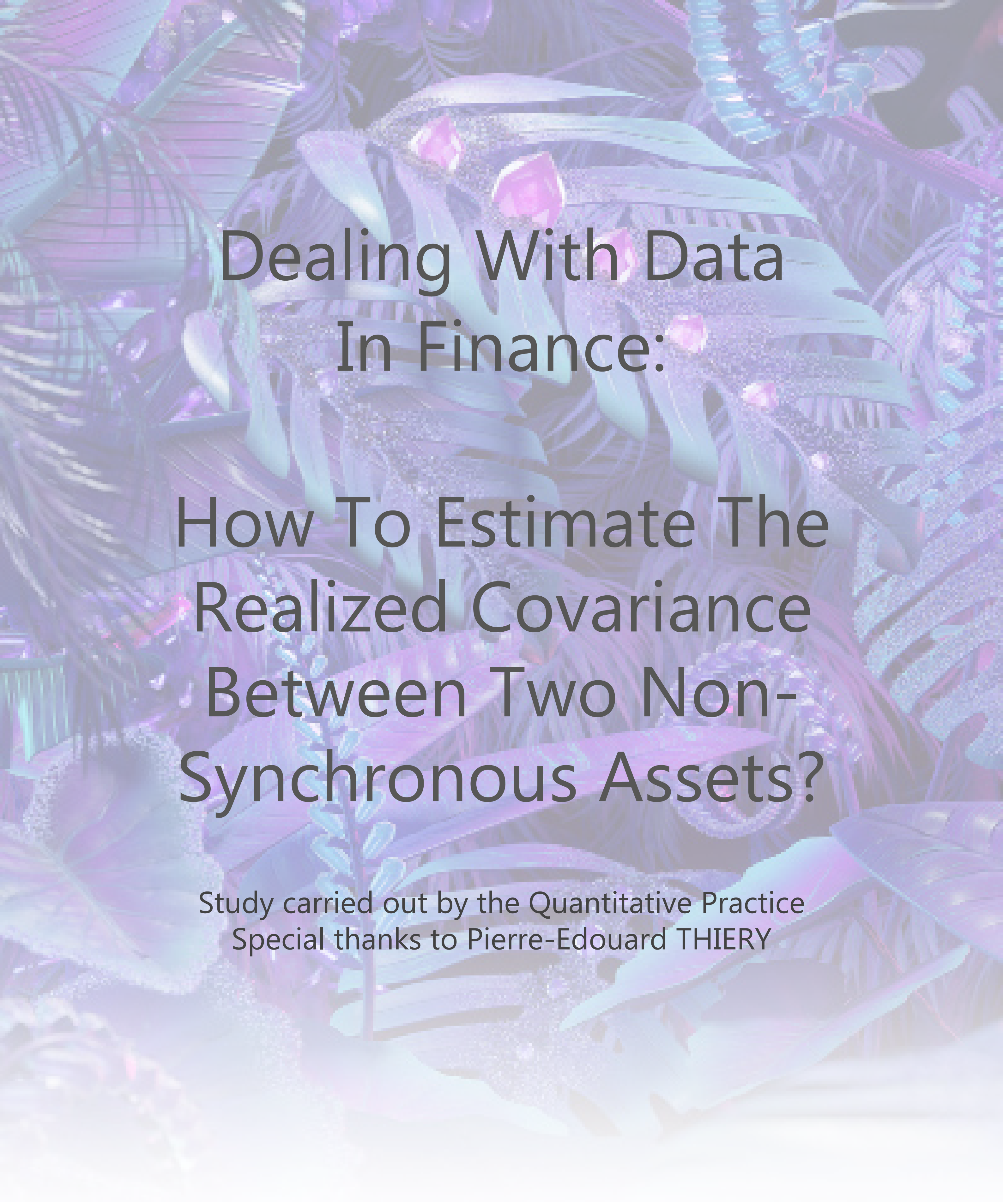

Dealing With Data In Finance:

How To Estimate The Realized Covariance Between Two Non-Synchronous Assets?

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

awalee

AVRIL 2021

Summary

Introduction

Finance relies heavily on numerical data, and this explains why mathematical and statistical approaches are so widespread in finance. Nonetheless numerical data in finance display lots of peculiarities insofar as they are mainly time-dependent. Therefore many classical statistical approaches may fail when directly implemented within a financial framework. The current development of the sub-field of Machine Learning called "financial Machine Learning" is partially due to the difficulties that may arise when dealing with time-dependent data.

In this paper we choose not to focus on the earmarks of financial Machine Learning, but instead on another issue regarding financial data: the "asynchronicity" of observations. When coping with only one underlying, the concept of "asynchronicity" is irrelevant: the observations are made at different points on the time line. This dates define a first time reference. A second underlying, with its own observations made at other dates, is necessary to define a second time reference, and then to introduce the idea of asynchronicity. If the two time references do not match, it means that the observations are non-synchronous. We also write "asynchronous" in the rest of this paper. This is of the utmost importance of finance inasmuch as data are time-dependent: if the second underlying is observed at a time t_j^2 which is between two observation times of the first underlying, t_i^1 and t_{i+1}^1, the observation made at t_j^2 is of a very different kind compared to the observations made at t_i^1 and t_{i+1}^1: the observation made at t_j^2 contains more information than the one made at t_i^1, and less than the one made at t_{i+1}^1. Since financial mathematics rely heavily on the information which is contained in the numerical data up to the present date, it seems pretty unwise to mix up data coming from two different time references.

Nonetheless asynchronicity remains one of the crude realities of financial data, whereas synchronicity is a theoretical fantasy. Therefore it matters to have a clear mathematical framework to deal with such data. In this paper we focus on the question of the covariance estimation between two asynchronous underlyings: after showing that too simple approaches to handle the covariance estimation fails, we will set forth an estimation procedure of the realized covariance between the two underlyings which is both consistent and unbiased

1 A Simple Approach To Deal With Asynchronous Data

In this first part, we set forth a simple approach to deal with asynchronous data when it comes to estimating the realized covariance between two underlyings.

Let us define the following mathematical framework: we consider two Ito processes denoted P^1 and P^2:

$$\begin{cases} dP_t^1 = \mu_t^1 dt + \sigma_t^1 dW_t^1 \\ dP_t^2 = \mu_t^2 dt + \sigma_t^2 dW_t^2 \end{cases}$$

where W^1 and W^2 are two Brownian motions on a probability space $(\Omega, \mathscr{F}, \mathbb{P})$. The correlation between those two Brownian motions is a deterministic function denoted ρ_t, meaning that:

$$d < W^1, W^2 >_t = \rho_t dt$$

We assume we work on a time period $[0, T]$. Our purpose is to estimate the realized covariance between the two processes P^1 and P^2 on $[0, T]$. If we denote θ this quantity, then

$$\theta = \int_0^T \sigma_t^1 \sigma_t^2 \rho_t dt$$

The theoretical results [1] regarding Ito processes posit that, if we consider a given grid of times $G_n = \{t_0, t_1, \ldots, t_n\}$ such that

$$0 = t_0 < t_1 < \cdots < t_n = T$$

then the following quantity

$$\sum_{i=0}^{n-1} \left(P_{t_{i+1}}^1 - P_{t_i}^1\right)\left(P_{t_{i+1}}^2 - P_{t_i}^2\right)$$

converges in probability towards θ when

$$\sup_{0 \leq i \leq n-1} |t_{i+1} - t_i| \xrightarrow{n \to \infty} 0$$

This theoretical result relies on the assumption that the two processes P^1 and P^2 are observed at synchronous times, but this is not always the case in real life situations.

As of now, we assume that several observations of the two processes within the time period $[0, T]$ are available, but those observations are not necessarily synchronous. If we denote m_1 the number of observations for the process P^1, the observations times for this process are:

$$0 \leq T_0^1 < T_1^1 < \cdots < T_{m_1}^1$$

Likewise, m_2 indicates the number of observation times for the process P^2, and those times are denoted T_i^2 for $0 \leq i \leq m_2$.

It is possible to devise a first estimator of the realized covariance between P^1 and P^2, which will be referred to as the "previous tick" estimator, using those asynchronous data. To do so, we first define two processes Q^1 and Q^2 on the time period $[0, T]$ as follows: for $t \in [0, T]$, we find the two consecutive observation times such that

$$T_i^k \leq t < T_{i+1}^k$$

and then we define:

$$Q_t^k = P_{T_i^k}^k$$

This definition only means that the two processes Q^1 and Q^2 are actually constant piecewise processes, created thanks to the observations of P^1 and P^2 respectively.

It is then fairly natural to define a discretization step h for the time period $[0, T]$ such that $T = mh$ with $m \in \mathbb{N}$ and to consider the following estimator for the realized covariance between P^1 and P^2:

$$V_h = \sum_{i=0}^{m-1} \left(Q^1_{(i+1)h} - Q^1_{ih}\right)\left(Q^2_{(i+1)h} - Q^2_{ih}\right)$$

This approach may seem rather elegant due to its simplicity; sadly it fails when it comes to properly estimating the covariance. We display in the following section a numerical implementation of this estimator on simulated data.

2 The Failure Of The Simple Approach: The Epps Effect

We show in this section that the previous tick estimator of the realized covariance fails: when the discretization step h converges towards 0, this estimator, instead of converging towards the realized covariance θ, converges towards 0.

Besides our implementation will also illustrate the fact that the previous tick estimator is constantly biased, even for higher discretization steps.

2.1 Definition of the two processes

For our implementation, we assume that the two processes P^1 and P^2 are merely proportional to Brownian motions:

$$\begin{cases} P^1_t = \sigma^1 W^1_t \\ P^2_t = \sigma^2 W^2_t \end{cases}$$

with $\sigma_1 = 0.1$, $\sigma_2 = 0.5$, and the correlation between the two Brownian motions W^1 and W^2:

$$d < W^1, W^2 >_t = \rho dt$$

with $\rho = 0.5$.

The time window chosen for our implementation is $T = 100$. All those values mean that the realized covariance is equal to:

$$\theta = \int_0^T \sigma_1\sigma_2\rho dt = 100 \times 0.1 \times 0.5 \times 0.5 = 2.5$$

2.2 Simulation of the asynchronous times and simulation of the observations

To generate the two sets of observation times, we consider a Poisson process $\mathscr{P}(\lambda)$ with $\lambda = 0.05$. By simulating twice such a Poisson process, this provides us with two set of times between 0 and T. Each set of times represents the observation times for one of our two processes P^1 and P^2. The times T^k_i are now known in our implementation.

It is then possible to simulate the two processes P^1 and P^2, which are basically two correlated Brownian motions multiplied by a given constant, at the observation times.

Once the observations of P^1 and P^2 are generated, it is then fairly straightforward to implement the previous estimator V_h for several values of the discretization step h.

2.3 Results of the previous tick estimation

Figure 1 displays the estimation of the realized covariance on $[0, T]$ for several values of h, from 1.0 to close to 0.

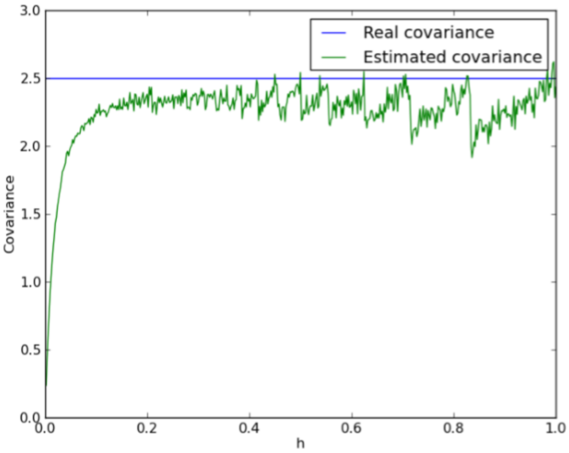

Figure 1: Estimation of the realized covariance using the previous tick estimator

This implementation exemplifies two important phenomena. First, when we use high-frequency data, i.e. when $h \to 0$, the estimation of the covariance converges towards 0. This effect is known as the Epps effect [2]. The estimator is no longer performing with high-frequency data.

Besides we observe that, for lower frequencies, the previous tick estimator appears to be biased. This is another of its shortcomings we would like to correct.

In the following section, we present another estimator for the realized covariance with asynchronous data which is both consistent and unbiased.

3 The Consistent And Unbiased Estimator Of The Realized Covariance

The framework is exactly the same as above: we wish to estimate the realized covariance between two processes P^1 and P^2, which are observed at asynchronous times T^k_i.

The observations times define two partitions of the time window $\left[0, T\right[$, if we assume that $T^1_0 = T^2_0 = 0$ and $T^1_{m_1} = T^2_{m_2} = T$. Those are the only two supplementary assumptions we make. The first partition is denoted π^1 and is given by the observation times of the process P^1: $\pi^1 = \left(I^i\right)_{0 \leq i \leq m_1 - 1}$ where, for $0 \leq i \leq m_1 - 1$

$$I^i = \left[T^1_i, T^1_{i+1}\right[$$

So the intervals of the first partition π^1 are given by the observation times of the first process P^1. We use a similar set of notations for the second process: the partition is denoted $\pi^2 = \left(J^j\right)_{0 \leq j \leq m_2 - 1}$, with

$$J^j = \left[T_j^2, T_{j+1}^2\right[$$

For a process X, we define the variation of the process on a given interval $\left[a, b\right[$:

$$\Delta X\left(\left[a, b\right[\right) = X(b) - X(a)$$

We can then define what we call the "robust" estimator of the realized covariance between two processes observed at asynchronous times:

$$U_n = \sum_{i=0}^{m_1-1} \sum_{j=0}^{m_2-1} \Delta P^1\left(I^i\right) \Delta P^2\left(J^j\right) 1_{I^i \cap J^j \neq \emptyset}$$

It is worth noticing that we have added a subscript n: n is merely a variable which represents the size of our partitions. $n \to \infty$ means that ever more observations are available and that the maximum length between two consecutive observation times converges towards 0.

U_n is a consistent and unbiased estimator of the realized covariance. We provide here, in a simple case, a heuristic of the proof that the robust estimator is indeed unbiased: we assume that $P^1 = \sigma_1 W^1$ and $P^2 = \sigma_2 W^2$ where W^1 and W^2 are two Brownian motions with:

$$d < W^1, W^2 >_t = \rho dt$$

Within this framework, the realized covariance is fairly simple:

$$\theta = \sigma_1 \sigma_2 \int_0^T \rho dt = \sigma_1 \sigma_2 \rho T$$

We also assume that the observation times are given by two Poisson processes. We denote Π the information which is known at all the observation times.

To compute $\mathbb{E}\left[U_n\right]$, we use the information Π; we can then write:

$$\mathbb{E}\left[U_n\right] = \mathbb{E}\left[\mathbb{E}\left[U_n | \Pi\right]\right]$$

$$= \mathbb{E}\left[\mathbb{E}\left[\sum_{i=0}^{m_1-1} \sum_{j=0}^{m_2-1} \Delta P^1\left(I^i\right) \Delta P^2\left(J^j\right) 1_{I^i \cap J^j \neq \emptyset} \middle| \Pi\right]\right]$$

$$= \mathbb{E}\left[\sum_{i=0}^{m_1-1} \sum_{j=0}^{m_2-1} \mathbb{E}\left[\Delta P^1\left(I^i\right) \Delta P^2\left(J^j\right) 1_{I^i \cap J^j \neq \emptyset} \middle| \Pi\right]\right]$$

Since the expression $1_{I^i \cap J^j \neq \emptyset}$ is fully determined when Π is known, we can write it out of the second expectation:

$$= \mathbb{E}\left[\sum_{i=0}^{m_1-1} \sum_{j=0}^{m_2-1} \mathbb{E}\left[\Delta P^1\left(I^i\right) \Delta P^2\left(J^j\right) \middle| \Pi\right] 1_{I^i \cap J^j \neq \emptyset}\right]$$

$$= \mathbb{E}\left[\sum_{i,j, I^i \cap J^j \neq \emptyset} \mathbb{E}\left[\Delta P^1\left(I^i\right) \Delta P^2\left(J^j\right) \middle| \Pi\right]\right]$$

$$= \mathbb{E}\left[\sum_{i,j, I^i \cap J^j \neq \emptyset} \mathbb{E}\left[\sigma_1 \Delta W^1\left(I^i\right) \sigma_2 \Delta W^2\left(J^j\right) \middle| \Pi\right]\right]$$

$$= \sigma_1 \sigma_2 \mathbb{E}\left[\sum_{i,j, I^i \cap J^j \neq \emptyset} \mathbb{E}\left[\left(W^1\left(T_{i+1}^1\right) - W^1\left(T_i^1\right)\right)\right.\right.$$

$$\left.\left.\times\left(W^2\left(T_{j+1}^2\right) - W^2\left(T_j^2\right)\right) \middle| \Pi\right]\right]$$

Since $d < W^1, W^2 >_t = \rho dt$, we know that we can write:

$$W_t^2 = \rho W_t^1 + \sqrt{1 + \rho^2} B_t$$

where (B_t) is a Brownian motion which is independent from W^1. So we can see that:

$$\left(W^1\left(T_{i+1}^1\right) - W^1\left(T_i^1\right)\right)\left(W^2\left(T_{j+1}^2\right) - W^2\left(T_j^2\right)\right)$$

$$= \rho\left(W^1\left(T_{i+1}^1\right) - W^1\left(T_i^1\right)\right)\left(W^1\left(T_{j+1}^2\right) - W^1\left(T_j^2\right)\right)$$

$$+ \sqrt{1 - \rho^2}\left(W^1\left(T_{i+1}^1\right) - W^1\left(T_i^1\right)\right)\left(B\left(T_{j+1}^2\right) - B\left(T_j^2\right)\right)$$

If we take the expectation knowing Π of this sum of two terms, the second expression is equal to 0:

$$\mathbb{E}\left[\sqrt{1 - \rho^2}\left(W^1\left(T_{i+1}^1\right) - W^1\left(T_i^1\right)\right)\left(B\left(T_{j+1}^2\right) - B\left(T_j^2\right)\right) | \Pi\right]$$

$$= \sqrt{1 - \rho^2} \mathbb{E}\left[\left(W^1\left(T_{i+1}^1\right) - W^1\left(T_i^1\right)\right) | \Pi\right]$$

$$\times \mathbb{E}\left[\left(B\left(T_{j+1}^2\right) - B\left(T_j^2\right)\right) | \Pi\right]$$

because the two Brownian motions W^1 and B are independent. The two expectations are equal to zero since, when Π is known, the two differences have the same law as a normal variable whose mean is equal to 0.

We now focus on the expectation knowing Π of the first term. It is pivotal to remind that $I^i \cap J^j \neq \emptyset$, for instance we can assume that:

$$T_i^1 < T_j^2 < T_{i+1}^1 < T_{j+1}^2$$

So:

$$\mathbb{E}\left[\rho\left(W^1\left(T_{i+1}^1\right) - W^1\left(T_i^1\right)\right)\left(W^1\left(T_{j+1}^2\right) - W^1\left(T_j^2\right)\right) | \Pi\right]$$

$$= \rho \mathbb{E}\left[\left(\underbrace{W^1\left(T_{i+1}^1\right) - W^1\left(T_j^2\right)}_{A} + \underbrace{W^1\left(T_j^2\right) - W^1\left(T_j^2\right)}_{B}\right)\right.$$

$$\left.\left(\underbrace{W^1\left(T_{j+1}^2\right) - W^1\left(T_{i+1}^1\right)}_{C} + \underbrace{W^1\left(T_{i+1}^1\right) - W^1\left(T_j^2\right)}_{D}\right) \middle| \Pi\right]$$

For the sake of clarity, we now only use the notation A, B, C and D. The expectation is equal to:

$$\rho\left(\mathbb{E}\left[AC | \Pi\right] + \mathbb{E}\left[AD | \Pi\right] + \mathbb{E}\left[BC | \Pi\right] + \mathbb{E}\left[BD | \Pi\right]\right)$$

Thanks to the common properties of the Brownian motion, we can see that three of those four terms are equal to 0:

$$\mathbb{E}\left[AD|\Pi\right] = \mathbb{E}\left[BC|\Pi\right] = \mathbb{E}\left[BD|\Pi\right] = 0$$

This stems from the fact that, for a Brownian motion B and times $t_1 < t_1 + h < t_2 < t_2 + h$

$$\mathbb{E}\left[\left(B_{t_2+h} - B_{t_2}\right)\left(B_{t_1+h} - B_{t_1}\right)\right]$$

$$= \mathbb{E}\left[\left(B_{t_1+h} - B_{t_1}\right)\mathbb{E}\left[\underbrace{\left(B_{t_2+h} - B_{t_2}\right)}_{\mathcal{N}(0,h)}\middle|\mathscr{F}_{t_2}\right]\right] = 0$$

It is then possible to write:

$$\rho\mathbb{E}\left[AC|\Pi\right] = \rho\mathbb{E}\left[\left(\underbrace{\Delta W^1\left(I^i \cap J^j\right)}_{\mathcal{N}\left(0,l\left(I^i\cap J^j\right)\right)}\right)^2\middle|\Pi\right]$$

$$= \rho \times l\left(I^i \cap J^j\right)$$

where l gives the length of the considered interval. If we sum over the indices i and j such that $I^i \cap J^j \neq \emptyset$, we find that:

$$\mathbb{E}\left[U_n\right] = \sigma_1\sigma_2\rho \sum_{i,j,I^i\cap J^j\neq\emptyset} l\left(I^i \cap J^j\right) = \sigma_1\sigma_2\rho T = \theta$$

This proof shows that, in a simple case, our robust estimator is indeed unbiased. The result is still valid in the more general case, when P^1 and P^2 are two Ito processes. The idea of the proof is very similar to the one in the simple case.

It is also possible to demonstrate the following result regarding our robust estimator:

$$U_n \xrightarrow[L^2]{n\to\infty} \theta$$

This result shows that our estimator converges towards the realized covariance, in the 2-nd mean, so also in the 1-st mean, in probability, and almost surely. To see that, we use the following result [3]:

$$\mathbb{E}\left[U_n^2\right] = \theta^2 + o(1)$$

We can then write:

$$\mathbb{E}\left[U_n^2\right] - \theta^2 = \mathbb{E}\left[U_n^2\right] - \mathbb{E}\left[U_n\right]^2 = var\left(U_n\right)$$

$$= \mathbb{E}\left[\left(U_n - \mathbb{E}\left(U_n\right)\right)^2\right] = \mathbb{E}\left[\left(U_n - \theta\right)^2\right] \xrightarrow{n\to\infty} 0$$

4 Applications Of The Robust Estimator: The Lead-Lag Effect

In this final section, we would like to insist on the interest of what we have called the "robust" estimator. Indeed, this estimator has been devised with a single purpose in mind, i.e. estimating the realized covariance of two processes when the observation times are asynchronous; but, as a matter of fact, it can be used in miscellaneous contexts.

We only display here a few ideas regarding what is knows as the lead-lag effect. This effect occurs when a delay seems to exists between two processes: a process P^2, called the lagger, replicates partially the variations of a process P^1, called the leader. It is then crucial to properly estimate the lead-lag parameter θ such that the variation of P^2 at a time t is very similar to the variation of P^1 at $t - \theta$.

We do not delve into the details in this paper, whose purpose is to focus on the estimation of the realized covariance with asynchronous data. For further mathematical context regarding the lead-lag effect, the reader can refer to [4]. We only want to shed a light on the fact that an estimator very similar to the one used for estimating the realized covariance can also be used to estimate the lead-lag parameter. It is written as a sum of kind

$$\sum_{i,j} \Delta P^1\left(I^i\right)\Delta P^2\left(J^j\right) 1_{I^i\cap J^j\neq\emptyset}$$

where the partition $\left(J^j\right)$ is built by translating the initial partition $\left(I^i\right)$ with a parameter ξ. The objective is then to find the value ξ that maximizes the above-mentioned sum: the optimal value for ξ then converges towards the lead-lag parameter.

This shows that the kinds of sums we have used in this paper can be helpful, not only when dealing with asynchronous data, but in every situation where different time references intervene.

Conclusion

In this paper, we decided to face one of the many particularities about financial data in real life: when working with several underlyings, it is quite common that the available data are asynchronous. Such a situation may arise when trying to estimate the realized covariance between two processes: this is the problem we have chosen to tackle in this paper.

Naive approaches to circumvent the issue of non-synchronous data are quite easy to implement, but they display lots of shortcomings, therefore justifying the need for a more robust approach. It is then necessary to devise an estimator of the realized covariance which is both consistent and convergent towards the desired quantity.

However focusing on the problem of the covariance estimation with asynchronous data provides some ideas when it comes to coping with other difficult mathematical questions that may arise in finance. When two time references appear in a problem of financial mathematics, such as estimating the lead-lag parameter, an approach similar to the one used to estimate the realized covariance, may prove to be helpful. It is indeed be possible to adapt such an estimator to various contexts, where the existence of multiple time references plays a key role in the description of the issue.

References

[1] Nicole El Karoui and Emmanuel Gobet. *Les outils stochastiques des marchés financiers*. Editions de l'école Polytechnique, 2015.

[2] T.W. Epps. Comovements in stock prices in the very short run. *American Statistical Association*, 1979.

[3] T. Hayashi and N. Yoshida. On covariance estimation of non-synchronously observed diffusion processes. *Bernoulli*, 2005.

[4] M. Rosenbaum, M. Hoffmann, and N. Yoshida. Estimation of the lead-lag parameter from non-synchronous data. *Bernoulli*, 2013.

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