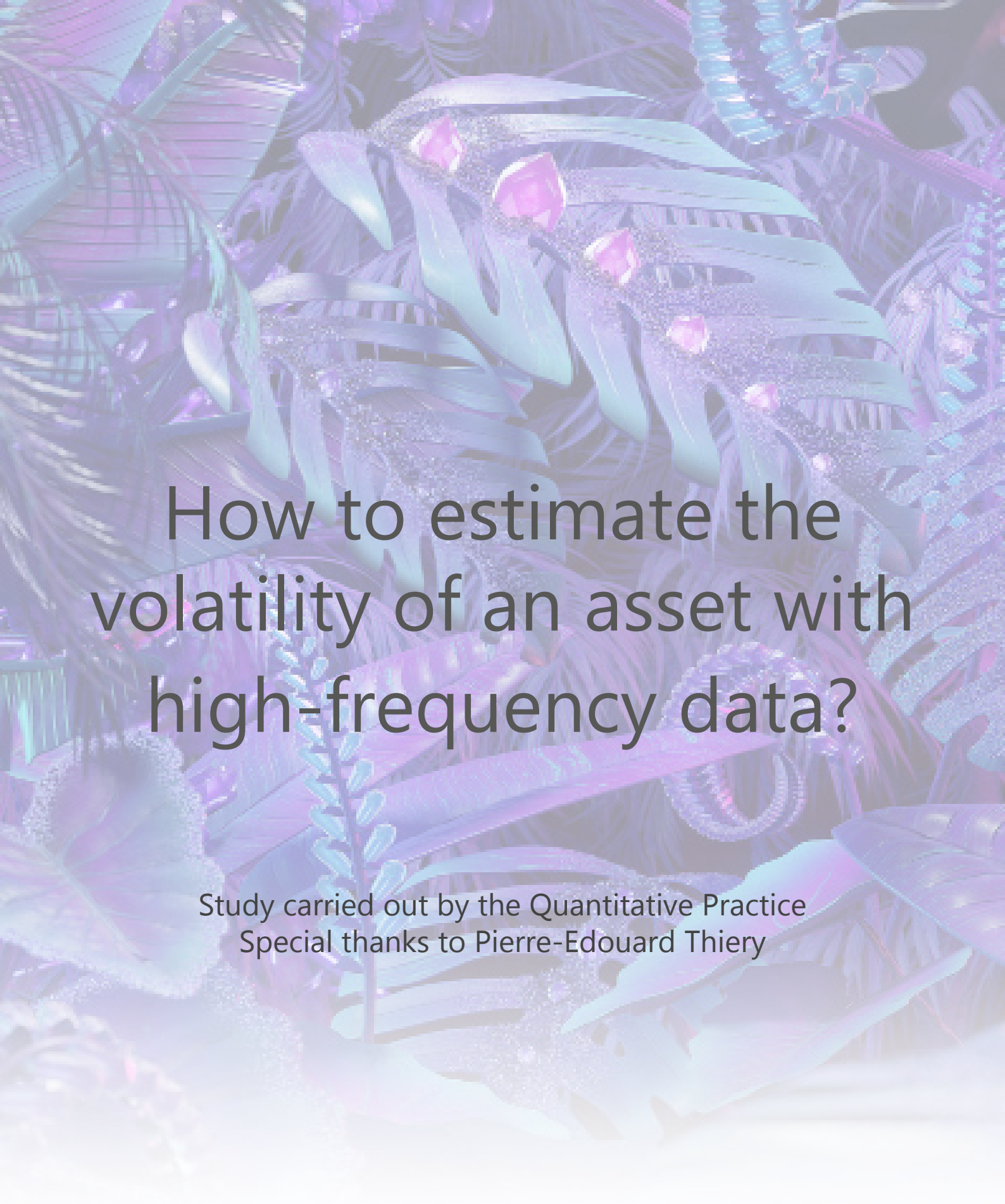

How to estimate the volatility of an asset with high-frequency data?

Study carried out by the Quantitative Practice
Special thanks to Pierre-Edouard Thiery

awalee

SEPTEMBRE 2020

Summary

Introduction

Estimating the volatility of an asset is of the utmost importance in financial mathematics, insofar as volatility plays a key role in almost every model: the assumption of a constant volatility, used in the famous Black-Scholes formula, is too simple to really take into account the realities of financial markets; more elaborate approaches are therefore required.

Nonetheless volatility estimation is not as easy as it may seem from a purely mathematical point of view. When using the widespread Itô modeling for the price of an asset, it is well-known that the cumulative volatility can be estimated thanks to the sum of squared returns over a given period of time. Even though this approach turns out to be sufficient in most cases, it may fail in certain contexts, for instance when working with high-frequency data.

High-frequency volatility estimation deserves indeed a specific treatment due to the peculiarities of what is called the "market microstructure". In this paper, after setting forth a few reminders regarding the common framework for volatility estimation, we will focus on the difficulties which arise when dealing with high-frequency data. It is then necessary to find new ways to estimate the high-frequency volatility of an asset. To solve this issue, we will present and implement several estimators, from the least to the most precise.

1 From The Classic Modeling To The Signature Plot Reality

First, we denote (S_t) the price process of a security, and we define the log price process $X_t = \ln(S_t)$. We make the following assumption on (X_t):

Definition 1 (Itô Modeling)
The log price process (X_t) is assumed to be an Itô process:

$$dX_t = \mu_t dt + \sigma_t dB_t$$

where (B_t) is a standard Brownian motion. The cumulative volatility of our security over a time period $[0, T]$ is given by

$$\int_0^T \sigma_t^2 dt$$

As of now, we consider that the upper bound of the time interval, i.e. T, is known, so we only work on $[0, T]$. **Our purpose is to estimate the cumulative volatility** $\int_0^T \sigma_t^2 dt$.

Under the required assumptions, the theory of stochastic processes states that [1]:

$$\sum_{i=0}^{N-1} \left(X_{t_{i+1}} - X_{t_i}\right)^2 \xrightarrow{N\to\infty} \int_0^T \sigma_t^2 dt$$

where the dates $(t_i)_{i=0..N}$ define a grid $\mathbf{G}$ such that:

$$\mathbf{G} = \{0 = t_0 < t_1 < \ldots < t_N = T\}$$

In our estimation procedure, the dates (t_i) are the ones where we observe a value for the process X. They are called the sampling dates.

We know that the sum of the squared returns converges to the cumulative volatility over $[0, T]$ when the sampling frequency increases.

However this approach is a bit simplistic. If it works in common cases, it is because the process X is not sampled too frequently. Indeed, when using empirical data, we observe that, when the sampling frequency is too high, the sum of the squared returns, instead of converging toward the cumulative volatility, increases. This behavior is only one of the many peculiarities which appear when dealing with high-frequency data. Such peculiarities are often referred to as the "stylized facts" [2] of high-frequency.

In order to properly present this empirical phenomenon, it is important to define what the signature plot is:

Definition 2 (Signature Plot)
If we denote (X_t) a process, its signature plot on the interval $[0, T]$ with step τ ($T = N \times \tau$) is defined as:

$$V_T(\tau) = \frac{1}{T} \sum_{n=0}^{N-1} \left|X_{(n+1)\tau} - X_{n\tau}\right|^2$$

It is merely the realized volatility over a time period $[0, T]$, using a step τ.

When using empirical data for the process (X_t), if we draw the signature plot by considering the function $\tau \longrightarrow V_T(\tau)$, we observe that the signature plot is a decreasing function of τ. This is called the signature plot effect.

The consequence of this empirical reality is that the aforementioned volatility estimator is not robust when working with high-frequency data. Mathematically, this means that the log return process is not a true semi-martingale in real life.

This phenomenon is due to what is called the "market microstructure". The term "market microstructure" refers to the phenomena which are observable only when working with high-frequency data, and which therefore challenge low-frequency modelings.

A simple approach to deal with the market microstructure consists in tweaking the modeling of the log return process. Since real data cannot be viewed as a true semi-martingale, we assume we actually observe a process (Y_t) which is derived from the true, invisible, process X:

Definition 3 (Additive Modeling)
The observed log price process is denoted (Y_t); it is made of two components:

$$Y_t = X_t + \epsilon_t$$

(X_t) is the true log price process, which is not visible, i.e. it is not possible to observe the process (X_t) at a given time. (ϵ_t) is an independent noise around the true returns. We have $\mathbb{E}[\epsilon_t] = 0$ and we denote $\mathbb{E}[\epsilon_t] = \epsilon^2$

Such a modeling accounts for the signature plot effect. To see that, let us define:

Definition 4 (Quadratic Variation)
The quadratic variation of a process U, which is observed on a given grid denoted $\mathbf{H}$

$$\mathbf{H} = \{0 = h_0 < h_1 < \ldots < h_n = T\}$$

is denoted $[U, U]^{\mathbf{H}}$*, and is defined by:*

$$[U, U]^{\mathbf{H}} = \sum_{k=0}^{n-1} \left(U_{h_{k+1}} - U_{h_k}\right)^2$$

Thanks to this definition, it is straightforward to define what the quadratic variation between two processes U and V is:

$$[U, V]^{\mathbf{H}} = \sum_{k=0}^{n-1} \left(U_{h_{k+1}} - U_{h_k}\right)\left(V_{h_{k+1}} - V_{h_k}\right)$$

So, by assuming that the observed data are actually noisy, we have:

$$[Y, Y]^{\mathbf{G}} = [X, X]^{\mathbf{G}} + 2[X, \epsilon]^{\mathbf{G}} + [\epsilon, \epsilon]^{\mathbf{G}}$$

which leads to:

$$\mathbb{E}([Y, Y]^{\mathbf{G}} | X) = [X, X]^{\mathbf{G}} + 2n\epsilon^2$$

This equation shows that, when the sampling frequency increases, meaning that n increases, the estimation of the high-frequency volatility using the mere quadratic variation of the observed data will not converge. The factor $2n\epsilon^2$ will cause the estimated values to soar when the number of available data within $[0, T]$ becomes too important.

For the sake of precision and the reader's interest, it is even possible to prove that [3]:

$$\frac{1}{\sqrt{n}}\left(\mathbb{E}([Y, Y]^{\mathbf{G}} - 2n\epsilon^2\right) \xrightarrow[n\to\infty]{L} 2\sqrt{\xi}\mathcal{N}(0, 1)$$

where $\xi = \mathbb{E}\left(\epsilon_t^4\right)$ and $\mathcal{N}(0, 1)$ a standard normal law.

To illustrate this, we have implemented a simple case. We consider that the upper bound for our time interval is $T = 1$. Given m an integer, we define the grid $\mathbf{G}^m$:

$$\mathbf{G}^m = \left\{\frac{k}{2^m},\ k = 0..2^m\right\}$$

This way we can assume that, when m is big enough, we work with high-frequency data. The process X is merely a Brownian path multiplied by a constant value; to generate the observed data Y at the dates given by the grid $\mathbf{G}^m$:

$$Y_{\frac{k}{2^m}} = X_{\frac{k}{2^m}} + \epsilon_{\frac{k}{2^m}}$$

we only simulate independent normal variables: $\epsilon_{\frac{k}{2^m}} \simeq \mathcal{N}(0, 0.00025)$.

We then estimate the cumulative volatility over $[0, 1]$ thanks to $[Y, Y]^{\mathbf{G}^m}$, for m from 0 to 24. Figure 1 displays the estimated values for the cumulative volatility:

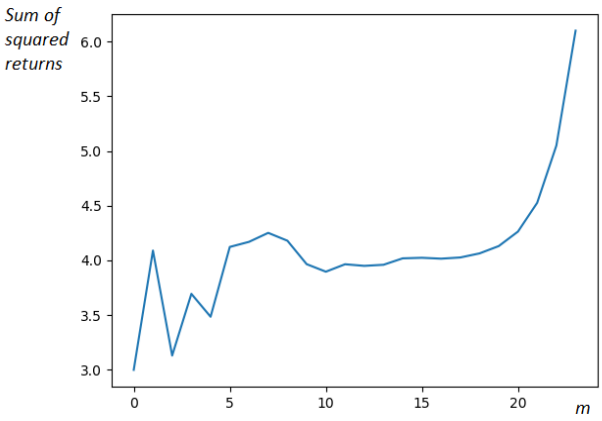

Figure 1: Cumulative volatility estimation as a function of m

We see that, when m is too small, the estimation is unprecise due to the small amount of data. Then, when m increases, the estimation converges towards the cumulative value, in our case 4. But when the sampling frequency is too high, above 20 in our case, the estimated value starts to increase.

The purpose of this paper is to find a way to properly estimate the cumulative volatility when working with high frequency data. We cannot consider the classic estimator based on the quadratic variation because of the market microstructure, which can, as a first approximation, be assimilated to a noise signal which impacts the observed data.

2 How To Devise A True And Unbiased Estimator Of The High-Frequency Volatility?

Since the classic estimator $[Y, Y]^{\mathbf{G}}$ is no longer a reliable estimator of the volatility in a high-frequency context, a first idea consists in disregarding some of our observations in order to work with lower-frequency data.

• The subsampled estimator

To do this, we consider a sub-grid $\mathbf{G}^s \subset \mathbf{G}$. If $\mathbf{G}$ contains n elements, we denote n^s the number of elements within $\mathbf{G}^s$. Of course $n^s < n$. For instance the data contained in $\mathbf{G}^s$ are sampled every five minutes, instead of every second in $\mathbf{G}$.

It is then fairly intuitive to define the following "subsampled" estimator.

Definition 5 (Subsampled Estimator)
The subsampled estimator is defined as follows:

$$[Y,Y]^{\mathbf{G}^s}$$

So it is only the quadratic version computed on a subgrid of our data.

It is important to keep in mind that the observed data are still polluted by a noise process, whose purpose is to take into account the market microstructure. Nonetheless the subsampled estimator is better than the classic one insofar as it is now as if we work with lower-frequency data.

Nonetheless it is pivotal to remark that this does not mean that the subsampled estimator is without shortcomings. Indeed, similarly to the results we have for the classic estimator, it is possible to show [3] that:

$$[Y,Y]^{\mathbf{G}^s} \xrightarrow[n\to\infty]{L} \int_0^T \sigma_t^2 dt + 2n^s\epsilon^2 + \mathcal{N}(0,1)\sqrt{4n^s\xi + \frac{2T}{n^s}\int_0^T \sigma_t^4 dt}$$

So, at low-frequency, we can assume that the subsampled estimator is a good estimator, albeit imperfect, of the cumulative volatility because n^s compared to ϵ^2 and ξ is such that $n^s\epsilon^2$ and $n^s\xi$ are almost equal to zero.

However the estimator is not very satisfactory inasmuch as we only jettison most of the available data. It would be much better to use all the available data.

It is possible to do so while staying true to the philosophy of the subsampled estimator. We devise a new estimator based on both subsampling and averaging. We will refer to it as the subsampled and averaged estimator.

• The subsampled and averaged estimator

Definition 6 (Subsampled and Averaged Estimator)
We divide the grid $\mathbf{G}$ *into* K *subgrids:*

$$\mathbf{G} = \bigcup_{k=1}^{K} \mathbf{G}^k$$

such that $\mathbf{G}^k \cap \mathbf{G}^q = \emptyset$ *when* $k \neq q$. *We write* $\bar{n}$ *the average size of the grids:*

$$\bar{n} = \frac{1}{K}\sum_{k=1}^{K} Card\left(\mathbf{G}^k\right)$$

The subsampled and averaged estimator, denoted $[Y,Y]^{avg}$, *is then defined as follows:*

$$[Y,Y]^{avg} = \frac{1}{K}\sum_{k=1}^{K}[Y,Y]^{\mathbf{G}^k}$$

Let us assume that the data in $\mathbf{G}$ are sampled every second. A simple way to define the subgrids of $\mathbf{G}$ consists in sampling the data, for instance, every five minutes. The number of grids is then given by the number of seconds which are contained within five minutes, i.e. $K = 300$.

The subsampled and averaged estimator improves on the subsampled estimator insofar as it is based on all the data which are available. Besides, it is possible to show that:

$$[Y,Y]^{avg} \xrightarrow[n\to\infty]{L} \int_0^T \sigma_t^2 dt + 2\bar{n}\epsilon^2 + \sqrt{4\frac{\bar{n}}{K}E + \frac{4T}{3\bar{n}}\int_0^T \sigma_t^4 dt}\,\mathcal{N}(0,1)$$

We see that, when using the subsampled and averaged estimator, the cumulative volatility is no longer noised by a factor proportionate to n, the total number of observations, but by a factor proportionate to $\bar{n}$, which is smaller by definition. Nonetheless this estimator is still biased; this leads us to consider a fourth and final volatility estimator.

Indeed we would like to rule out the factor $2\bar{n}\epsilon^2$ which appears with the subsampled and averaged estimator. This can be done by combining the subsampled and averaged estimator and the classic one, whose bias factor is $2n\epsilon^2$.

• The unbiased subsampled and averaged estimator

Definition 7 (Unbiased Estimator)
The unbiased estimator, denoted $[Y,Y]^{unb}$, *of the cumulative volatility is obtained by combining the subsampled and averaged estimator and the classic estimator together:*

$$[Y,Y]^{unb} = [Y,Y]^{avg} - \frac{\bar{n}}{N}[Y,Y]^{\mathbf{G}}$$

Once again, it is then possible to prove the following result regarding this final estimator: if K is chosen such that $K = cn^{\frac{2}{3}}$

$$[Y,Y]^{unb} \cong \int_0^T \sigma_t^2 dt + \frac{1}{n^{\frac{1}{6}}}\sqrt{\frac{4Tc}{3}\int_0^T \sigma_t^4 dt + \frac{8\epsilon^4}{c^2}}\,\mathcal{N}(0,1)$$

3 Implementation Of The Final Estimator On Simulated Data

To illustrate the advantages of the final estimator, it has been implemented in the context of simulated data. In this section we display our results.

Framework (Simulated Data)
First, we remind the reader that the process X *is not visible in real life. In our case, the process* X *is merely a Brownian motion multiplied by the constant value* 2. *We simulate the process* X. *We set* $T = 1$, *so we work on the time interval* $[0,1]$.

Since X *is only a Brownian motion, we can rewrite definition 1:*

$$dX_t = 2dB_t$$

meaning that σ_t *is a constant process whose value is equal to* 2. *Thus the cumulative volatility is theoretically equal to:*

$$\int_0^T \sigma_t^2 dt = \int_0^1 2^2 = 4$$

When it comes to the observed data Y, we assume we have observations at the points $\frac{k}{2^m}$, where m is equal to 21. So, with our notations:

$$\mathbf{G} = \left\{ \frac{k}{2^m},\ k = 0..2^m \right\}$$

To create the observed data Y at the points in $\mathbf{G}$, we only add a normal noise. For k in $0..2^m$:

$$Y_{\frac{k}{2^m}} = X_{\frac{k}{2^m}} + \epsilon_{\frac{k}{2^m}}$$

where the random variables $\epsilon_{\frac{k}{2^m}}$ are just independent normal variables, with mean 0 and standard deviation $\epsilon = 0.00025$.

We simulated 100 different paths, and each time we estimated the cumulative volatility with both the classic and the final estimator.

Figure 2 displays the 100 estimated values for the cumulative volatility when using the first and naive estimator. We see that the estimation is not satisfactory, since all the values are within the interval $[4.25, 4.28]$, whereas the cumulative value is theoretically equal to 4. This illustrates the signature plot effect. The average estimated value, computed on the 100 simulated paths, is equal to 4.2626152773485515.

In the first part of our paper, we mentioned that

$$\mathbb{E}([Y,Y]^{\mathbf{G}}\,|X) = [X,X]^{\mathbf{G}} + 2n\epsilon^2$$

With $\epsilon = 0.00025$ and $n = 2^{21}$, it is easy to check that

$$2n\epsilon^2 \approx 0.26$$

which is very close to the distance between the average estimated value and the theoretical one.

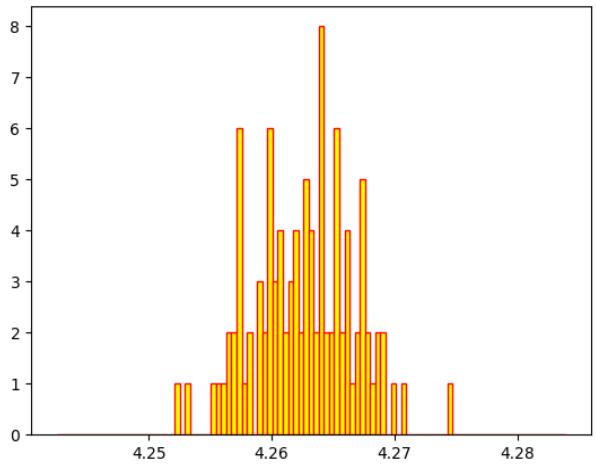

Figure 2: Estimated values for the cumulative volatility when using the classic estimator

When it comes to the final estimator, the subsampling has been carried out using 2^{21-8} subgrids containing 2^8 points. Indeed we used the following subgrids, for $k \in \left[0, 2^{21-8} - 1\right]$:

$$\mathbf{G}^k = \left\{ \left(\frac{q}{2^{21-8}} + k \right) \frac{1}{2^{21}},\ q = 0 \dots 2^8 - 1 \right\}$$

Figure 3 displays the 100 estimated values for the cumulative volatility when using the unbiased subsampled and averaged estimator. The estimation is much better, the values being approximately centered around 4, illustrating that this estimator is unbiased. If we compute the averaged estimated value over the 100 paths, we find indeed 4.023208490333749, which is much closer to 4.

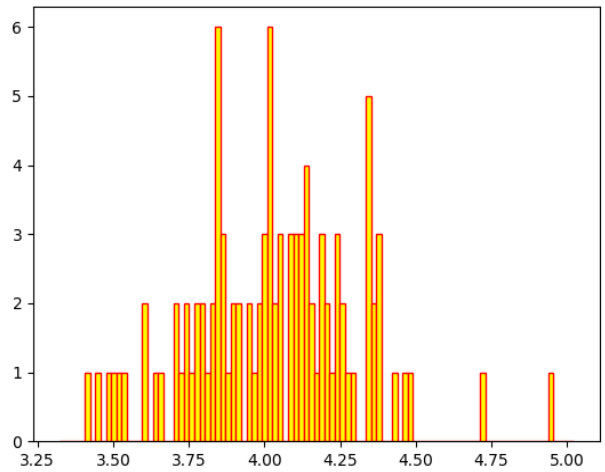

Figure 3: Estimated values for the cumulative volatility when using the unbiased subsampled and averaged estimator

Conclusion

When it comes to the practical implementation of mathematical results in finance, one should be particularly careful when dealing with high-frequency data, otherwise one may face some unexpected outcomes. This is due to the market microstructure. The peculiarities of the high-frequency world are often referred to as the "stylized facts" of high-frequency. In this paper we have mainly insisted on the signature plot effect, which is probably the most famous of the various high-frequency stylized facts.

One of the consequences of the signature plot effect is that it is no longer possible to estimate the cumulative volatility of an asset thanks to a mere quadratic variation. It is then important to develop a mathematical model which manage to reproduce this empirical effect. One of the simplest models is the additive one; it assumes that the true data are invisible, and that the observer only sees a noisy signal. Within this framework, it is possible to develop ever more precise estimators in order to estimate the cumulative volatility.

We have purposefully kept our approach rather simple, but there are several aspects which could lead to further investigation. For instance, even though the additive model is fairly simple, it fails in perfectly modeling the signature plot effect. Indeed, in reality we observe that, when the sampling frequency increases, the signature plot increases and converges. In the additive model, when the sampling frequency increases, the signature plot diverges towards the infinity. Furthermore, when considering subsampling estimators, it is also possible to optimize the subsampling param-

eters, i.e. the number of subgrids we choose to consider, in order to improve the quality of the estimator. All those questions could be addressed in order to better take into account the singular reality of the high-frequency world.

References

[1] Bougerol P. *Calcul Stochastique des martingales continues*. Université Pierre et Marie Curie, 2015.

[2] Bacry E., Delattre S., Hoffmann M., and Muzy J F. Modelling miscrostructure noise with mutually exciting point processes. *Quantitative Finance*, pages 13, 65–67, feb 2013.

[3] Zhang L., Mykland Per A., and Ait-Sahalia Y. A tale of two time scale: Determining interated volatility with noisy high-frequency data. *American Statistical Association*, pages 1407–1411, Dec 2005.

A propos d'Awalee

Cabinet de conseil indépendant spécialiste du secteur de la Finance.

Nous sommes nés en 2009 en pleine crise financière. Cette période complexe nous a conduits à une conclusion simple : face aux exigences accrues et à la nécessité de faire preuve de souplesse, nous nous devions d'aider nos clients à se concentrer sur l'essentiel, à savoir leur performance.

Pour accomplir cette mission, nous nous appuyons sur trois ingrédients : habileté technique, savoir-faire fonctionnel et innovation.

Ceci au service d'une ambition : dompter la complexité pour simplifier la vie de nos clients.

«*Run the bank*» avec Awalee !

Contactez-nous

Ronald LOMAS
Partner
rlomas@awaleeconsulting.com
06 62 49 05 97

awalee

est une marque de

canopee group