VOLATILITY ESTIMATION USING HIGH-FREQUENCY FINANCIAL DATA

Research Report

April 10, 2024

Badiss Ben Abdallah, Gaetan Ecrepont, Samson Gourevitch

ÉCOLE POLYTECHNIQUE

IP PARIS

CONTENTS

Abstract

In this paper we will tackle the issue of volatility estimation when dealing with high-frequency financial data. We will use EUR/USD tick-by-tick forex data to test our models. We will first discuss stylized facts of high-frequency forex data. Then we will look into non-naive high-frequency integrated variance estimators. Finally we will show how point processes can be used to model high-frequency price dynamics.

INTRODUCTION

The increasing availability of quality high-frequency financial data has enabled better modeling of intraday asset prices dynamics. Although diffusion models are rather accurate at daily timescales, they fail to reproduce some of the stylized facts exhibited by high-frequency price data such as the divergence of the signature plot at very short timescales [3] or the Epps effect [4].

In particular, estimating the volatility of an asset's price process at very high frequency is hard because of the discrete nature of price variations and the impact of microstructure noise [1].

We will first review stylized facts exhibited by high-frequency financial data in Section 1. Then we will study different estimators of variance at short timescales in Section 2. Finally, in Section 3 we will use point processes to model price dynamics.

The code used to produce our results and highlight other interesting statistical facts can be found at *github.com/gaetanX21/hf-volatility*

1
STYLIZED FACTS IN HIGH-FREQUENCY FOREX DATA

In this section, we will rely on tick data from the EUR/USD index dated March 4th, 2024.

1.1 Strong first order negative autocorrelation

Autocorrelation, or serial correlation, indicates the correlation between successive observations within a time series. A positive first-order return autocorrelation indicates a tendency for a return to align with the direction of the previous retur, while a negative first-order autocorrelation suggest a reversal.

Unlike in liquid markets, where significant autocorrelation in price movements is typically absent due to market efficiency, high-frequency data, often exhibit a pronounced first-order negative autocorrelation. This can be particularly observe at shorter time horizons in foreign exchange markets. This phenomenon is attributed to the discrete nature of transaction prices and bid-ask spreads. Detecting and removing such unwanted temporal dependence is crucial in further analysis.

Table 1: Autocorrelation Values of Log Returns for EUR/USD index March 4th 2024

Time Lag	**1 min**	**5 min**	**7 min**	**10 min**
Autocorrelation	-0.0512	-0.0245	0.0057	-0.0048

1.2 Intraday seasonality

The U-shaped pattern observed in various exchange-traded instruments, ranging from stocks to commodities, is a common occurrence throughout the trading day. Typically, volatility tends to spike shortly after opening and then gradually decreases as the day progresses, only to rise again just before the market closes. Similar U-shaped patterns could be observed in return rates , bid-ask spreads and other measures.

Market Open Fuels Volatility: Indeed, at the start of the trading day, market players have already assimilated the information available overnight. Upon market opening, they adjust their positions and simultaneously place orders to either establish new positions or close out existing ones. This concentrated order flow leads to heightened volatility at the market open.

"Wait and See" policy: As the trading session progresses, activity tends to moderate, with a potential lull around midday. However, volatility picks up again as the closing hour approaches. Investors make final adjustments based on the day's developments, causing another spike in price fluctuations.

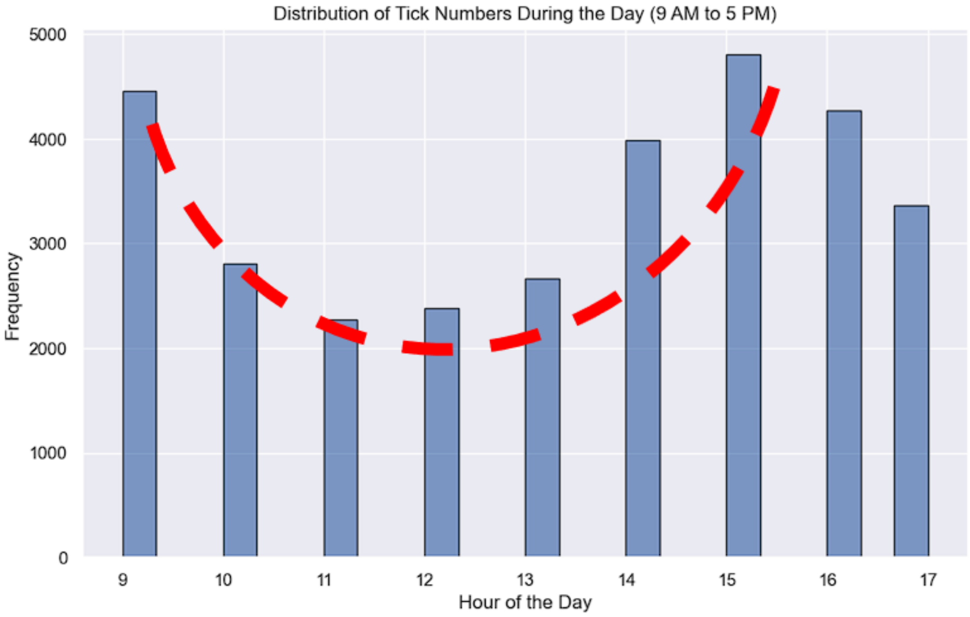

Figure 1: U-Shape during trade day for EUR/USD March 4th 2024

We can observe the famous "lunch break" effect with the lowest reported activity being between 11 am and 1 pm.

1.3 Distribution properties

In financial analysis, the widespread use of the normal distribution is prevalent in risk and performance evaluation. This distribution, approximated through the central limit theorem, portrays returns with events deviating from μ by more than 5 σ as extremely rare. Yet in constrast to traditional modeling, high frequency data often display fat-tailed return distribution across various financial instruments.

These can be explained by several factors. The divergence between these distribution actually relies on the wrongness of central limit theorem's key assumptions. Traders' decisions seem to be more influenced by extreme prices in the sort term, leading to more frequent realization of rare values. Moreover, real-world spikes such as political changes defy mathematical randomness and contribute to the prevalence of extreme price changes.

The examination of these fat tails has become increasingly significant, particularly in the realm of financial asset risk management.

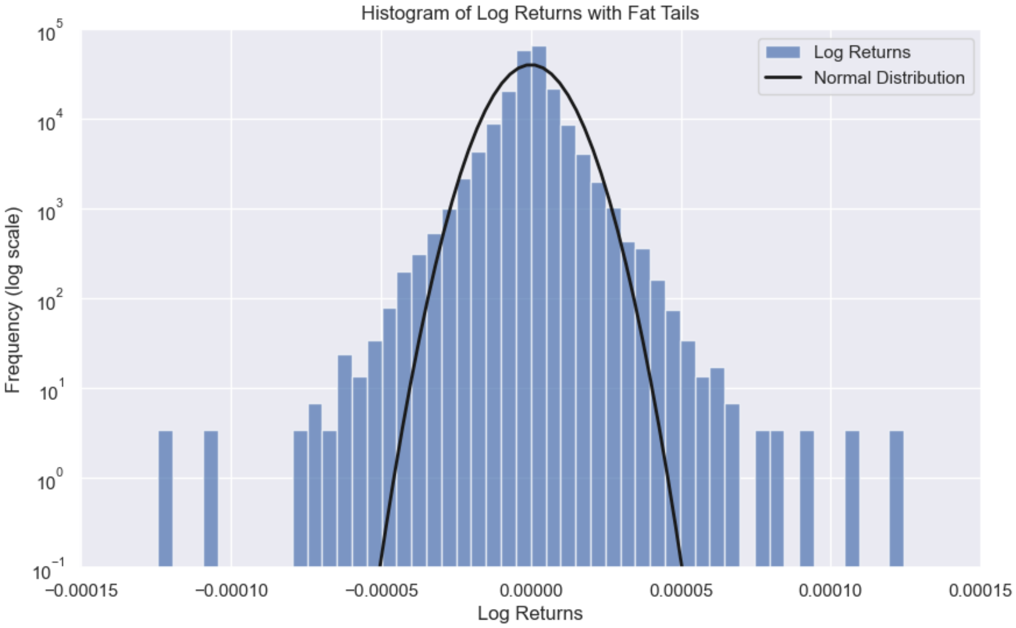

Figure 2: Histogram of Log Returns for EUR/USD on March 4th 2024

2 Estimating Volatility at Short Timescales

Because of the idiosyncrasies of high-frequency financial data, often subsumed under the umbrella term *microstructure noise*, estimating the volatility of the price process at short timescales is an arduous task. In theory, a higher sampling rate should increase the estimator's accuracy because of the larger number of data points. However, when taking into account the fact that the observed price is a noisy version of the latent price, one quickly realizes that naive volatility estimators are doomed at high frequencies.

In this section, we will first explain why naive estimators diverge at high frequency, then we will show that low-frequency estimators actually fare better. We will finally describe an unbiased volatility estimator proposed in [2] which takes the best of both world between high-frequency and low-frequency. All proofs have been omitted and can be found in [2].

2.1 Naive High-Frequency Estimator

Let S_t be the price process of a security and suppose that $X_t = \log S_t$ follows an Itô process:

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

where B_t is a standard Brownian motion. We seek to estimate the integrated variance of the process X_t over the time interval $[0, T]$, i.e. $IV = \int_0^T \sigma_t^2 dt$. A natural estimator is to use the sum of squared returns, also known as realized variance:

$$\hat{IV}_{\text{naive}} = \sum_{t_i} (X_{ti+1} - X_{t_i})^2$$

Indeed, we know by stochastic calculus that this estimator converges towards IV almost surely as the sampling frequency of the t_i goes to infinity.

However, it has been shown empirically that this naive estimator behaves poorly at higher frequencies, typically under a five minutes, again because of microstructure effects. A simple way to understand this phenomenon is to incorporate *observation error* in the process. That is, we observe $Y_{t_i} = X_{t_i} + \epsilon_{t_i}$ instead of the latent process X_t. In this model, the ϵ_{t_i} are independent centered random variables with finite variance $\mathbb{E}[\epsilon^2] = S^2$ accounting for the contamination of the latent price by microstructure effects. Within this framework, we now have

$$\hat{IV}_{\text{naive}} \xrightarrow{L^1} IV + 2nS^2$$

where n is the total number of points sampled. Since IV is bounded, we see that increasing the sampling frequency increases the bias of the estimator and indeed it diverges linearly with n as the sampling frequency increases. Thus, when using ultra high-frequency data e.g. tick-by-tick, $2nS^2 \gg IV$ such that the naive estimator becomes entirely irrelevant. In fact, when $2nS^2 \gg IV$, IV_{naive} estimates the microstructure noise $2nS^2$ instead of the integrated variance IV.

2.2 Improved Low-Frequency Estimator

One easy way to deal with the poor performance of the naive estimator at high frequencies is to sample the data sparsely. For instance, if we have second-by-second data, we can sample every 300 points bring down the sampling frequency to 5 minutes. In this case, we can show that

$$\hat{IV}_{\text{sparse}} \approx^{\mathcal{L}} IV + 2n_{\text{sparse}} S^2 + \left[4n_{\text{sparse}} \mathbb{E}[\epsilon^4] + \frac{2T}{n_{\text{sparse}}} \int_0^T \sigma_t^4 dt\right]^{1/2} N(0, 1)$$

In this setting, we see that reducing n_{sparse} effectively reduces the bias of the naive estimator. If we also take into account the variance of the estimator, we can choose n_{sparse} to minimize the MSE of the estimator, yielding

$$n^*_{\text{sparse}} = \left(\frac{T}{4S^4} \int_0^T \sigma_t^4 dt\right)^{1/3}$$

2.3 Subsampling and averaging

Although the above low-frequency approach reduces the bias of the integrated variance estimator, it is still frustrating because it only uses a small portion of the available data. For instance, subsampling every 5 minute on second-by-second data amounts to discarding 299 observations every 300. A better approach is to separate the n observations into K low-frequency bins and then compute $\hat{IV}^{(i)}_{\text{sparse}}$ over each bin B_i for $1 \leq i \leq K$. We then consider the average

$$\hat{IV}^{\text{avg}}_{\text{sparse}} = \frac{1}{K}\sum_{i=1}^{K} \hat{IV}^{(i)}_{\text{sparse}}$$

The most natural way to build the bins is to choose a low sampling frequency such as 5 minutes, and then choose B_1 to be the first observation in every 5-min interval, B_2 to be the second observation in every 5-min interval, etc. as illustrated in Figure 3.

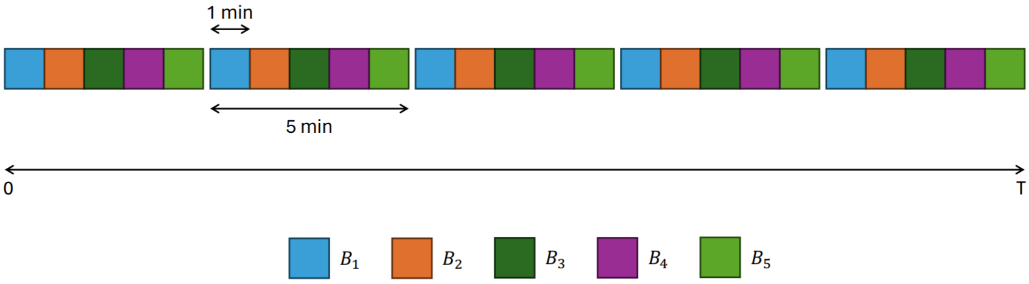

Figure 3: Building $K = 5$ bins out of 1-min sampled data.

Thus, averaging over K bins of average size $\bar{n}$, one can show that

$$\hat{IV}^{\text{avg}}_{\text{sparse}} \approx^{\mathcal{L}} IV + 2\bar{n}S^2 + \Big[4\frac{\bar{n}}{K}\mathbb{E}[\epsilon^4] + \frac{4T}{3\bar{n}}\int_0^T \sigma_t^4 dt\Big]^{1/2} N(0,1)$$

We see that $\hat{IV}^{\text{avg}}_{\text{sparse}}$ is still a biased estimator, but the bias $2\bar{n}S^2$ now scales with the average size of the bins $\bar{n}$ instead of the total number of points n, and because $\bar{n} = \frac{n}{K} < n$ the bias effectively diminishes as K increases. Again, if we look at both bias and variance, we can optimize $\bar{n}$ for the MSE of $\hat{IV}^{\text{avg}}_{\text{sparse}}$, and we then set $K^* = \frac{n}{\bar{n}^*}$ as the optimal numbers of bins. In this case we can show that

$$\bar{n}^* = \Big(\frac{T}{6S^4}\int_0^T \sigma_t^4 dt\Big)^{1/3}$$

2.4 Combining High-Frequency with Low-Frequency

Although the sparse-average estimator uses all the data while avoiding the high-frequency sampling issue, it is still a biased estimator, with bias $2\bar{n}S^2$. However, recall that the high-frequency naive estimator measures precisely the microstructure-induced bias $2nS^2$. Thus, we can use the high-frequency estimator $\hat{IV}_{\text{naive}}$ to de-bias the low-frequency estimator $\hat{IV}^{\text{avg}}_{\text{sparse}}$ by defining

$$\hat{IV}_{\text{best}} = \hat{IV}^{\text{avg}}_{\text{sparse}} - \frac{\bar{n}}{n}\hat{IV}_{\text{naive}}$$

In this setting, if the number of bins is selected as $K = cn^{2/3}$, one can show that

$$\hat{IV}_{\text{best}} \approx^{\mathcal{L}} IV + n^{-1/6}\Big[\frac{8}{c^2}S^4 + c\frac{4T}{3}\int_0^T \sigma_t^4 dt\Big]^{1/2} N(0,1)$$

Although this estimator converges at rate $n^{-1/6}$ only, it is unbiased and allows for sampling at arbitrarily high frequencies. In addition, c can be optimized for MSE, yielding

$$c^* = \Big(\frac{T}{12S^4}\int_0^T \sigma_t^4 dt\Big)^{1/3}$$

2.5 Assessing estimator performance on synthetic data

We will now compare the performance of our four estimators, from worse to best: $\hat{IV}_{\text{naive}}$, $\hat{IV}_{\text{sparse}}$, $\hat{IV}^{\text{avg}}_{\text{sparse}}$ and $\hat{IV}_{\text{best}}$. To do so, we will use synthetic data generated using the Heston stochastic volatility model [5]. This way, we will know the true value of the integrated variance $IV = \int_0^T \sigma_t^2 dt$ and we will be able to compare it to the estimated value $\hat{IV}$ for each estimator. Following the Heston model, the asset price S_t will obey the following dynamic

$$dS_t = S_t \sigma_t dB_t^S$$

$$d\sigma_t^2 = \kappa(\theta - \sigma_t^2)dt + \sigma\sigma_t\delta dB_t^\sigma$$

Note that we set $\mu_t = 0$ for the sake of simplicity. In addition, the two Brownian motions B^S and B^σ should be negatively correlated to account for the so-called *leverage effect*, which states that negative returns tend to increase future volatility. Note that the variance σ_t^2 follows a CIR process and if the Feller condition $2\kappa\theta > \sigma^2$ are met, then the variance remains positive and cannot reach zero. In this setting, θ is the long-term variance equilibrium while κ is the pullback strength towards that equilibrium. Finally, σ is the volatility of the CIR process and as such it can be thought of as the volatility of the volatility, or *vol of vol*.

Following [6], we use parameters $\sigma = 0.5$, $\kappa = 5$, $\theta = 0.1$ and $\rho = -0.3$ for the correlation between B^S and B^σ. We also set $S = 0.001$ the standard deviation of the microstructure noise, such that $\mathbb{E}[\epsilon^2] = S^2$. We also set $\sigma_0^2 = \theta = 0.1$ as the initial variance.

We use an Euler integration scheme to simulate our stochastic system, and since we want high-frequency data we use a time step $dt = 1s$. We also set $T = 24h$, such that we have a total of $n = 86400$ samples.

In this setting, the naive high-frequency estimator will use a 1-second sampling frequency i.e. the highest possible, while the naive low-frequency estimator will use a 5-minute sampling frequency, i.e. 1 observation kept every 300. The sparse-average estimator and the best estimator will both use $K = 300$ bins, such that the resulting per-bin sampling frequency is $86400/300 = 288s$ i.e. one observation every 2min48s.

The simulation protocol is as follows:

1. we generate M independent price paths using the Heston model described above

2. for each price path $X^{(i)}$, we compute the integrated variance as the discrete integral of the $(\sigma_t^{(i)})^2$, and we compute the four estimates $\hat{IV}_{\text{naive}}^{\text{HF}}$, $\hat{IV}_{\text{naive}}^{\text{LF}}$, $\hat{IV}_{\text{sparse}}^{\text{avg}}$ and $\hat{IV}_{\text{best}}$

3. for each estimator we plot the histogram of the relative error $\frac{\hat{IV}-IV}{IV}$

The resulting histograms are given in Figure 4. As expected, we see that the high-frequency naive estimator behaves extremely poorly, with a relative approximation error consistently above 500. The low-frequency estimator is two orders of magnitude better in terms of performance but is still highly biased, with a relative approximation error consistently greater than 1.5. The sparse-average estimator is only slightly better than the low-frequency naive estimator, with a similar positive bias but lower variance overall. Finally, the best estimator is the only unbiased estimator and also has the lowest variance, making it far better than the preceding estimators. In our Monte Carlo simulation, we find that the relative error $\frac{\hat{IV}_{\text{best}}-IV}{IV}$ is centered with a standard deviation of about 0.07, meaning that the relative error is below 14% in 95% of cases.

Thus, we can conclude that $\hat{IV}_{\text{best}}$ is a satisfying estimator of integrated variance which has moderate variance thanks to its low-frequency part and no microstructure-induced bias thanks to its high-frequency part.

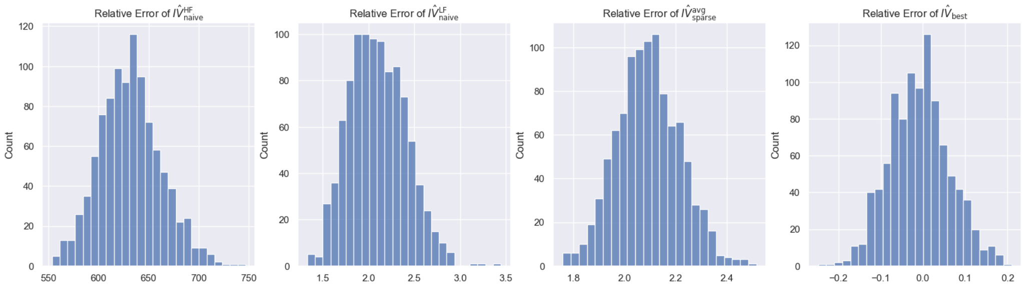

Figure 4: Histogram of the relative error $\frac{\hat{IV}-IV}{IV}$ for the four estimators proposed using Monte Carlo simulations with $M = 1000$.

2.6 Implementation on tick-by-tick forex data

In this last subsection, we will look at high-frequency, tick-by-tick forex data for the EUR/USD pair obtained from TrueFX. We will first measure integrated variance using the naive estimator $\hat{IV}_{\text{naive}}$ at various frequencies to check that our estimations do indeed diverge as the sampling frequency goes to infinity. We will then measure integrated variance using the best estimator $\hat{IV}_{\text{best}}$ and compare the results.

To avoid the abnormal spread occurring on forex between 22:00 and 22:30 (because of large banks performing their daily IT maintenance during this time slot), we will look at just 23 hours of non-stop EUR/USD data collected from 02/04/2024 18:00 EST to 02/05/2024 17:00 EST. Our dataset is tick-by-tick data containing only the best bid and best ask at a given time.

We have a total of 681,912 ticks, representing an average of 8.24 ticks/s. We could define S_t to be the logarithm of the mid price $\text{mid} = \frac{\text{ask}+\text{bid}}{2}$ but in practice this leads to an artificial reduction of volatility (since we're averaging). Thus, we will take the logarithm of the best bid for S_t. Using the best ask yields identical results.

Our experiment protocol is the following:

1. we define a range of sampling frequencies from $\tau = 100ms$ to $\tau = 20s$ using $100ms$ intervals

2. for each of τ, we resample S_t to this frequency. [note that when a resampling bin has several ticks we take the first (i.e. earliest) one, and when a resampling bin has no ticks we use forward fill (i.e. the fill the value with the latest value available)]

3. we estimate integrated variance over our whole dataset using the naive estimator and the best estimator, both sampling at frequency τ [note that for the sake of completeness we use $K = 10, 100, 1000$ for the best estimator]

4. we plot $\left(\tau, \hat{IV}(\tau)\right)$ for each estimator

The results are illustrated in Figure 5. As expected, the naive estimator diverges at high sampling frequencies, as early as $\tau \sim 10s$. On the contrary, the best estimator remains stable at higher frequencies for all values of K, though for $K = 10$ we see a small bias at the highest sampling frequency ($\tau = 100ms$). Interestingly, the best estimators for $K = 100$ behave slightly differently, with the $\hat{IV}_{\text{best}}^{K=100}$ monotonously increasing as τ decreases, whereas $\hat{IV}_{\text{best}}^{K=1000}$ remains very stable up to $\tau \sim 10s$ then decreases and reaches a minimum at around $\tau \sim 2s$ and then increases again. In this case we obviously do not know the integrated variance so we cannot be sure which of these two estimators is more accurate.

3
A NEW MODEL TO EXPLAIN HIGH-FREQUENCY VOLATILITY

3.1 Modeling the price by a Hawkes process

Up until now, we have modeled the so-called *microstructure noise* (as it is often done in litterature) with the introduction of a latent price and an additive noise. However, although this approach can give interesting results, it doesn't quite reproduce the high frequency effects that we are interested in. In this report, we will focus on one specific effect which affects the shape of the signature plot. This signature plot is simply the plot of the realized volatility against the sampling frequency.

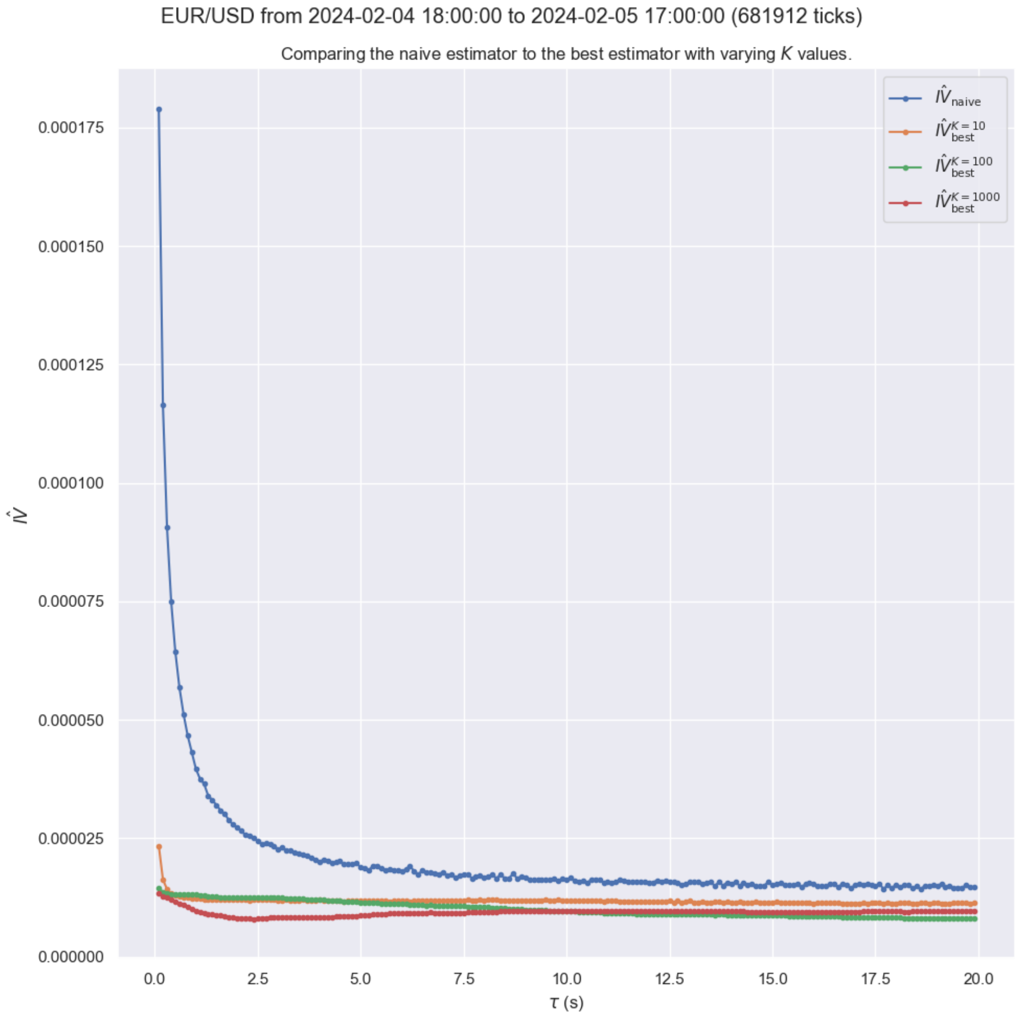

Figure 5: Integrated variance estimated using naive estimator and best estimator at various frequencies.

One the the problems of the additive noise model is that it doesn't really give insight into the nature of this noise and its micro-structure properties and the introduction of a noise at high frequency may seem a bit artificial.

To remediate those issues, we introduce a new dynamic which explains price movements using mutually exciting Hawkes processes.

Let $X(t)$ be the price of our asset, $N_1(t)$ and $N_2(t)$ two point processes that represents the count of positive and negative price jumps up until time t. We then have :

$$X(t) = N_1(t) - N_2(t) \tag{1}$$

.

The idea is that $N_1(t)$ and $N_2(t)$ are mutually exciting at short time scales. This means that at high frequency, when the price goes up a little bit, it tends to go down right after and thus we model our mean reverting property. At longer time scales, the model should behave as a Brownian motion to match the Black-Scholes model. This is indeed the case if $N_1(t)$ and $N_2(t)$ are independent Poisson processes with intensity μ for instance. Indeed, in this case

$$\lim_{t\to\infty} \frac{1}{\sqrt{T}} X(tT) = \sqrt{2\mu} B(t) \tag{2}$$

where $B(t)$ is a Brownian motion.

This approach can thus give a unified view of the asset's price behaviour while reproducing (as we will show later) the effects we observe at both high and low frequencies.

These mutually exciting processes can be naturally modelled by Hawkes processes in order to introduce those time scale differences (which will be modeled by the shape of the kernels). Let λ_1 and λ_2 be the intensities of N_1 and N_2.

The mean reverting property we aim for suggest to set the self-exciting kernels to 0 and simply keep the base excitations as well as the cross exciting kernels. Therefore, we have

$$\lambda_i = \mu_i + \int_{-\infty}^{t} \phi_{ij}(t-u) dN_j(u), i \in 1, 2. \tag{3}$$

As the situation appears to be symmetric if the price goes down or up (it may not be completely true but we make this approximation), we take $\phi_{12} = \phi_{21} = \phi$ and $\phi_{11} = \phi_{22} = 0$. As it is done in [1], we use an exponential kernel for simplicity and because this can easily relate to the time scale we are considering. Therefore, we have $\phi(t) = \beta \cdot \exp(-\alpha t)$ where $\alpha, \beta \in \mathbb{R}_+^*$.

To make sure this approach is sensible, we can compare the plots of real tick data versus data generated by our model.

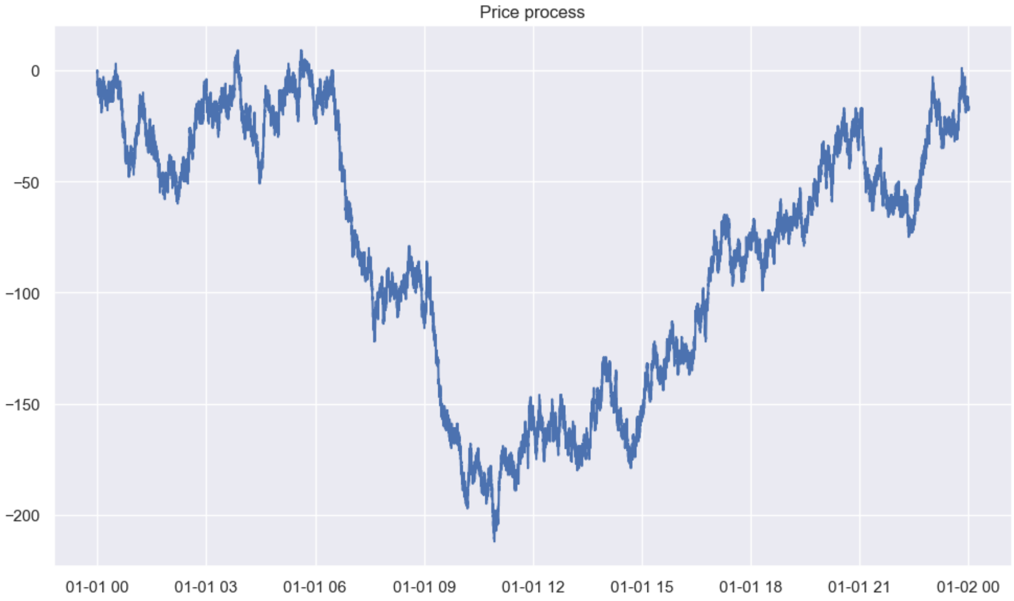

Figure 6: Plot of EUR/USD tick data

We see that they look similar. Although this is no proof, it is reassuring that our model gives similar outputs if we want to fit it to real data later.

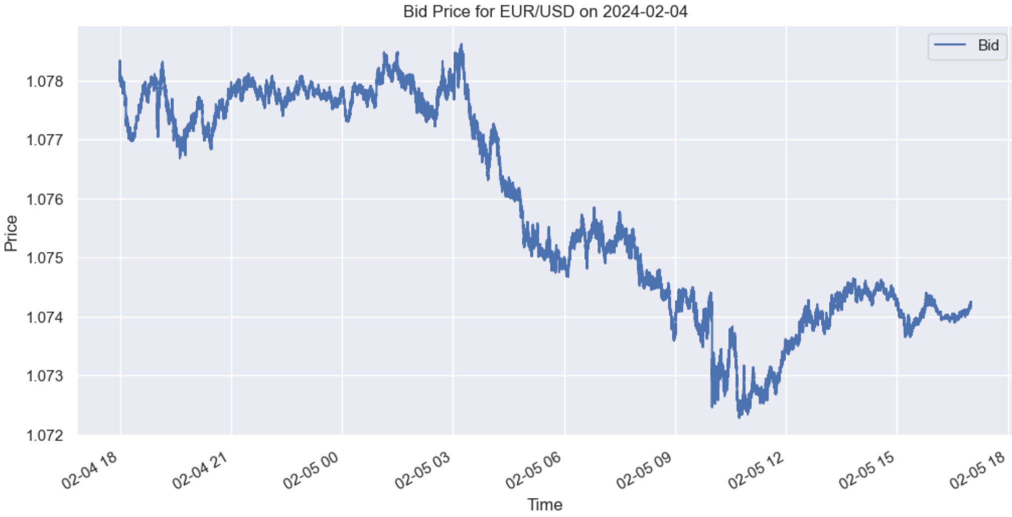

Figure 7: Plot of our simulated Hawkes process with arbitrary parameters

3.2 Explaining the signature plot

We introduce an estimator of the volatility at the frequency $\frac{1}{\tau}$

$$\hat{C}(\tau) = \frac{1}{T}\sum_{n=0}^{T/\tau}(X((n+1)\tau) - X(n\tau))^2 \tag{4}$$

The main behavior that we are trying to address with this new model is the shape of $\hat{C}(\tau)$ as $\tau \to 0$. Indeed, the classical Black-Scholes model would predict a flat line for the $C(\tau) = \mathbb{E}[\hat{C}(\tau)]$ function as the volatility is constant. This is what we can see in the figure as τ grows larger.

However, using our model, we can actually compute analytically the form of $C(\tau)$ (the proof is omitted and can be found in [1]). Under the right stability condition for a Hawkes process, meaning if $||\phi||_1 < 1$, we have

$$C(\tau) = \Lambda\Big(\kappa^2 + (1-\kappa^2)\frac{1-\exp(-\gamma t)}{\gamma t}\Big) \tag{5}$$

Where $\Lambda = \frac{2\mu}{1-||\phi||}$, $\kappa = \frac{1}{1+||\phi||}$ and $\gamma = \alpha + \beta$.

This equation is quite interesting. Firstly, we show that for the right values of μ, α and β this can fit pretty well our $\hat{C}(\tau)$ estimator of the volatility for every time-scale. Secondly, this gives us a way to estimate the parameters of our model.

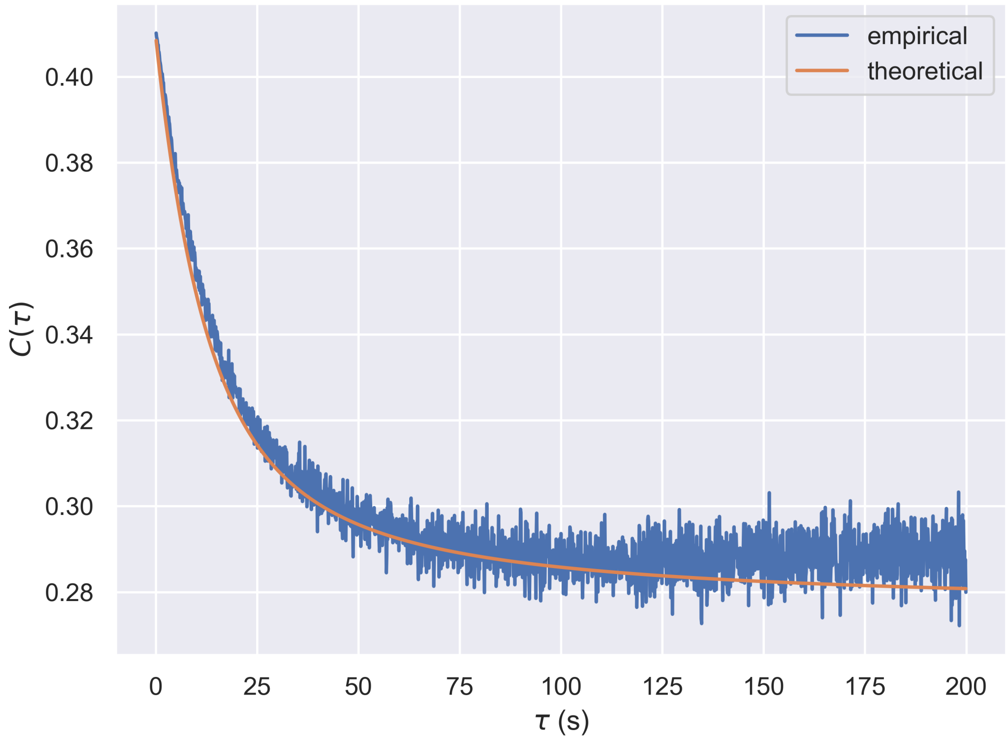

Figure 8: Plot of $\hat{C}(\tau)$ for the EUR/USD tick data

3.3 Fitting the model on real data

Now that we have an analytical closed form of $C(\tau)$ we can compute the parameters of our model by simply taking the argmin of the mean squared error between the real data and our theoretical data.

We have implemented this approach on EUR/USD tick data which yields the curves of figure 4.

This approach can be complemented by a maximum likelihood estimation of our model directly by looking at the asset's price. This is done using the closed formula given in [7] which we implemented for our two dimensional Hawkes process. Indeed, we have for each dimension m :

$$\ln L(m(\{t_i\})) = T - \sum_{i=1}^{N}\sum_{n=1}^{M}\sum_{j=1}^{P} \alpha_j^{mn}\beta_j^{mn}\left[1 - e^{-\beta_j^{mn}(T-t_i)}\right] + \sum_{t_l^m} \ln\left[\lambda_{m0}(t_l^m) + \sum_{n=1}^{M}\sum_{j=1}^{P}\alpha_j^{mn} Rj^{mn}(l)\right] \tag{6}$$

Where the t_i or the time-points of process m, P is 1 in our case and $R_j^{mn}(l) = \sum_{t_n^k < t_m^l} e^{-\beta_j^{mn}(t_m^l - t_n^k)}$.

And the maximum likelihood estimate (MLE) of our process defined in (1) is given by:

$$\theta_{\text{MLE}} = \arg\min_{\theta}(L_1(\theta) + L_2(\theta))$$

The implementation that can be found on our github seems unconsistent however because it is too unstable. So we have only showed the results of the first method, using the MSE estimator and the signature analytical form.

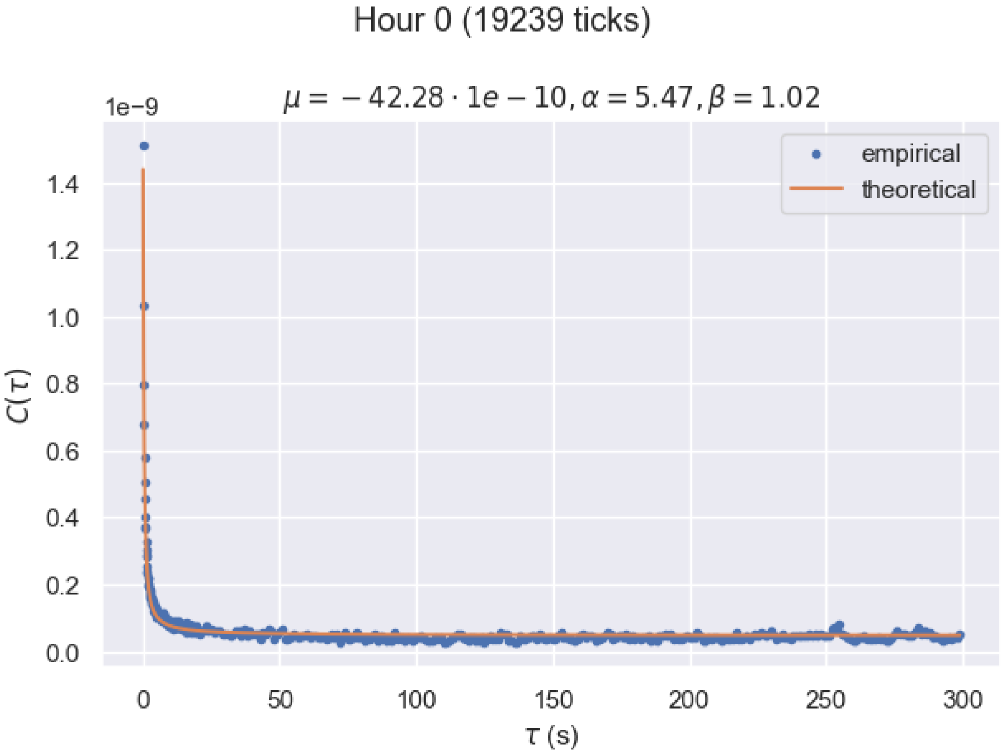

Figure 9: Plot of the MLE estimator found by fitting the theoretical curve to real EUR/USD tick data

4 CONCLUSION

In this report, we explored different methods to estimate volatility at short timescales in the context of high-frequency financial data. We began by discussing the challenges posed by microstructure noise, which makes naive volatility estimators unreliable at high frequencies. We introduced the concept of naive high-frequency estimators and demonstrated their limitations due to microstructure effects, leading to biased estimates.

To address these challenges, we have reviewed improved low-frequency estimators found in litterature that sample the data sparsely, reducing bias but still maintaining some level of accuracy.

Furthermore, we presented a hybrid approach that combines high-frequency and low-frequency estimators to mitigate bias effectively. This approach leverages the strengths of both methods and results in an unbiased estimator suitable for high-frequency data analysis.

We implemented these methods using synthetic data generated from the Heston stochastic volatility model and demonstrated their performance through Monte Carlo simulations. The best estimator showed promising results with moderate variance and no microstructure-induced bias, making it suitable for practical applications.

Finally, we discussed a novel modeling approach using mutually exciting Hawkes processes to explain price movements at different time scales. This approach provided insights into the nature of microstructure noise and allowed for analytical predictions of volatility behaviors, as demonstrated through the signature plot analysis.

We briefly mentionned the Epps effect which states that the correlation between asset price is very low on microscopic scales while they converge to their "regular" value at more long-term scales. Although we did not implement it, [1] shows that the Hawkes process model correctly reproduces this effect and thus seems consistent with our analysis.

REFERENCES

[1] Bacry, E. *Modeling microstructure noise with mutually exciting point processes* (2018).

[2] Zhang, L. et al. *A Tale of Two Time Scales: Determining Integrated Volatility With Noisy High-Frequency Data* (2018).

[3] Hansen, P. et al. *Realized Variance and Market Microstructure Noise* (2004).

[4] Epps, T. W. *Comovements in stock prices in the very short run* (1979).

[5] Heston, S. *A closed-form solutions for options with stochastic volatility* (1993).

[6] Aït Sahalia Y. *High frequency market microstructure noise estimates and liquidity measures* (2009).

[7] Ioane Muni Toke, *An Introduction to Hawkes Processes with Applications to Finance*, February 4th, 2011