.3. What is the limit of:

$$\sqrt{n} \left| \sum_{i=1}^{n} \frac{1}{n} \sigma_{i/n}^2 - \int_0^1 \sigma_s^2 ds \right|$$

  What does it imply about the second term?

### C. Convergence of the first term

Now, we focus on the first term and see how we can apply Lindberg's CLT.

Define $X_{n,i} \equiv \frac{1}{n} \sigma_{i/n}^2 (Z_i^2 - 1)$. This will be our triangular array.

- C.1. Compute $\mathbb{E}[X_{n,i}]$. Does it have zero mean? Why is it important to have zero mean?

- C.2. Compute $\mathbb{E}\left[X_{n,i}^2\right]$. Remember that $\sigma$ is Riemann integrable.

- C.3. Verify Lindberg's condition.
  - C.3.1 Use Holder's inequality to bound $\mathbb{E}\Big[\Big\|X_{n,i}^2\Big\| \mathbb{1}_{\{\|X_{n,i}\|>\varepsilon\}}\Big]$ by the multiplication of two terms.
  - C.3.2. Use Markov's inequality to further bound one of the two terms.
  - C.3.3. Show that:

$$\mathbb{E}\Big[\Big\|X_{n,i}^2\Big\| \mathbb{1}_{\{\|X_{n,i}\|>\varepsilon\}}\Big] \leq \frac{1}{\varepsilon^{\frac{2p}{q}}} \frac{1}{n^{2p}} \sigma_{i/n}^{4p} \mathbb{E}\Big[\Big(Z_i^2 - 1\Big)^{2p}\Big]$$

    Then substitute $p = q = 2$.
  - C.3.4. Is $\sigma$ bounded? If so, let $M$ be its upper bound. Then, show that:

$$\sum_{i=1}^{n} \mathbb{E}\Big[\Big\|X_{n,i}^2\Big\| \mathbb{1}_{\{\|X_{n,i}\|>\varepsilon\}}\Big] \leq \sum_{i=1}^{n} \frac{1}{\varepsilon^{\frac{2p}{q}}} \frac{1}{n^{2p}} \sigma_{i/n}^{4p} \mathbb{E}\Big[\Big(Z_i^2 - 1\Big)^{2p}\Big] \leq \frac{1}{n^3}\left(\frac{M^8 \cdot 60}{\varepsilon^2}\right)$$

  - C.3.5. Conclude that Lindberg's condition is satisfied
- C.4. Use Lindberg's CLT to obtain the convergence result for the $X_{n,i}$ we defined above.

**D. Combining the Convergence of Both Terms**

Show that:

$$\sqrt{n}\left(\sum_{i=1}^{n} r_{n,i}^2 - \int_0^1 \sigma_s^2 ds\right)$$

can be split into the two terms we analyzed before. Apply the convergence results for each to get the main result. Congratulations if you made it this far!