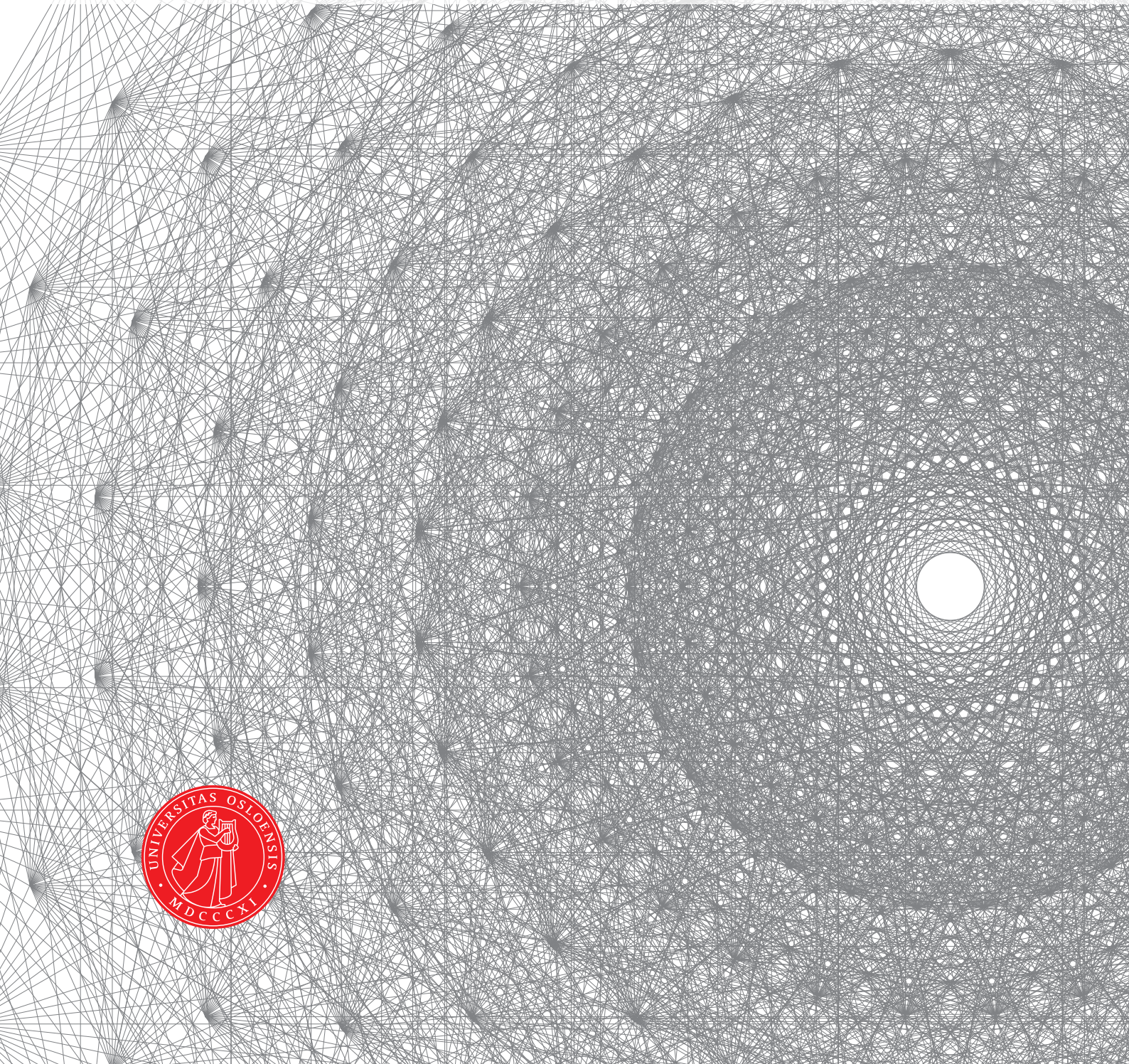
UiO : Department of Mathematics
University of Oslo

Stochastic Volatility and High-Frequency Financial Data

Summer Internship, 2023

Lucas Haubert
École Nationale Supérieure des Mines de Saint-Étienne



STOCHASTIC VOLATILITY AND HIGH-FREQUENCY FINANCIAL DATA

LUCAS HAUBERT

SUMMER INTERNSHIP REPORT

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF OSLO
NORWAY
2023

Series of dissertations submitted to the
Faculty of Mathematics and Natural Sciences, University of Oslo
Special edition

Without ISSN number

Cover: University of Oslo
Print production: University of Oslo (if printed)

Acknowledgements

First and foremost, I would like to express my wholehearted appreciation and gratitude to my supervisor Prof. Giulia di Nunno for her persistent and unceasing support during my Erasmus exchange semester and my research internship at the University of Oslo. Studying and working in Oslo would not have crossed my mind just a few months ago. In less than a year, I have been able to experience life as a resident of Oslo, discover a new working environment, practice English and Norwegian with my friends and learn a lot more. I was able, with Mrs. di Nunno, to study Stochastic Analysis and Stochastic Differential Equations during Fall 2022, before applying my knowledge in a research internship during Summer 2023 to the study of Stochastic Volatility using High-Frequency Financial Data. This human, academic and professional experience have played a key role in my orientation and the construction of my profile.

Next, my special thanks goes to my second tutor, Assoc. Anton Yurchenko-Tytarenko, for his guidance in programming and his support in my understanding of the mathematical concepts underlying the study of volatility. I really appreciated his accessibility and pedagogy. It was a pleasure to work with him during my research internship. My gratitude also goes to all of the researchers in the Department of Mathematics of the University of Oslo with whom I have been able to exchange about mathematics, its applications and many other things.

I would also like to thank my family and my friends who encouraged me in my work, not only during my internship and semester at the University of Oslo, but since the beginning of my studies in France. I am convinced that without them, especially my father and my mother, I would not have been able to pursue such a prestigious academic path, which now allows me to steer myself in a direction that fully satisfies me.

Finally, I would like to acknowledge my school in France, the École Nationale Supérieure des Mines de Saint-Étienne, for sponsoring this study and this internship. Without this scholarship, I would only be dreaming to study and practice a research internship abroad, in such an expensive country. Also, thanks to the University of Oslo, especially the Department of Mathematics, for all facilities provided and making the journey in Norway possible.

Lucas Haubert
Oslo, August 10, 2023

Contents

Executive Summary

In finance, it is common to work on a tool called *volatility* in order to best describe the behaviour of the markets. Volatility is an instrument that measures the extent and speed of price changes over a given period. Often written σ, it features in a large number of equations in financial mathematics. For over a century, several models have emerged to describe volatility. Starting, as is often the case in science, with simple models, the description of σ has expanded over time, until the arrival of the *stochastic volatility* models that are the subject of this research paper.

Modern approaches in quantitative finance involve what is called *high-frequency financial data* (HFFD). It refers to a type of financial market data that is recorded and updated at an extremely rapid pace. HFFD capture the micro-level dynamics of trading activities and market fluctuations, providing insights into the rapid changes. For all these reasons, this type of data is widely used in finance, particularly to work on stochastic volatility models. For instance, HFFD can be used to answer the following question:

Can volatility be written as a function of the price of the underlying asset?

This question concerns the test of the *Markovianity* hypothesis. Jean Jacod, Professor at Sorbonne University, and Yacine Aït-Sahalia, Professor at Princeton University, have worked on the subject to develop a statistical test to answer this question. The aim of this paper is therefore to implement the Markovianity test, in order to study various financial products. To this end, a preliminary study of volatility models and simple tools such as quadratic variation are necessary before building the mathematical framework of the two researchers' test.

***

Chapter 1 first focuses on the definition of volatility and stochastic volatility in order to consider relevant models. A detailed study of each leading model, along with their advantages and disadvantages, provides the basis for the representation of σ for the rest of this paper.

Using the knowledge acquired in the first part, Chapter 2 proposes an initial study of stochastic volatility using quadratic variation. This chapter is relevant in that it refines the models used to describe volatility, in addition to providing graphical representations of this coefficient. Convergence properties involving quadratic variation and the continuous semi-martingale model are then established. Despite its good approximations, the model of continuous semi-martingales needs to be extended to take account of two decisive phenomena: *micro-structure noise* and *jumps*. This chapter thus introduces the challenges involving such elements.

Chapter 3 brings together the key elements of the mathematical framework of Jacod and Aït-Sahalia's Markovianity test. The features of the observations of HFFD are first introduced, before to settle the construction of the test that is based on local times, estimators and convergence results. The results then come from a test statistic T_n, whose value is used to reject or validate the Markovianity hypothesis.

Finally, Chapter 4 presents the results of the test in several tables, in order to answer the question above. These tests are performed using HFFD, which the preparation process is described in Appendix A. To carry out the tests, particularly when considering micro-structure noise, several methods can be used in order to compute T_n. These methods are detailed in Appendix B.

CHAPTER 1

What is stochastic volatility and why is it used in finance ?

In science, the study of random phenomena requires knowledge of stochastic analysis. Stochastic analysis is an extension of probability theory, aimed at studying time-dependent random phenomena. In physics, it may be desirable to model the movement of a "large" particle immersed in a liquid and subject to no interaction other than shocks with the "small" molecules of the surrounding fluid. The position of the particle is then described by a stochastic process, i.e. a datum that depends on time and the experimental scenario. In this case, such a process is called a Brownian motion. In finance, it may be also desirable to model random phenomena. To optimize a portfolio or find a strategy to minimize risk, it is often necessary to look at the variation in the values of different financial products, described by such stochastic processes. In this respect, volatility is an instrument that measures the extent and speed of price changes over a given period. Often written as σ, it features in a large number of equations in financial mathematics. The aim of this first chapter is to describe the historical models involving σ, in order to set out the theoretical framework on which this research paper is based.

1.1 Deterministic models

Financial mathematics is a recent branch of applied mathematics. In 1900, French mathematician Louis Bachelier suggested an initial model of the price of a share as part of his doctoral thesis: "*Théorie de la spéculation*" [3]. He first suggested to model stock price as some random process $S = \{S(t), t \in [0, T]\}$. After analysing empirical data from the Paris stock market, he came to the conclusion that increments $S(t+\Delta) - S(t)$ have means that are close to zero and standard deviations of the order $\sqrt{\Delta}$. By translating this property into modern mathematical language, the price of the share is given by the following formula:

$$S(t) = S(0) + \sigma W(t) \tag{1.1}$$

with $S(0)$ being the initial price of the stock, $\sigma > 0$, and $W = \{W(t), t \in [0, T]\}$ being a standard Brownian motion (the same, in one dimension, that describes the movement of the particle in physics as presented above).

It was Samuelson who proposed a simple but very important modification of Bachelier's approach: he used (1.1) to model price logarithms and not the prices themselves. That solved the most obvious problem: a Brownian motion is a Gaussian process and hence can take negative values with positive probability whereas stock prices are inherently non-negative. After a small adjustment with a linear trend, Samuelson's model took the form of a geometric or "relative economic" (the term used by Samuelson himself) Brownian motion:

$$S(t) = S(0) \exp\left((\mu - \frac{\sigma^2}{2})t + \sigma W(t)\right) \tag{1.2}$$

with $\mu \in \mathbb{R}$ and $S(0)$, σ, $W(t)$ defined as in (1.1). The representation as a stochastic differential equation is given by (1.3):

$$dS(t) = \mu S(t)dt + \sigma S(t)dW(t). \tag{1.3}$$

This log-normal process (1.2)–(1.3) subsequently became a mainstream choice for stock price models for the next couple of decades. Even now, when there are multiple arguments against the geometric Brownian motion, practitioners still use it as a benchmark model or a good "first approximation".

It is interesting to note that, in addition to the market model, Bachelier also considered the problem of option pricing and eventually derived an expression that can be called a harbinger of the now famous Black-Scholes formula. Of course, his reasoning was not based on the no-arbitrage principle and had a number of shortcomings inherent in any pioneering work. The correction of those shortcomings became the subject of a number of studies in the 60s. Samuelson himself also heavily contributed to that topic with Robert Merton [19] where they suggested to consider a warrant/option payoff as a function of the price of the underlying asset, and it can be said that these works were only a few steps away from the real breakthrough made by Black, Scholes and Merton himself just a few months later.

In 1973, two revolutionary papers appeared: "*The pricing of options and corporate liabilities*" [4] by Fischer Black and Myron Scholes and "*Theory of rational option pricing*" [18] by Robert Merton. The main result of Black, Scholes and Merton can be formulated as follows: if a stock follows the model (1.2)–(1.3), then, under some assumptions, the discounted no-arbitrage price of a standard European call option V evolves as a function of the current time t and current price S and must satisfy a partial differential equation of the form:

$$\frac{\partial V}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} + rS\frac{\partial V}{\partial S} - rV = 0, \tag{1.4}$$

with a boundary condition:

$$V(T, S) = \max\{0, S - K\}, \tag{1.5}$$

where r denotes the instantaneous interest rate that is assumed to be constant, T is the maturity date of an option and K is its exercise price. Moreover, the equation (1.4)–(1.5) turns out to have an explicit solution of the form:

$$V(t, S(t)) = S(t)\Phi(d_+(t, S(t))) - Ke^{-r(T-t)}\Phi(d_-(t, S(t))), \tag{1.6}$$

where:

$$d_+(t, S(t)) := \frac{\log\frac{S(t)}{K} + (T-t)(r + \frac{\sigma^2}{2})}{\sigma\sqrt{T-t}},$$

$$d_-(t, S(t)) := \frac{\log\frac{S(t)}{K} + (T-t)(r - \frac{\sigma^2}{2})}{\sigma\sqrt{T-t}},$$

$$\Phi(x) := \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{y^2}{2}}\, dy.$$

The ideas of Black, Scholes and Merton revolutionized mathematical finance and enjoyed empirical success. However, this was only the beginning of a journey full of challenges. Indeed, the infamous "Black Monday" market crash happened on October 19, 1987.

The classical Black–Scholes–Merton model relies on a number of rather abstract assumptions that are not met on the real life market: specific dynamics for prices, no transaction costs, ability to buy and sell any amount of assets etc. However, being unable to reflect the reality perfectly is not always a big deal; after all, "all models are wrong, but some are useful" would have said the statistician George Box. In fact, Black, Scholes and Merton themselves were very well aware of this; and what really mattered was the successful empirical performance of the model, as a first approximation.

However, after the mentioned 1987 crash, it became crystal clear that something was very wrong with the log-normal paradigm, and something had to be done. In fact, on October 19, 1987, the two month S&P 500 futures price fell 29%. Under the log-normal hypothesis, this is a -27 standard deviation event with probability 10^{-160}, which represents an almost impossible scenario. Clearly, this almost impossible price which left thousands of investors destitute was already a good argument to reconsider financial modeling approaches. But, except for that shock (which in principle could be branded as a single anomaly), Black Monday brought something even more annoying from the theoretical perspective: the *volatility smile*.

As a tricky parameter, the volatility σ is the only parameter in the Black-Scholes formula (1.6) that is not observable. Maturity date T and exercise price K are given in specifications of the given option contract, the price $S(t)$ can be taken directly from the market, and σ has to be somehow "guessed" from the market data. In 1986, Latané & Rendelman [16] proposed an elegant method to do that. Fix some t together with the corresponding price $S(t)$ and consider the Black-Scholes option price (1.6) as a function $V_t := V_t(\tau, \kappa, \sigma)$ of the *log-moneyness* $\kappa = \log \frac{K}{S(t)}$, the *time to maturity* $\tau := T - t$ and the volatility σ.

Next, take the *actual market price* $\widetilde{V}_t$ of the corresponding option and notice that, since V_t is supposed to coincide with $\widetilde{V}_t$, the volatility σ can be found from the equation:

$$V_t(\tau, \kappa, \sigma) - \widetilde{V}_t = 0. \tag{1.7}$$

The solution $\hat{\sigma} = \hat{\sigma}_t(\tau, \kappa)$ to this equation is called the *implied volatility* and, if the stock model (1.2)–(1.3) indeed corresponds to reality well enough, $\hat{\sigma}_t(\tau, \kappa)$ should be approximately constant for options with the same underlying asset but differing maturities T and strikes K (and hence τ and κ). Unfortunately, this is not the case. For instance, $\hat{\sigma}_t(\tau, \kappa)$ turns out to change with τ for fixed κ. There seems to be an easy fix of (1.2) to account for this type of variation and, in fact, Merton actually considered such a modification in his original paper. Namely, if the volatility $\sigma = \sigma(t)$ is a deterministic function of time, one can obtain a version of (1.6) of the form:

$$V(t, S(t)) = S(t)\Phi(\bar{d}_+(t, S(t))) - Ke^{-r\tau}\Phi(\bar{d}_-(t, S(t))),$$

where:

$$\bar{d}_+(t, S(t)) := \frac{-\kappa + (r + \frac{1}{2}\bar{\sigma}^2(t,\tau)\tau}{\bar{\sigma}(t,\tau)\sqrt{\tau}},$$

$$\bar{d}_-(t, S(t)) := \frac{-\kappa + (r - \frac{1}{2}\bar{\sigma}^2(t,\tau)\tau}{\bar{\sigma}(t,\tau)\sqrt{\tau}},$$

$$\bar{\sigma}^2(t,\tau) := \frac{1}{\tau}\int_t^{t+\tau} \sigma^2(s)ds.$$

The counterpart of the equation (1.7) then takes the form: $V_t(\tau, \kappa, \bar{\sigma}(t,\tau)) - \widetilde{V}_t = 0$ and its solution $\hat{\sigma}_t(\tau, \kappa)$ is allowed to vary in τ for fixed κ. One may even argue that it is reasonable to assume that σ changes with time, as "there is nothing inconsistent about expecting high volatility this year and low volatility next year" as would say famous quant Emanuel Derman. As for the variation in κ for fixed τ, luckily, the implied volatility remained relatively flat (at least, the variation was subtle enough to be ignored), exactly until the above-mentioned Black Monday crash in 1987. Since that time, investors started observing notable variability of the implied volatility in κ with very clear convex patterns (see Fig. 1.1) which were eventually called "volatility smiles".

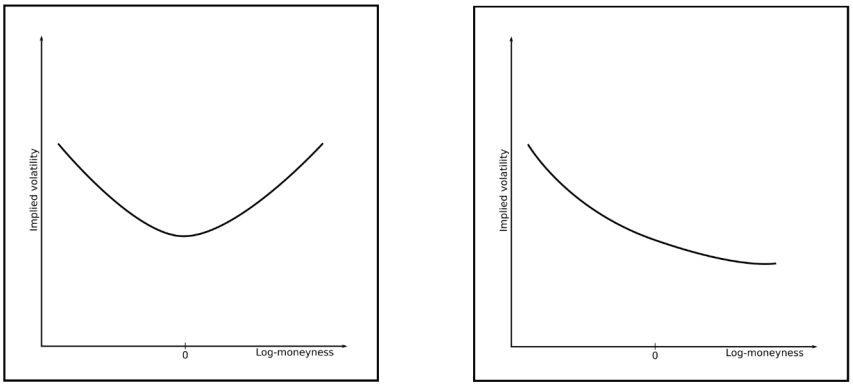


Figure 1.1: Idealised volatility smiles: the left figure represents the general form of volatility smiles for foreign currency options; the right one depicts a typical implied volatility smile for equity options

Such a behaviour was consistent, had a direct negative impact on empirical performance of the Black-Scholes formula and could not be explained by the price dynamics (1.2)-(1.3) - a very annoying combination for a theoretical framework. Moreover, one should not forget about the variability of implied volatility in τ : ideally, one would like a model that mimics the interplay between τ and κ, i.e. represents the behaviour of the *entire volatility surface* $(\tau, \kappa) \mapsto \hat{\sigma}_t(\tau, \kappa)$. This creates many additional effects to reproduce:

• First, the smile amplitude decreases very slowly as τ increases, as noted in [6] (see e.g. Fig. 1.2);

• Second, the observed *at-the-money volatility skew* defined as

$$\Psi(\tau) := |\frac{\partial}{\partial \kappa}\hat{\sigma}_t(\tau, \kappa)|_{\kappa=0} \tag{1.8}$$

is known to behave as $O(\tau^{-\beta})$ when $\tau \to 0$ (see e.g. Fig. 1.3).

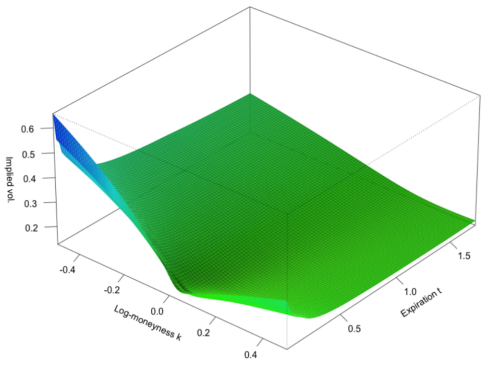


Figure 1.2: Shape of the S&P volatility surface

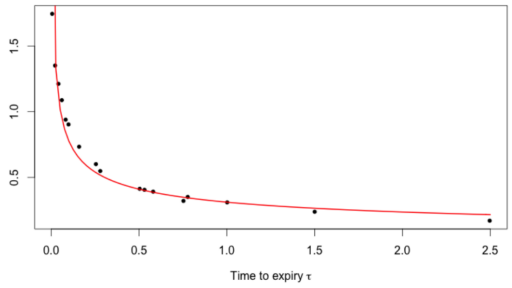


Figure 1.3: At-the-money volatility skew

Figure 1.2 is taken from [13], Figure 1.1 and represents the shape of the S&P volatility surface on June 20, 2013. Figure 1.3 is taken from [13], Figure 1.2. The dots represent estimates of the S&P volatility skews and the curve is the power-law fit $\Psi(\tau) = O(\tau^{-0.4})$.

In fact, it is not easy to conceive a model that reproduces jointly these two effects and the search for such a model is a a difficult puzzle for both theorists and practitioners.

As proved above, smile effect makes a spectacular point against the geometric Brownian motion (GBM) dynamics (1.2)–(1.3), but it is definitely not the only argument in place. In fact, objections to log-normality of prices appeared long before 1987, perhaps as early as the log-normal model itself. For instance, the empirical studies of Mandelbrot [17] (1963) and Fama [10] (1965) pointed out that tails of price distribution are much fatter than the ones expected from a log-normal random variable (in order to account for this, Mandelbrot suggested modeling price log-returns with α-stable distributions). Some other interesting phenomenon not grasped by the GBM is the so-called *leverage effect*: negative correlation between variance and returns of an asset. A third next empirical contradiction to the Black-Scholes-Merton framework is that any basic financial time series analysis reveals clusters of high and low volatility episodes. This clustering effect is often quantified by analyzing the autocorrelation function of absolute log-returns, i.e. $\text{corr}(|R(t)|, |R(t+\tau)|)$, where $R(t)$ is defined as $\log(\frac{S(t+\Delta)}{S(t)})$ for some given time scale Δ (which may vary between a fraction of a second for tick data to several days).

Of course, we cannot list all stylized facts about market behaviour contradicting the GBM dynamics due to the vast amount of material. Two things are clear though: first, the log-normal model is way too simple and does not reflect a lot of important qualitative features of the market and, second, the behaviour of financial time series is incredibly complex and requires fairly ingenious modeling approaches. It is then relevant to search for a model that is complex enough but easy to understand and work on. The class of *stochastic volatility models*, which is directly related to the framework of the internship, seems to meet our expectations.

1.2 Stochastic models

As noted in the previous section, the "trickiest" parameter of Black-Scholes formula is the volatility σ. Empirical observations show that it varies with time, is correlated with the current price level, has clusters of low and high values and seems to have a long memory. Another important phenomenon is the so-called excess volatility [7], [9]: the variability in asset prices cannot be fully explained only by changes in "fundamental" economic factors. All these stylized facts together lead to an idea to modify (1.3) as:

$$dS(t) = \mu S(t)dt + \sigma(t)S(t)dW(t) \tag{1.9}$$

with the volatility $\{\sigma(t), t \geq 0\}$ being a random process that is only imperfectly correlated with the Brownian motion W. This approach can be traced back to discrete-time model of Clark (1973) [5] where asset prices were considered as subordinated stochastic processes with the time change being used to represent trading volumes and information arrival. Early contributors to continuous time stochastic volatility modeling include Hull & White, Wiggins, Scott, Stein & Stein, Heston and others. Each of them proposed a precise differential notation for $\sigma(t)$. As examples, one can mention:

• Hull & White who assume that the squared volatility $\sigma^2 = \{\sigma^2(t), t \geq 0\}$ is itself a geometric Brownian motion, i.e. price and volatility satisfy stochastic differential equations of the form:

$$dS(t) = \mu S(t)dt + \sigma(t)S(t)dW(t)$$

$$d\sigma^2(t) = \theta_1\sigma^2(t)dt + \theta_2\sigma^2(t)dB(t)$$

respectively, where B and W are two Brownian motions that are allowed to be correlated to account for the leverage effect,

• Wiggins who suggests a slightly more general dynamics of the form:

$$dS(t) = \mu S(t)dt + \sigma(t)S(t)dW(t)$$

$$d\sigma(t) = f(\sigma(t))dt + \theta\sigma(t)dB(t)$$

where f is a given function on $\mathbb{R}$,

• Or Heston who introduces the SDE of the form:

$$dS(t) = \mu S(t)dt + \sqrt{\sigma(t)}S(t)dW(t)$$

$$d\sigma(t) = \theta_1(\theta_2 - \sigma(t))dt + \theta_3\sqrt{\sigma(t)}dB(t)$$

Stochastic volatility models turned out to have an additional important advantage: they have an ability to reproduce, to some extent, "smiley" patterns of the implied volatility. However, one must acknowledge that the models of the researchers listed above leave a lot of room for improvement when it comes to accuracy of grasping volatility surfaces (see e.g. [12] for a detailed overview of empirical performance of the classical stochastic volatility models). Therefore, there is no surprise that a lot of effort was made to advance the stochastic volatility framework further to account for all such inconsistencies.

Still on the subject of volatility modeling, many questions have arisen concerning the link between stochastic volatility and the price of the underlying asset (or the value of the financial product in general). The subject of this research paper falls squarely within the scope of these issues. It therefore makes sense to study several facets linked to variations in the value of financial products, in order to obtain information about the volatility coefficient σ. In finance, high-frequency financial data (HFFD) are used to study the values of such products, and hence their variations. The following sections are dedicated to the study of price variations of several assets/indexes using HFFD, in order to test the Markovianity hypothesis. As a way to work in the most general context from formula (1.9), the price of an asset (resp. the value of an index) is modeled as follows:

$$dS(t) = a(t)S(t)dt + \sigma(t)S(t)dW(t) \tag{1.10}$$

with a and σ being stochastic processes on the time line $[0, T]$ defined by the underlaying HFFD.

CHAPTER 2

Quadratic variation estimator: a first step in the study of volatility

The study of volatility is closely linked to the study of variations in the value of the underlying financial product. Indeed, it is an instrument that measures the extent and speed of price changes over a given period. In order to draw up a portrait of this instrument, the study of the quadratic variation of the process that describes the evolution of the asset price is a relevant first step. The aim of this section is then to exploit the quadratic variation tool to deduce, using HFFD, as much information as possible about the volatility of the financial products to be analyzed (AAPL, AMZN, DJIA and BTC).

2.1 Quadratic variation and continuous semi-martingales

As written before in (1.10), the price (OPEN price) S of the given financial assets/index is modeled as:

$$dS(t) = a(t)S(t)dt + \sigma(t)S(t)dW(t) \tag{2.1}$$

with a and σ being stochastic processes on the time line $[0, T]$ defined by the underlaying HFFD. This framework (that is called *continuous semi-martingale* model for $X = log(S)$) is useful when it comes to work with quadratic variations of S. In fact, the limited number of elements in a HFFD data set only allows to work with the *estimator of the quadratic variation* of S.

Let P be a stochastic process such that, for all $t \in [0, T]$,

$$P(t) = P(0) + \int_0^t \alpha(s)\, ds + \int_0^t \beta(s)\, dW(s)$$

and define the discrete times t_0, t_1, ..., t_n such that $t_0 = 0$, $t_n = T$ and $\forall i \in [\![0, n-1]\!], t_{i+1} - t_i = |\pi|$ with $|\pi|$ a constant. From papers of Y. Aït-Sahalia (as [1]), the estimator of the quadratic variation converges in probability to integrated β^2 when $n \to +\infty$ ($\iff$ $|\pi| \to 0^+$), i.e.:

$$\sum_{k=0}^{n-1} (P(t_{k+1}) - P(t_k))^2 \xrightarrow[n\to+\infty]{\mathbb{P}} \int_0^T \beta^2(s)\, ds \tag{2.2}$$

Moreover by means of Itô formula, it is possible to write (2.1) as:

$$S(t) = S(0) \exp\left(\int_0^t b(s)\, ds + \int_0^t \sigma(s)\, dW(s) \right), \text{ where } b(s) := a(s) - \frac{\sigma^2(s)}{2} \tag{2.3}$$

Hence defining $X := log(S)$ and applying (2.2) gives a capital relation between the *log-prices* and the volatility of the underlaying asset/index:

$$\sum_{k=0}^{n-1} (X(t_{k+1}) - X(t_k))^2 \xrightarrow[n\to+\infty]{\mathbb{P}} \int_0^T \sigma^2(s)\, ds \tag{2.4}$$

This convergence will be used to estimate the volatility (or its square integrated value). However, it would be naive to think that real markets obey such a simple representation. For this reason, the remainder of this section focuses on the precautions to be taken when studying quadratic variation.

First observe, because the convergence in (2.4) is given by $|\pi| \to 0^+$, that:

$$\sum_{k=0}^{\lfloor \frac{n}{2} \rfloor - 1} (X(t_{2i+2}) - X(t_{2i}))^2 \xrightarrow[n\to+\infty]{\mathbb{P}} \int_0^T \sigma^2(s)\, ds$$

$$\sum_{k=0}^{\lfloor \frac{n}{2} \rfloor - 1} (X(t_{2i+3}) - X(t_{2i+1}))^2 \xrightarrow[n\to+\infty]{\mathbb{P}} \int_0^T \sigma^2(s)\, ds$$

Hence, the mean of these two sums also converges to the integrated volatility. It is then possible to generalize this property. For any integer $m \geq 1$ and $k \in [\![0, m-1]\!]$, the following result is given by the previous reasoning:

$$V_m := \frac{1}{m} \sum_{k=0}^{m-1} \sum_{i=0}^{\lfloor \frac{n}{m} \rfloor - m} (X(t_{(i+1)m+k}) - X(t_{im+k}))^2 \xrightarrow[n\to+\infty]{\mathbb{P}} \int_0^T \sigma^2(s)\, ds \tag{2.5}$$

Note: If the partition $(t_0, t_1, ..., t_n)$ is fine enough and the model is adequate, then we should have $V_1 \approx V_2 \approx ... \approx V_m$ for a given m. However, the plots of the V_m's from the different HFFD data sets (AAPL, AMZN, DJIA and BTC) are giving an unexpected result:

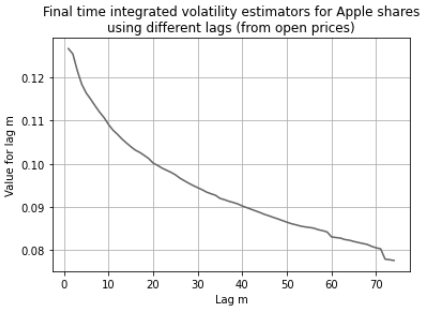


Figure 2.1: AAPL values for V_m

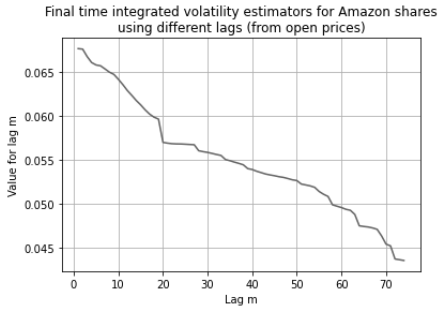


Figure 2.2: AMZN values for V_m

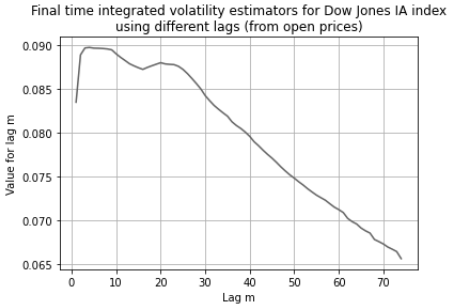


Figure 2.3: DJIA values for V_m

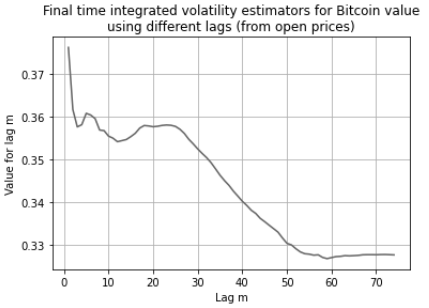


Figure 2.4: BTC values for V_m

These plots can be used to argue against the semi-martingality of X. In the literature, this phenomenon is often explained by means of the *micro-structure noise*: according to this paradigm, we do not observe the values of semi-martingale X but only:

$$Y(t_k) = X(t_k) + \eta_{t_k} \tag{2.6}$$

where η is some contaminating micro-structure noise that appears due to bid-ask bounce, discontinuity of price changes etc. It is important to note that micro-structure noise (that will be written "MN" in the rest of the paper) manifests itself in the high-frequency domain.

It is relevant to explain, in an analytic way, why it is necessary to consider time lags bigger than 1 minute (from 15 to 30 minutes) in general in order to get rid of the effects of MN. As a reminder, one observe the data as in (2.6), then the estimator of the quadratic variation is given by:

$$\sum_{k=0}^{n-1}(Y(t_{k+1})-Y(t_k))^2=\sum_{k=0}^{n-1}(X(t_{k+1})-X(t_k))^2+2\sum_{k=0}^{n-1}(X(t_{k+1})-X(t_k))(\eta_{t_{k+1}}-\eta_{t_k})+\sum_{k=0}^{n-1}(\eta_{t_{k+1}}-\eta_{t_k})^2$$

The converging direction is clearly given by the first term, while η is "contaminating" the others. The second term can be managed by assuming $\mathbb{E}(\eta_{t_k})=0$ for all k and (η_{t_k}) being a family of independent random variable (which is not absurd when considering MN). Then its expectation is 0. In addition, if v^2 is the variance of each η_{t_k}, the expectation of the last term is given by $2v^2n$ which obviously goes to $+\infty$ with n. On the one hand, it is therefore preferable not to take n to big (in order to avoid MN effects), but on the other, the convergence of the first term in the equation above suggests to take n big enough.

This reasoning leads to the following conclusion: the best way to estimate the quadratic variation (with respect with X), and then estimate the integrated volatility, is to deal with a time lag m between each data record that respects such an equilibrium. Figures 2.1, 2.2, 2.3 and 2.4 gets stabilized around $m=20$. That corresponds to the best estimation of the integrated volatility at final time.

2.2 Estimating the integrated volatility with HFFD

The idea of relation (2.5) can be used to work with the square integrated volatility term σ. Indeed, let's consider the indexes $k_t\in[\![0,m-1]\!]$ and $i_t\in[\![0,\lfloor\frac{n}{m}\rfloor-m]\!]$ such that $t_{(i+1)m+k}$ (with $m=20$) is the last time in $(t_0,t_1,...,t_n)$ before t. Then the following approximation holds:

$$\frac{1}{m}\sum_{k=0}^{k_t}\sum_{i=0}^{i_t}(X(t_{(i+1)m+k})-X(t_{im+k}))^2\approx\int_0^t\sigma^2(s)\,ds \tag{2.7}$$

It is then possible to implement (2.7) into a program to construct the histories of such an integral, from time 0 to T, that is defined according to each asset/index HFFD:

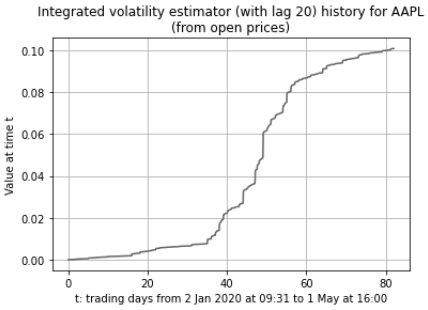


Figure 2.5: AAPL integrated σ^2 estimation

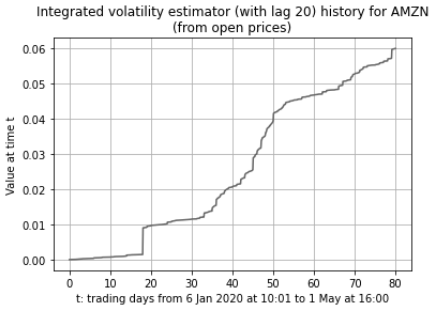


Figure 2.6: AMZN integrated σ^2 estimation

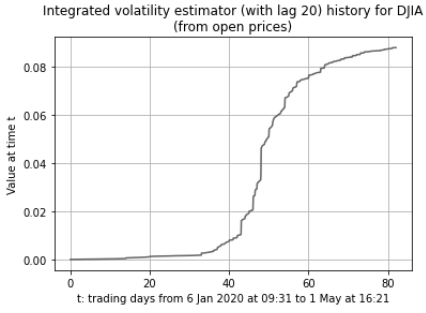


Figure 2.7: DJIA integrated σ^2 estimation

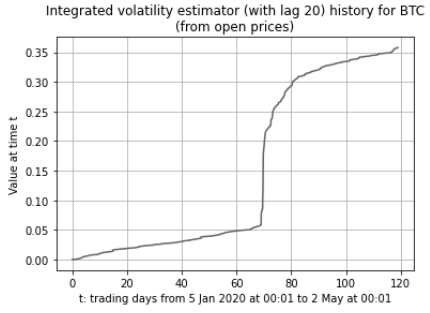


Figure 2.8: BTC integrated σ^2 estimation

Figures 2.5, 2.6, 2.7 and 2.8 provide relevant information about the variations of the log-prices processes, an then the coefficient σ. Basic notions of analysis can be used to evaluate the behavior of this coefficient. For example, a clear increase in σ^2 (and therefore in estimated volatility) can be observed a little over a month after the start of January. All practitioners agree that this is due to the Covid-19 crisis that hit the world at the start of 2020.

Because of the integration/derivation link between integrated volatility and the σ coefficient, it is also possible to estimate historical values for the latter. This is why the method of integrated volatility is used. For instance, Figure 2.6 provides a great deal of information about Amazon business at the end of January 2020. Due to the a noticeable *jump* in the values of the integrated volatility, it shows that Amazon surely experienced a major macro-economic event at that time. Indeed, the company financial report [2], written in late January 2020, about the last Quarter of 2019. The company has announced very positive results regarding its business momentum for the end of 2019, and therefore the months ahead. In the eyes of investors, this created an emulation that led to a sharp rise in Amazon share prices, as well as a jump in estimated volatility. Figures 2.9 and 2.10, that represent AMZN prices and AMZN integrated volatility over the same period of time, illustrate this event perfectly:

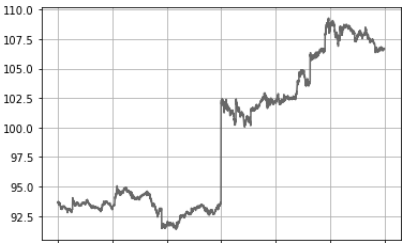


Figure 2.9: AMZN prices

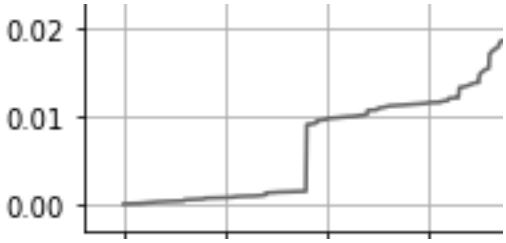


Figure 2.10: AMZN integrated volatility

Clearly, there is a correlation between asset price and volatility over this time window. This observation is also supported by the macro-economic analysis above. It then makes sense to wonder whether there is a simple functional relationship between the observed asset price S and the volatility coefficient σ. With a few assumptions, this hypothesis is called *Markovianity*. Chapters 3 and 4 of this paper focus on the study of this hypothesis.

However, Markovianity hypothesis is not the only inspiration induced by bringing together the two graphs above. In science, it makes sense to question the results obtained from an experiment based on a certain model. Indeed, volatility σ, as defined in a continuous semi-martingale framework (2.1)-(2.3), makes a significant jump at the same time as the explosion in Amazon share prices. This leads to the following question: what to represent using the coefficient σ? Being strongly linked to Brownian motion W, the volatility coefficient acts as an amplifier for the action of W. However, the latter is well-suited to modeling permanent chaotic events in financial markets, which very rarely induce such significant jumps in value for S. It therefore seems advisable to attribute such jumps to a phenomenon other than the volatility-Brownian motion coupling. For this reason, there is a class of stochastic processes that are useful to model such variations: *jump processes*.

Over the past decades, stochastic processes with jumps have become increasingly popular for modeling market fluctuations, both for risk management and option pricing. For instance, R. Cont and P. Tankov have written a whole book [8] about such processes and their applications in finance. On the occasion of an international conference [14], J. Jacod gives a definition for such processes. A stochastic process $J = \{J(t), t \in [0, T]\}$ is said to be a jump process if it can be written as:

$$J(t) = \sum_{i \geq 1} \Gamma_i 1_{\{T_i \leq t\}} \tag{2.8}$$

where Γ_i is $\mathscr{F}_{T_i}$-measurable and (T_i) is a sequence of stopping times such that $T_i \to +\infty$. Such a tool now makes the jump in value observed for S on the AMZN asset more realistic. In such an alternative situation, the coefficient σ would be uncorrelated from this jump. The rest of this chapter presents the consequences of implementing jump processes in the continuous semi-martingale model on what the quadratic variation of X represents.

2.3 Jump processes and quadratic variation

As indicated in the end of the previous section, the choice of the model must be taken into account when analyzing the results of the resulting experiments. Considering jump processes J, as in (2.8), to express S is equivalent to write (2.3) as the following:

$$S(t) = S(0) \exp\left(\int_0^t b(s)\, ds + \int_0^t \sigma(s)\, dW(s) + \sum_{i \geq 1} \Gamma_i 1_{\{T_i \leq t\}} \right) \tag{2.9}$$

In a paper [11] written by J. Fan and Y. Wang, (2.4) is then changed as follows:

$$\sum_{k=0}^{k_t} (X(t_{k+1}) - X(t_k))^2 \xrightarrow[n \to +\infty]{\mathbb{P}} \int_0^t \sigma^2(s)\, ds + \sum_{i=1}^{i_t} \Gamma_i^2 \tag{2.10}$$

A second term relating to the jump process is therefore added to the integrated volatility as a result of the convergence of the quadratic variation of X. This implies, as indicated in the previous section, that the volatility coefficient σ plays a different role in the description of the quadratic variation. Considering σ as an amplifier for the action of W, the situation presented in Figures 2.9 and 2.10 can be explained using the new term introduced in (2.10).

It is now useful to deal with the consequences of the choice of this model in order to analyze σ better. First, it should be pointed out that volatility is always studied in the context of the presence of MN. Consequently, one should exploit (2.10) with a lag $m = 20$ as before, in order to best estimate the effective convergence result. Hence, the following approximation, inspired from (2.7) holds:

$$\frac{1}{m} \sum_{k=0}^{k_t} \sum_{i=0}^{i_t} (X(t_{(i+1)m+k}) - X(t_{im+k}))^2 \approx \int_0^t \sigma^2(s)\, ds + \sum_{i=1}^{i_t} \Gamma_i^2$$

Second, since quadratic variation no longer represents integrated volatility, few adjustments are necessary in order to work on either the continuous sum (that is relative to σ,) or the discrete one (that is relative to the jump process J). This is the problem Fan and Wang's paper addresses. In order to estimate first the jump variation (second term in the convergence in (2.10)), they: "first apply some wavelet method to the observed HFFD and locate all jumps in the sample path of the log-process X=log(S) and then use the estimated jump locations to estimate jump size for each estimated jump. Finally the jump variation is estimated by the sum of squares of all estimated jump sizes". Then, they can estimate the integrated volatility by getting rid of the jump variation.

The study of the integration of jump processes in the model of S leads to the following conclusion: the analysis of the results obtained after the same calculation procedure can vary according to the chosen representation of S. In the previous situation, the convergence results respects the following change, when one chose to write S with a jump process J:

$$\int_0^t \sigma^2(s)\, ds \mapsto \int_0^t \sigma^2(s)\, ds + \sum_{i=1}^{i_t} \Gamma_i^2$$

This may imply the need to make certain adjustments, as explained above. Fortunately, this problem no longer arises in the remainder of this paper, since J. Jacod and Y. Aït-Sahalia, in constructing their test of the Markovianity hypothesis, have taken into account the possible presence of jumps in the values of the assets/indices considered. By defining a jump truncation threshold v, skilfully chosen to distinguish what is a jump process from what is a volatility-Brownian coupling effect, the jump action is cancelled, which allows the test user to focus on the coefficient σ.

CHAPTER 3

The mathematical framework of J. Jacod and Y. Aït-Sahalia's Markovianity test

Assuming the continuous semi-martingale model, Figures 2.9 and 2.10 show a clear correlation between the volatility coefficient σ and the price of the underlying asset. The question then arises of the existence of a functional relationship between σ and S (or X). This is followed by a reflection on what this coefficient represents, and the role of the hypothetical presence of jump processes, defined in equation (2.8). This chapter therefore focuses on the study of the volatility in such an environment, in order to test what is called *Markovianity* hypothesis.

3.1 Can log-prices be described by essentially markovian processes?

It is first necessary to define Markovianity hypothesis. For this, J. Jacod and Y. Aït-Sahalia (who will be called by "J and AS" in the rest of the paper) present in [14] and [15] the mathematical framework associated with the study of such a hypothesis. Inspired by (2.9), they consider the log-price process X as in the following. For all $t \in [0, T]$, consider:

$$X(t) = X(0) + \int_0^t b(s)\, ds + \int_0^t \sqrt{c(s)}\, dW(s) + J(t) \tag{3.1}$$

such that b and c are two stochastic processes on $[0, T]$ and J is a jump process, defined as in (2.8), with finitely many jumps on every finite time interval. In fact, b the represents what is called the *drift* of X, and $\sqrt{c}$ represents the *volatility* coefficient σ. According to this framework, Markovianity can therefore be defined as follows:

$$\text{There exists some positive function } a \in \mathscr{C}^1(\mathbb{R}, \mathbb{R}) \text{ such that } c(t) = a(X(t)) \tag{H_0}$$

This hypothesis is called *local volatility*, *null hypothesis* or also *Markovianity*. Yet it is important to distinguish between a process that obeys (H_0) and a Markov process. In fact, Markov property considers the null hypothesis, but also a similar functional relationship between the drift b and the price and some special structure for J. Then if X is a Markov process, it can be written as:

$$X(t) = X(0) + \int_0^t \mu(X(s))\, ds + \int_0^t a(X(s))\, dW(s) + J(t)$$

However, the available HFFD allows to study model (3.1) according to one single scenario ω, which makes the characterization $b(t) = \mu(X(t))$ impossible (see more details in [15]). In order to keep the same underlying idea, without considering this last detail, the expression "Markovianity" therefore refers to (H_0).

An additional remark may be made on the choice of writing the model by J and AS. The volatility coefficient σ is not written in (3.1), but is represented by $\sqrt{c}$ for some given process c. In fact, this does not change the nature of the test, since c and σ are positive processes, and the square root is a bijection of $\mathbb{R}^+$. This is a notation chosen by J and AS that subsequently simplifies mathematical writing in the following. Using the same reasoning and using the bijections square-root/square function and exponential-logarithm, the relation in (H_0) can be written as:

$$c = a_1(X) \iff c = a_2(S) \iff \sigma = a_3(X) \iff \sigma = a_4(S) \text{ for some } a_1, a_2, a_3, a_4 \in \mathscr{C}^1(\mathbb{R}, \mathbb{R})$$

In order to set up a statistical test that aims to validate or reject the null (hypothesis), several features of the observations must be noticed:

• **A finite horizon:** X is observed over a finite time interval, which is taken to $[0,T]=[0,1]$. Hence, (H_0) can not be tested and gives way to (H_0') defined as follows:

There exists some positive function $a \in \mathscr{C}^1(\mathbb{R},\mathbb{R})$ such that $c(t)=a(X(t))$ for all $t\in[0,1]$ (H_0')

• **A single path:** The HFFD allow to study only one path ω for each asset/index (this is why the drift can not be characterized).

• **Discrete observations:** Only values $X(i\Delta_n)$ are observed, for $i=0,1,...,n$ and regular sampling is considered, in order to make the construction of the test easier. Hence, Δ_n is equal to $\frac{1}{n}$ as the time horizon is given by $[0,1]$.

• **Possible noisy observations:** In many cases, the data $Y(i\Delta_n)$ are observed, where $Y(t)=X(t)+\eta_t$, with η being the so called *micro-structure noise* (MN) as introduced in (2.6).

• **High-frequency setting:** For a fixed time lag Δ_n, nothing can be said about σ. It is necessary to consider $\Delta_n \to 0$.

These features are giving relevant information about what is observed, and then who to process with HFFD.

The setting of the test also requires some assumptions concerning the process X and the MN η in order to exploit some good mathematical properties:

For the process: On some filtered space $(\Omega,\mathscr{F},(\mathscr{F}_t),\mathbb{P})$ with X written as in (3.1):

• b is locally bounded and adapted, c is càdlàg and adapted, and $(c(t))$ and $(c(t^-))$ do not vanish.

• $\mathbb{E}((c((t+s)\wedge\tau_m)-c(t\wedge\tau_m))^2|\mathscr{F}_t)\leq K_m s$ for a sequence $\tau_m\to+\infty$ of stopping times.

• Under (H_0), the function a does not vanish.

• J is adapted and and written as in (2.8). In fact, it is possible to assume that J is a pure jump semi-martingale, with jump activity strictly less than 1. In that case, the same procedure holds, up to adjusting the so called *jump threshold* v (see Section 3.3) to the jump activity index.

For the MN:

• Conditionally on $\mathscr{F}_\infty$, the variables η_t are independent and centered (which corresponds to the assumption in the analysis of the quadratic variation with MN in page 10).

• For all $p\geq 0$, the process $\mathbb{E}(|\eta_t|^p|\mathscr{F}_\infty)$ is pre-locally bounded.

• For all t, $\mathscr{L}(\eta_t|\mathscr{F}_t)=\mathscr{L}(\eta_t|\mathscr{F}_\infty)$ (equality in conditional laws).

Note that in finance, a key feature is the *rounding effect*. Consider that prices are written in cents, then the observed value $Y(t)$ is a sort of rounded value of $X(t)$, and $\eta_t=Y(t)-X(t)$. It was shown by J and AS that η therefore does not satisfy the required assumptions. However, there exists a model satisfying the above and incorporating rounding: let (ϵ_t) be a family of i.i.d. variables, independent of $\mathscr{F}_\infty$ and uniform on $[0,1]$. Then:

$$Y(t)=\lfloor X(t)\rfloor \text{ if } \eta_t\geq X(t)-\lfloor X(t)\rfloor, \text{ and } Y(t)=\lfloor X(t)\rfloor \text{ otherwise}$$

satisfies the assumptions.

Nota Bene: It is important to note that these features and assumptions do not bind the user of the statistical test, since they are necessary conditions for the construction of it. The content of this page is therefore used for the proper functioning of the Markovianity test which is presented in the Section 3.2.

3.2 The Markovianity test

Markovianity test is a complex test, which takes into account several parameters and relies on many intermediate results. It is therefore reasonable to gradually describe the construction of the test statistic and the decision paradigms.

***

Since the object of the paper is to work on (H_0) (or rather (H_0')), it is first necessary to establish an alternative hypothesis of the null. It seems natural to take:

$$\text{The property: "}\exists a \in \mathscr{C}^1(\mathbb{R},\mathbb{R}) \text{ positive such that } c(t) = a(X(t)) \text{ for all } t \in [0,1]\text{" fails} \qquad (H_1)$$

However, a tractable alternative should be such that, if one has "full information" (i.e. if the function $t \mapsto X(t)$ is perfectly known on $[0,1]$) then one can decide which one of the null or the alternative hypothesis holds. This is not true with (H_1) because only one path is observed, so if $R = \{X(t), t \in [0,1]\}$ is the range of the process on $[0,1]$, then it is possible that $c(t) = a(X(t))$ for $X(t) \in R$, but not necessarily otherwise. Then the best one can do is to use the following *random* alternative hypothesis:

$$\text{Given } \omega,\ \exists x \in R(\omega),\ \exists (s,t) \in [0,1]^2\text{: } X(s,\omega) = X(t,\omega) = x \text{ and } c(s,\omega) \neq c(t,\omega) \qquad (\Omega_1)$$

***

A key ingredient of the test is the *local time* L_t^x of X at each level x. It is defined as a semi-martingale:

$$L_t^x = |X(t)-x| - |X(0)-x| + \int_0^t sign(X(s)-x)\,(dX(s)-dJ(s)) - \sum_{i\geq 1}(|X(T_i)-x| - |X(T_i^-)-x|)1_{\{T_i\leq t\}}$$

where (T_i) is the sequence of stopping times in the definition (2.8) of the jump process J. The process (L_t^x) admits a version that is continuous in (x,t), and the support of the measure dL_t^x is the set $\{t : X(t) = X(t^-) = x\}$.

Local times are used to construct the elements U and S as defined below:

$$U(x,p) := \int_0^1 c(s)^{\frac{p}{2}-1}\, dL_s^x, \text{ for } p \in \mathscr{P} := \{0,2,4\} \text{ and } x \in \mathbb{R} \qquad (3.2)$$

$$S^x := \int_0^1 \int_0^1 (c(t)-c(s))^2 c(t)^{-1} c(s)^{-1}\, dL_t^x\, dL_s^x = 2(U(x,0)U(x,4) - U(x,2)^2), \text{ for } x \in \mathbb{R} \qquad (3.3)$$

such that: under the null (H_0), $S^x = 0$ for all x, and under the alternative (Ω_1), $S^x > 0$ for at least one x. Indeed, local times are first built to detect when $X(t) = X(t^-) = x$ some some (x,t), then under the null, $(c(t)-c(s))^2 = 0$ when the measures dL_t^x and dL_s^x are not equal to 0, hence $S^x = 0$ for all x. Else, because of the continuity property of the local time version, $S^x > 0$ for at least one x.

An idea for testing (H_0') would then be to use the integral of S^x on $\mathbb{R}$ (that is well defined because $x \mapsto S^x$ vanishes outside a compact set) defined below:

$$S := \int_{\mathbb{R}} S^x\, dx \text{ (by using the estimators } \hat{S}_n := \int_{\mathbb{R}} \hat{S}_n^x\, dx)$$

However, in order to construct the test, a Central Limit Theorem (CLT) is needed for $(\hat{S}_n)$, at least under the null. Now, a CLT holds for $(\hat{U}(x,p)_n)$, but the limiting process (indexed by x) in this CLT has a structure of a white noise, so it does not translate into a CLT for $(\hat{S}_n)$. It is then necessary to replace (Ω_1) by the following restrictive alternative, for any given finite subset $\mathscr{X}$ of $\mathbb{R}$:

$$\text{Given } \omega,\ \exists x \in \mathscr{X},\ \exists (s,t) \in [0,1]^2\text{: } X(s,\omega) = X(t,\omega) = x \text{ and } c(s,\omega) \neq c(t,\omega) \qquad (\Omega_1(\mathscr{X}))$$

In practice R is known and $\mathscr{X}$ is a regular grid $\{\alpha + i\beta : i = 0,1,...,N\}$ such that $R \subset [\alpha, \alpha + N\beta]$ and β is small. Since $x \mapsto S^x$ is continuous, ($\Omega_1(\mathscr{X})$) is "almost" the same as (Ω_1).

J and AS first present the construction of the test statistic in case of absence of MN. This is the simplest version and it subsequently serves as a common thread in the case of the presence of MN.

***

First consider a non-negative $\mathscr{C}^1$ *kernel function* f on $\mathbb{R}$ with support in $[-1,1]$ with Lebesgue integral equal to 1. Then take a sequence of *bandwidths* $h_n > 0$ and a sequence of *jump truncation thresholds* $v_n > 0$ such that, for some $\epsilon > 0$,

$$\frac{h_n^3}{\Delta_n} \to 0, \quad \frac{\Delta_n}{h_n^2} \to 0, \quad v_n \to 0 \quad \text{and} \quad \frac{\Delta_n^{1/2-\epsilon}}{v_n} \to 0$$

It means that the bandwidths h_n tend to 0, but "not to fast". Moreover, the thresholds v_n are built to get rid of hypothetical jumps (see (3.4)). The sequence (h_n) is then used to construct the approximation of the Dirac mass of f at 0:

$$f_n(x) = \frac{1}{h_n} f(\frac{x}{h_n})$$

Considering the p^{th} absolute moment of a standard normal random variable m_p, it is possible to construct a sequence $(\hat{U}(x,p)_n)$ of estimators of $U(x,p)$, for a given $p \in \mathscr{P} := \{0,2,4\}$, as defined in (3.2):

$$\hat{U}(x,p)_n := \Delta_n^{1-\frac{p}{2}} \frac{1}{m_p} \sum_{i=1}^{n} f_n(X((i-1)\Delta_n) - x)|\Delta_i^n|^p 1_{\{|\Delta_i^n| \leq v_n\}} \tag{3.4}$$

where the discrete increments Δ_i^n are equal to the log-returns $X(i\Delta_n) - X((i-1)\Delta_n)$. It is then possible to construct the estimators $\hat{S}_n^x := 2(\hat{U}(x,0)_n\hat{U}(x,4)_n - \hat{U}(x,2)_n^2)$ by using (3.3). J and AS have shown in [14] that the convergence:

$$\hat{U}(x,p)_n \xrightarrow[n\to+\infty]{\mathbb{P}} U(x,p)$$

is indeed valid, which leads to the following of the test construction in the case of no MN presented in page 15.

***

Estimating S^x in the presence of MN first requires the introduction of an additional parameter: the *pre-averaging window size* k_n, which is an even integer and follows the rules below, for some $\epsilon > 0$:

$$k_n^5\Delta_n^3 \to +\infty, \quad \frac{h_n^3}{k_n\Delta_n} \to 0, \quad \frac{k_n\Delta_n}{h_n^{2+\epsilon}} \to 0, \quad v_n \to 0 \quad \text{and} \quad \frac{(k_n\Delta_n)^{1/2-\epsilon}}{v_n} \to 0$$

As indicated in the definition of k, the estimation of S^x, in the presence of MN, lies in the use of the *pre-averaging* method. Remember that in this situation, the process $Y = X + \eta$ is observed, for some micro-structure noise η. The first step is then to exploit the HFFD for Y, in order to construct its average over a given window, as defined below:

$$\bar{Y}_i^n := \frac{1}{k_n} \sum_{j=0}^{k_n-1} Y((i+j)\Delta_n) \tag{3.5}$$

The reason for using the pre-averaging method is as follows: the MN process η is such that (η_t) is a i.i.d. set of random variables and the η_t's are centered (see more details in page 12). Thus, considering the average of the observations of Y, according to a well-chosen pre-averaging window, makes it possible to get rid, on average, of MN effects. The idea of the estimation of S^x is then to implement the $\bar{Y}_i^n$'s, over a set of well-chosen windows, in formula (3.4), in order to construct the estimator of $U(x,p)$ in the "noisy" case. There are therefore several ways to construct such an estimator. For instance, J. Jacod suggests in [14] to define $\hat{Y}_i^n$ as follows:

$$\hat{Y}_i^n := \frac{1}{k_n^2} \sum_{j=0}^{2k_n-1} (Y((i+j)\Delta_n) - \bar{Y}_i^n)^2 \tag{3.6}$$

in order to give explicit expressions for $\hat{U}(x,0)_n$, $\hat{U}(x,2)_n$ and $\hat{U}(x,4)_n$ that depend on $\bar{Y}_i^n$'s and $\hat{Y}_i^n$'s.

Now that the main elements of test construction have been introduced, it is time to define the test statistic, whose value will be used to validate or reject (H_0) (or rather (H_0')). First define a weighted sum of S^x, for $x \in \mathscr{X}$ and $r \in \mathbb{R}$, namely:

$$\hat{\Phi}_n := \sum_{x\in\mathscr{X}} (\hat{U}(x,2)_n)^r \hat{S}^x_n, \text{ which converges in probability to } \Phi := \sum_{x\in\mathscr{X}} (U(x,2))^r \hat{S}^x$$

Under the null hypothesis, $\Phi = 0$ and the stable convergence holds:

$$\sqrt{(h_n/k_n\Delta_n)}\hat{\Phi}_n \to Z \tag{3.7}$$

where Z is a $\mathscr{F}$-conditionally centered Gaussian random variable, with variance:

$$\Sigma = \frac{8\beta}{3}\sum_{x\in\mathscr{X}} a(x)(L^x_1)^{3+2r}, \text{ with estimators given by } \hat{\Sigma}_n := \frac{8\beta}{3}\sum_{x\in\mathscr{X}} \hat{U}(x,0)_n(\hat{U}(x,4)_n)^2(\hat{U}(x,2)_n)^{2r}$$

As a reminder, a is the $\mathscr{C}^1$ function in (H_0') and β is a parameter defined as $\beta := \int_{-1}^{1} f^2(x)dx$.

It is now possible to define the *test statistic* T_n as follows:

$$T_n := \sqrt{h_n/k_n\Delta_n}\frac{\hat{\Phi}_n}{\sqrt{\hat{\Sigma}_n}} \tag{3.8}$$

The decision between the local volatility hypothesis (H_0') and the random alternative hypothesis ($\Omega_1(\mathscr{X})$) is made on the basis of the values taken by T_n, depending on the asset/index under consideration, the number of historical data n and the presence or absence of MN.

***

The principle for evaluating the test statistic is as follows: since $\hat{\Phi} \xrightarrow{\mathbb{P}} \Phi$, $\hat{\Sigma} \xrightarrow{\mathbb{P}} \Sigma$, given the convergence (3.7) and the shape of T_n in (3.8), the result above gives the main key of the Markovianity test:

$$T_n \xrightarrow{\mathscr{L}} N(0,1) \text{ under the null } (H_0'), \text{ and } T_n \xrightarrow{\mathbb{P}} +\infty \text{ under the alternative } (\Omega_1(\mathscr{X})) \tag{3.9}$$

Then, with $z_{1-\alpha/2}$ being the α-quantile of $N(0,1)$[1], it is possible to define the rejection region $C_n = \{T_n > z_{1-\alpha/2}\}$ of (H_0') with an error α, that also corresponds to the acceptation zone for ($\Omega_1(\mathscr{X})$). Hence, (3.9) can be translated as follows: The critical regions C_n have the asymptotic level $\alpha/2$ under the null, i.e. $\mathbb{P}(C_n) \to \alpha/2$. Moreover, the C_n's are asymptotically consistent with ($\Omega_1(\mathscr{X})$), meaning that: $\mathbb{P}((C_n)^c \cap \Omega_1(\mathscr{X})) \to 0$. It is then possible to draw up an overview of the test, as well as the characteristic areas of each hypothesis, in Figure 3.1. b/c (resp. a/c) represents the quantile $z_{1-\alpha/2}$ (resp. $-z_{1-\alpha/2}$) for a given α, such that the probability to get between a/c and b/c is $1-\alpha$. Hence, given the error α, C represents the critical zone for (H_0'), and B represents the validation zone for it. Under the alternative, $\Phi > 0$, hence it is expected that T_n is positive (at least). Therefore, zone A only represents the case where $T_n < -z_{1-\alpha/2}$, what happens with a low probability, and then does not represent a validation zone for the alternative. In fact, it only concerns very rare cases.

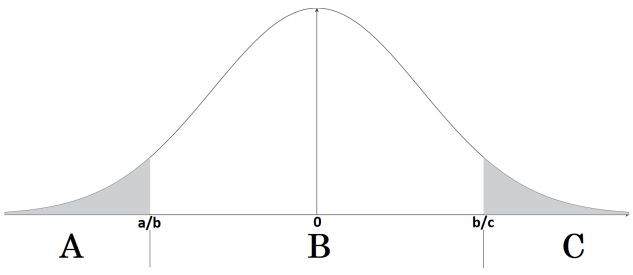


Figure 3.1: Overview of the test on a $N(0,1)$ distribution

[1]The notation with $1-\alpha/2$ is used here, whereas J. Jacod uses a notation with α. What follows is therefore strongly inspired by the work of J and AS, but uses different notations in order to bring out an overall coherence.

3.3 Choosing the tuning parameters

Now that the principle and construction of the Markovianity test have been presented, J and AS give some suggestions for choosing the tuning parameters. The purpose of this section is then to introduce these suggestions in order to implement the test in a Python program.

***

The kernel function f

J and AS first claim that the choice of the kernel function is rather immaterial in practice, while it follows the conditions given in page 14. It is chosen to work with the Quartic function, defined as:

$$f(x) := \frac{15}{16}(x^2 - 1)^2$$

with support $[-1, 1]$. It is an efficient kernel function, with coefficient $\beta = 5/7$ (square integrated f).

***

The bandwidth h_n

The rate of convergence of $\hat{\Phi}_n$ gets better when the bandwidth is increased, basically because more data is used to compute $\hat{U}(x,p)_n$. A good choice is to take:

$$h_n \asymp \frac{\Delta_n^{1/3}}{\log(1/\Delta_n)} \text{ in the non noisy case, and } h_n \asymp \Delta_n^{1/6} \text{ in the presence of MN.}$$

Note: $u_n \asymp v_n$ means that the sequences (u_n) and (v_n) are asymptotically equivalent, i.e. have the "same" asymptotic behaviour to within one multiplicative constant. It means that is is possible to change the values of h_n to get relevant results. This method is used in Chapter 4.

***

The pre-averaging window size k_n in the noisy case

The rate of convergence of $\hat{\Phi}_n$ gets better when the bandwidth is decreased, for the same reason as before. However, it is important to have k_n big enough in order to get rid of MN effects. J and AS suggest to take:

$$k_n \asymp \frac{\log(1/\Delta_n)}{\sqrt{\Delta_n}}$$

Note: It is needed that k_n is an even integer (see page 14). The implemented window size in the underlaying Python program is then the last even integer that is bigger than $\log(1/\Delta_n)/\sqrt{\Delta_n}$. It is easy to show that the asymptotic behaviour is the same and the convergence conditions of page 14 holds.

***

The coefficient r that regulates the weighted sums $\hat{\Phi}_n$ and $\hat{\Sigma}_n$

Looking at the definitions of $\hat{\Phi}_n$ and $\hat{\Sigma}_n$ can lead to the following: increasing r puts more weight on often visited sites. In fact, taking r large avoids small sample discrepancies that affect not often visited sites , but also typically decreases the power of the test because it reduces the number of sites $x \in \mathscr{X}$ which have a significant influence on the test statistic. Under the alternative, $\hat{S}_n^x$ roughly varies as $(L_1^x)^2$, so "equal weights" means taking $r = -2$, which is good for the power of the test, but bad under the null, since the true size of the test could be relatively far from the nominal size. A good compromise is perhaps to take $r = 0$, so that:

$$\hat{\Phi}_n = \sum_{x \in \mathscr{X}} \hat{S}_n^x \approx \int_{\mathbb{R}} \hat{S}_n^x dx \text{ (up to a normalizing constant)}$$

which is exactly the type of quantity which one want to test whether it vanishes or not.

The truncation threshold v_n

Using v_n is useful to get rid of hypothetical jumps in order to estimate $U(x,p)$. This parameter does not come in the rate of convergence, then the choice is open. In practice, J and AS suggest to use the same thresholds as for estimating the integrated volatility in the presence of jumps, that is:

$$v_n \approx \gamma\sqrt{\Delta_n} \text{ in the non noisy case, and } v_n \approx \gamma\sqrt{k_n\Delta_n} \text{ in the presence of MN}$$

where γ is "3 to 5 times a rough average of the volatility". As explained on page 10, estimating volatility in the presence of jumps requires some tricks about quadratic variation and jump detection. In fact, the integrated volatility, as an estimation, is obtained by considering the quadratic variation and reducing the latter because of the presence of jumps. As a reminder, what is called "integrated volatility" is in fact the integral of σ^2 according to Lebesgue measure, over the time horizon $[0,T]$.

The average of σ^2 is given by:

$$\langle\sigma^2\rangle := \frac{1}{T}\int_0^T \sigma^2(s)ds$$

Since the values of σ^2 are in fact quite little (see Figures 2.5, 2.6, 2.7 and 2.8) and the goal is to consider a "rough average of the volatility", it is not absurd to consider $\langle\sigma\rangle^{(rough)}$ as follows:

$$\langle\sigma\rangle^{(rough)} := \sqrt{\frac{1}{T}\int_0^T \sigma^2(s)ds}$$

The pros of such a definition is that it can be easily approximated with the V_m's as defined in (2.5). In fact, the V_m's represent estimations of the integrated volatility computed from the quadratic variation, since jumps are not in the model yet. Here, it is necessary to lower this value, since jumps must be considered, according to J and AS. Their idea is the to consider "3 to 5 times" something that is close to $\langle\sigma\rangle$, which is itself something a little smaller than $\sqrt{V_m}$, for some lag m. The idea is then the following: define γ as:

$$\gamma := K\sqrt{V_m}$$

for some constant K that is not bigger than 3. Take $K = 2$ for instance. The important thing is to define a threshold, then it is not absurd to exploit so much approximations. Moreover, differentiating an actual "little" jump from a variation due to a standard Brownian motion may be difficult. This is the reason why so many simplifying approximations are made. Finally, it is important to decide which lag m value to use. In fact, m is taken equal to 1 if no micro-structure noise is considered, ans is taken about 20 when MN belongs to the model. In fact, V_1 and V_{20} are of the same order of magnitude (see Figures 2.1, 2.2, 2.3 and 2.4) and since the choice of γ is quite open, one could choose both of them. To keep the code consistent, it is chosen to differentiate between the two cases (although this is a matter of detail).

***

The grid $\mathscr{X}$ that is used to construct the alternative ($\Omega_1(\mathscr{X})$)

When the cardinal of $\mathscr{X}$ is small, the procedure uses a limited number of points x, which lowers the power of the test. It is the important to take $|\mathscr{X}|$ large enough. In practice $\mathscr{X}$ is a regular grid $\{\alpha + i\beta : i = 0, 1, ..., N\}$ such that the random range R is in $[\alpha, \alpha + N\beta]$.

Knowing that the goal of $\mathscr{X}$ is to "simulate" the set of real numbers, one might be tempted to choose an extremely small β, like 10^{-100}. However, this method poses two major problems. First, the time required to calculate the test statistic increases linearly with $|\mathscr{X}|$. It can therefore be very slow to carry out the various statistical tests. Also, the construction of J and AS relies on an inequality relationship between the minimum distance between two points and the bandwidth h_n expressed as follows: $\beta > 2h_n$. It is then easy to construct $\mathscr{X}$: compute N such that the definition of the grid and the inequality are respected.

***

The choice of the tuning parameters is now done. These are the main elements that determine the success of the statistical test. The following chapter then focuses on the results of the test, in order to decide whether the volatility can be written as a function of the price or not.

CHAPTER 4

The results of the test

The present part contains the results of the various statistical tests explained in Chapter 3. Several values of the test statistic T_n are then presented, since Markovianity test has several parameters and is deployed in different situations. As indicated in Section 3.3, tuning parameters must be designed, as well as the choice of the model (with or without MN). The rest of this section is organized as follows: section 4.1 is devoted to the study of the test statistic when MN is not taken into account, then section 4.2 lists the test results when MN is considered, using several methods that are detailed in Appendix B. Finally, section 4.3 concludes this research paper by providing an answer to the following question: is the local volatility hypothesis validated ?

4.1 The simplest model: without micro-structure noise

As pointed out in the title of this section, not taking MN into account means performing a "simpler" calculation, at least easier to figure out. First, it is convenient to recall the principle of the test in such a situation. Section 3.1 begins by presenting the test framework and, above all, the hypotheses to be met by the data and processes studied. Note that these features do not bind the user, since they are necessary conditions for the construction of the test. Then no pre-processing concerning the data is needed. Such a work is introduced in Appendix A and is only relative to the quality of the observations of the prices. Markovianity test can thus be implemented by using directly (3.4).

Note that Section 3.3 gives some key point about the bandwidth h_n. The use of the asymptotic equivalence $\asymp$ gives some liberty about the choice of h_n. Looking at the construction of the grid $\mathscr{X}$, it is easy to see that the number of elements in the grid is proportional to $1/h_n$. This means that the more h_n is little, the more the grid is thin, and then the more the quality of the test is high (since $(\Omega_1(\mathscr{X}))$ is supposed to "simulate" (Ω_1)). By multiplying the suggestion $\Delta_n^{1/3}/\log(1/\Delta_n)$ by 0.1 to construct h_n, the asymptotic convergence still holds and there is a sufficient number of elements in the grid $\mathscr{X}$, and the results of the test are as follows :

Symbol	n	$\lvert\mathscr{X}\rvert$	T_n
AAPL	31955	711	18.059
AMZN	28794	628	39.420
DJIA	33701	816	12.638
BTC	167487	3321	38.684

Table 4.1: T_n values in the non-noisy case

According to the deciding result (3.9) and the reasoning on Figure 3.1, it is almost impossible to encounter such results in the situation where the null hypothesis (H_0') is valid. Indeed, if the null is valid, then $T_n \xrightarrow{\mathscr{L}} N(0,1)$. This means that the hypothesis of local volatility is clearly rejected, at least when MN is not considered in the underlaying model. The aim of the following section is thus to test (H_0') in the case of presence of MN. This is in fact the main purpose of this paper, since Chapter 2 already shows in page 7 the existence of some micro-structure noise in the studied HFFD.

4.2 Different methods with micro-structure noise

The Markovianity test first focuses on the absence of MN, before to consider MN in the framework. This last step is the most important, since it describes better the observations, in particular in Chapter 2. The second part of page 14 gives the idea of the computation of T_n. The main key is the so-called *pre-averaging* method, and it is now convenient to work with average data given by (3.5) as:

$$\bar{Y}_i^n := \frac{1}{k_n} \sum_{j=0}^{k_n-1} Y((i+j)\Delta_n)$$

where $Y = X + \eta$ is the observed process and η is the MN. The advantages in considering such a trick is detailed in the presentation of the test. Then, J and AS suggest to construct another datum $\hat{Y}_i^n$ given by (3.6) as:

$$\hat{Y}_i^n := \frac{1}{k_n^2} \sum_{j=0}^{2k_n-1} (Y((i+j)\Delta_n) - \bar{Y}_i^n)^2$$

in order to get explicit expressions for the $\hat{U}(x,p)_n$'s that are given in Appendix B. It is then possible to compute the test statistic. Table 4.2 gives the values of T_n with a multiplying factor equal to 0.001 to construct the bandwidth h_n (see more details in Section 4.1) :

Symbol	n	$\lvert\mathscr{X}\rvert$	T_n
AAPL	31955	1215	-3.270
AMZN	28794	1103	-3.008
DJIA	33701	1376	-3.598
BTC	167487	3718	-2.882

Table 4.2: T_n values in the noisy case with Jacod's formulas

Referring to the reasoning on Figure 3.1, the values of T_n are in zone A. It first means that it is highly unlikely that the null hypothesis is valid. However, the datum Φ is positive under the alternative ($\Omega_1(\mathscr{X})$). These values for T_n (large absolute value with negative sign) then occur very rarely. However, the four assets/index are in this situation. As noted in Appendix B, it then reveals an inconsistency in the process of the test. This is the reason why other methods must be tested.

It is first noticeable that the asymptotic equivalence is also used in the definition of the window k_n. In fact, the values of k_n as suggested in the tuning parameters Section 3.3 are very large (between 1000 and 4000). The same test can thus be made with $k_n \approx k_n/100$ in order to reduce the pre-averaging window, and then tend to a 20-window as in Chapter 2. Unfortunately, similar results a given (that is big absolute value and negative sign for T_n).

***

Such results naturally lead to a reflection about the used tools. As explained in Appendix B, there is a problem among the different steps of the test. This can be first due to a lack of clarity in the sources, notably J. Jacod's lecture (see [14]). Mathematical proofs and more details on the choice of the tuning parameters would be welcome, for example. Second, there can be a problem due to the data or the pre-processed data (see Appendix A). First, there the transition from no-MN hypothesis to MN hypothesis can bring "errors", meaning that the HFFD had already been pre-processed to get rid of MN. Since the data were collected from market prices, very likely with rounding effects or others (that can imply MN), this hypothesis should not, for the moment, be seen as solely responsible for the "bad" test results. Also, the problem could come from the pre-processing first step. Obviously, this process can add mistakes in the new data set. However, some intermediate calculations have been made on this new data set (see Appendix A again), and any problem occurred. For all of these reasons, it is then relevant to implement alternative methods in the computation of T_n. The rest of this section is then dedicated to the corresponding results.

The first alternative method is the most intuitive one. The idea is the following: consider the $\bar{Y}^n_{ik_n}$'s instead of the $X(i\Delta_n)$'s. The formula for $\hat{U}(x,p)_n$ is then given by (B.1). As before, it is convenient to choose the adequate parameters. In such a situation, the process of the test is inspired from the non noisy case, hence h_n is built with a multiplying coefficient equal to 0.05 (so that the grid $\mathscr{X}$ contains enough elements). Table 4.3 gives the values of T_n in such a situation :

Symbol	n	$\lvert\mathscr{X}\rvert$	T_n
AAPL	31955	1420	-2.740
AMZN	28794	1254	-1.074
DJIA	33701	1630	-1.725
BTC	167487	6640	-2.390

Table 4.3: T_n values in the noisy case (first alternative method)

The values of T_n are now a bit different as in Table 4.2. Indeed, these are certainly negative, but closer to 0, so that it would be absurd to reject the null for at least "AMZN". For this reason, and also the fact that Jacod's framework does not valid the null at all, it is relevant to consider other ways to compute T_n.

***

The second alternative is very close to the first one. The idea is to reduce the window k_n, as tried in the first part of this section, but now with the intuitive approach of the first method. The details are given in Appendix B. Table 4 then gives the values of T_n in such this framework :

Symbol	n	$\lvert\mathscr{X}\rvert$	T_n
AAPL	31955	1093	-17.763
AMZN	28794	965	-16.943
DJIA	33701	1254	-18.615
BTC	167487	5108	-48.840

Table 4.4: T_n values in the noisy case (second alternative method)

This time, the results are similar to those in Table 4.2, in that the values of T_n are large in absolute terms, but negative. It seems then necessary to find other ways to compute T_n, in order to catch positive values. That said, the big absolute values of the test statistic, in addition to Jacod's results, tend to reject the null hypothesis. It is then expected from the following to also reject (H_0'), or at least not to provide values which could correspond to a $N(0,1)$ distribution.

***

The previous methods, inspired from pre-averaging, have been useful to give a direction for the decision between the null (H_0') and the alternative ($\Omega_1(\mathscr{X})$). It is now relevant to search for another way to compute T_n that still respects the constraint of MN. The idea of pre-averaging is to consider several values of the observed price in order to pre-process them. It is possible to inspire from this method, i.e. to consider means of relevant data. Those "relevant data" are built so that the effects of MN vanish. However, for this reasoning to be valid, the underlying convergence property must be respected. Indeed, the "relevant data" is in fact something that converges to what has to be estimated. This idea is encountered in Chapter 2 in the construction of V_m. The following is thus inspired by this process.

As stated in Appendix B, one last method is given by imitating the construction of the estimators of the quadratic variation (V_m) in Chapter 2. It is given by the equation (B.5). The idea is then to consider data records m by m to construct the $\hat{U}(x,p)_n^{(k)}$ in order to average them. With $m = 20$, Table 4.5 gives the results of the test with this framework :

Symbol	n	$\lvert\mathscr{X}\rvert$	T_n
AAPL	31955	888	-9.720
AMZN	28794	784	-11.551
DJIA	33701	1019	-6.146
BTC	167487	4150	-31.694

Table 4.5: T_n values in the noisy case (third alternative method)

The values of T_n confirm the intuition developed in page 20. Large negative values for T_n do not allow to consider the null hypothesis as valid.

4.3 Conclusion of the paper

The content of Sections 4.1 and 4.2 is relevant to draw a conclusion for this paper. It is first and foremost important to recall the key elements of such a work.

In order to answer the question "Can volatility be written as a function of the price of the underlying asset ?", several steps have been completed. The development of Chapter 1 is useful in that a first model, the continuous semi-martingale (built from a Geometric Brownian Motion completed in (1.10) by the "stochasticity" of σ) is set up, making it possible to describe several log-price and volatility characteristics in Chapter 2 with additional elements such as MN and jumps. It is then possible to construct the Markovianity test protocol in Chapter 3. Due to the constraints of the observations and the nature of the HFFD, two hypotheses, (H_0') and ($\Omega_1(\mathscr{X})$), derived from (H_0) and (H_1), are constructed to represent the answers "yes" and "no" to the problematic of this paper. After a detailed mathematical construction, the principle of the test is established: it is based on the analysis of the values of a test statistic T_n, which converges to a normal distribution $N(0,1)$ under the null, and diverges under the alternative.

The results presented in Tables 4.1, 4.2, 4.3, 4.4 and 4.5 seem to confirm the following proposition: it is absurd to consider the null as valid. Despite the unforeseen circumstances encountered in the presence of MN, the "large negative" values of T_n allow to reject (H_0') quite naturally. This provides an answer to the problem of local volatility : **The volatility of the AAPL, AMZN, DJIA and BTC assets/indexes, described by their relative HFFD, cannot be expressed by a simple functional relationship with their price/value.** That said, it is important to note that this result was expected by J and AS in [14] and [15]. Moreover, local volatility hypothesis (H_0) is the simplest model for describing prices and the coefficient σ. This justifies the use of more sophisticated stochastic volatility models. However, local volatility models are useful when it comes to consider products in which the underlaying's volatility is predominantly a function of the level of this latter, such as interest rates derivatives. In that case, (H_0) is not rejected.

After studying the local volatility hypothesis throughout these parts, several comments are in order. Firstly, a data science project yields consistent results, provided that the input data is consistent. For this reason, a large part dedicated to the HFFD pre-processing was necessary. A second key to the success of such a project is the choice of models used. The example of jump processes in Chapter 2 is a good example, since considering the process J or not can radically change the interpretation given to the coefficient σ. The presence of J in the framework of the test then makes the model more complete. Finally, it is worth remembering that no model is perfect. In this way, each model can be fine-tuned to match the reality of the observations and the desired features, as long as it remains simple enough to handle. This is the very reason for the existence of this research paper about local volatility, as well as the emergence of many different stochastic models.

Appendices

APPENDIX A

High-Frequency Financial Data (HFFD)

The research project on the study of volatility is based on the use of HFFD. These data are used to test whether the volatility of an asset (or an index value) can be written as a function of its price (or its value). This is, under some approximations, what we call Markovianity. It is therefore crucial to understand them, to analyze their structure and coherence, before exploiting them. This is the first technical achievement of the project, since it involves all the reliability of the research paper.

A.1 Initial HFFD

The project focuses on the study of three assets and one index that are Apple stocks, Amazon stocks, Bitcoin and Dow Jones Industrial Average index, which symbols are respectively AAPL, AMZN, BTC and DJIA. These are all financial instruments, but they differ in nature and operation. The Dow Jones is the world's oldest stock market index. It is based on the market capitalization of the 30 largest companies listed on the New York Stock Exchange. The companies included evolve over time. It is made up of a weighted average of the share prices of these companies. The main purpose of the Dow Jones is to give a general idea of stock market performance. It cannot be bought or sold directly, but investors can trade derivative contracts based on the index. Shares, such as those in Amazon (AMZN) or Apple (AAPL), represent partial ownership of a company. When one buy shares in a company, one become a shareholder, entitled to a share of the company's profits and voting rights at general meetings. The value of shares fluctuates according to various factors, such as the company's financial performance, market conditions and investor expectations. Shares can be bought and sold on stock exchanges. Bitcoin (BTC) is a crypto-currency, a form of digital currency based on a technology called blockchain. Unlike stocks, Bitcoin is not tied to corporate ownership. It was created as a decentralized digital currency and is not regulated by a central authority such as a bank or government. Bitcoins can be bought, sold and used for peer-to-peer transactions. The value of Bitcoin is determined by supply and demand on crypto-currency exchange platforms, and can be highly volatile. Those differences in the nature of the studied financial instruments are likely to lead to different results in terms of volatility behavior, or simply in the form of the data initially collected.

These various financial products have some common points, particularly in the measurement of their dynamics using high-frequency data, often grouped under the same attributes, as shown in Table A.1.

Attribute	**Value**
<TICKER>	Asset (or index) ticker (ex. BTSX.BTC/USD for Bitcoin)
<PER>	Number of minutes in a trading period (always 1)
<DATE>	Date of the trading period (ex. 20200105 for 5 Jan 2020)
<TIME>	Beginning time of the trading period (ex. 093100 for 09:31:00)
<OPEN>	Open price (or index value) of the trading period (in USD)
<HIGH>	Same with higher price
<LOW>	Same with lower price
<CLOSE>	Same with close price
<VOL>	Exchange volume during the trading period

Table A.1: Attribute description for a HFFD table.

Collected on the website [website], these HFFD take the form of a .csv file (Excel table), each containing several thousand data records. Nine columns describe each Excel table (AAPL, AMZN, DJ and BTC) and represent the attributes <TICKER>, <PER>, <DATE>, <TIME>, <OPEN>, <HIGH>, <LOW>, <CLOSE> and <VOL>. An example data record, taken from AAPL table, is shown below.

US1.AAPL,1,20200102,093100,296.2500000,296.3000000,295.4000000,295.8000000,25345

A few remarks can be made about these initial HFFD. Firstly, we have decided to conduct our study over a common time period for each financial product, namely from the beginning of January 2020 to the beginning of May 2020. This period is very interesting, as the global economy has experienced some very strong episodes at this time, such as the direct consequences of the Covid-19 crisis. Before being high-frequency financial data, these data represent records of the current economic landscape. It is therefore important to choose to study this landscape through a characteristic period of time.

It is also worth explaining why these data are HFFD. With a computer, it is easy to show that all the data records have PER=1, which means that the data recording periods are relatively short (one minute), compared with the extent of the overall measurements (around 4 full months). This also explains why the underlying Excel tables contain a large number of data records (around 25,000 for AMZN, and almost 170,000 for BTC).

A.2 Modification of initial data

As in any data science project, it is relevant to ensure that the collected data is usable. Indeed, the success of such a project lies in the right choice of algorithmic and mathematical methods to be used, but also and above all in the quality of the data to be exploited. In the context of HFFD, it is crucial to ensure that they correspond to what is expected of data of this nature.

DISCLAIMER: The purpose of this appendix, as with the report in general, is to present the technical solution, implemented on computer, for carrying out the Markovianity test of J. Jacod and Y. Aït-Sahalia's. It is therefore relevant to detail this solution in the report. However, in order to lighten the paper, the underlaying Python code is available on [Github address or other, ask tutors for details].

The first step in data exploitation and pre-processing is to import the data into a relevant work environment. Python, and its numerous associated libraries, is a pertinent tool to complete the project. The aim is to import the data contained in the Excel tables AAPL, AMZN, DJIA and BTC into a Python environment (here Spyder). As Python is an object-oriented language, it is easy to create classes, so financial data can be easily represented and exploited. That is the aim of the first file of the project, called init_data, which role is to collect and pre-process the initial financial data. The main idea behind data collection is the use of the *pandas* library, which is exploited in the creation of the InitData class that initially allows, with the help of an initialization function, the creation of the dataframe related to the corresponding Excel table. Objects aaplData, amznData, djiaData and btcData are then created. In order to search for specific data using the index of a given record, the function getColumn(self,columnName) is also often used.

After a quick look at the values given for the prices (or values) of the various assets (or indexes), we can notice some outliers contained in the aaplData and amznData objects: the prices (OPEN, HIGH, LOW, and CLOSE) are much larger than they should be, by comparison with the corresponding values in other data bases. This can be explained by the following phenomenon: stock splits by Apple and Amazon. A stock split happens when a company increases the number of its shares to boost the stock's liquidity. Indeed, a stock split on a 4-for-1 basis happened on August 28, 2020 with Apple, and a 20-for-1 other stock split happened on June 6, 2022 with Amazon. This means that for Apple (resp. Amazon), there are 4 (resp. 20) times as many shares as before August 28, 2020 (resp. June 6, 2022). As the selected financial data were recorded before these dates, the values in the OPEN, HIGH, LOW and CLOSE columns of aaplData (resp. amznData) must be divided by 4 (resp. 20). This is what the function multColumnFactor(self,columnName,factor) does.

It is now relevant to focus on the distribution of data records and what does it mean. By printing len(aaplData.getColumn("OPEN")), we obtain that aaplData contains 31,837 data records (also 24,360 for Amazon, 33,701 for Dow Jones and 165,299 for Bitcoin). AAPL, AMZN and DJIA are subject to financial markets, that are opened during trading days, and closed during weekends. Then the size of BTC table is obviously much larger than other ones.

On closer inspection of the data, it is possible to see that some moments are not represented in the set of records. For example, one can go from a minute t_i to another minute t_{i+2}, without considering t_{i+1}. This can be a big problem when it happens within the same trading day, as we want to work on high-frequency data. In addition, it is also interesting to work on regular measurement intervals to implement the Markovianity test (see in Chapter 3). The function fillMissingMinutes(self,symbol) is then implemented in the definition of the class InitData. Its role is to fill in the "missing" financial data so that the data records for each trading day are uniformly distributed. This is very useful when it comes to displaying price trends, or performing volatility tests. In order not to induce any change in price changes, the fillMissingMinutes function assigns to the missing minutes the price (or value) of the underlying financial instrument at the last moment preceding them and which is entered in the initial data. It also assigns the value VOL=0 for all new data record, so that the exchanges are not biased.

After a serie of tests, allowing to design and test the effectiveness of the function fillMissingMinutes, Excel files are created in a new directory and will serve as usable data. These new files are then converted again according to the definition of a new class called CleanData, in the file called clean_data, using the *pandas* library again. The new aaplData, amznData, djiaData and btcData objects then contain uniformly distributed data records for each trading day. However, it is also important to check whether the clean data gives an almost constant time line, i.e. a time line, given by the columns DATE and TIME, that reflects well the real time. With such a property, we can affirm that the trading days are uniformly distributed in the real time line. Moreover, we have built the clean data so that the records are also uniformly distributed in each trading days. That means that the plots given by the clean data time line are analog to plots that would have been produced with a continuous time line. An efficient way to check this uniformity property is to represent the real time compared to the clean data time line by, for example, plotting a Y-X curve, with Y axis being the real time in days spent since reference time, and X axis being the number of trading days spent since the same reference time trading days. This is the role of the function plotDatesFromCutValues. The following figures show such results.

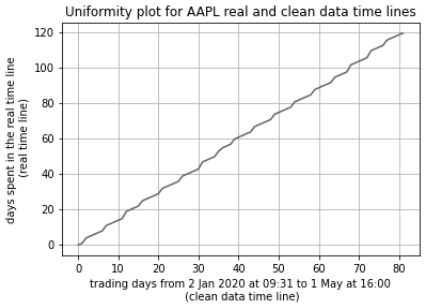


Figure A.1: Uniformity curve for AAPL

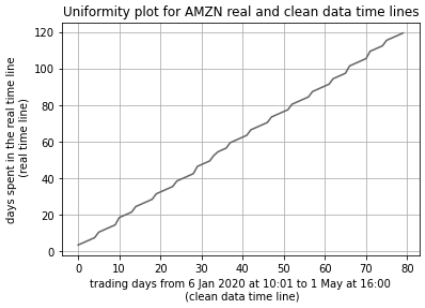


Figure A.2: Uniformity curve for AMZN

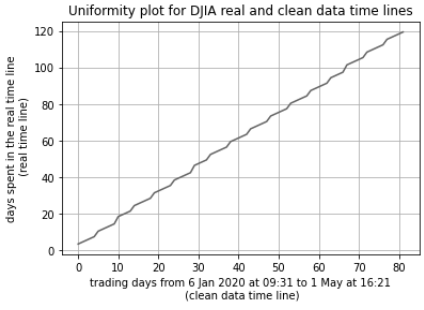


Figure A.3: Uniformity curve for DJIA

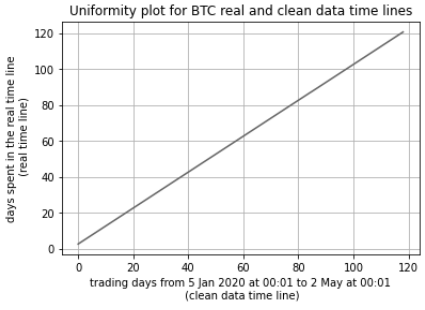


Figure A.4: Uniformity curve for BTC

The four plots, for each asset/index, show an almost linear relationship between the real spent time, and the spent trading days. It shows that with a global point of view, the trading days are uniformly distributed. This result was obviously expected, since there is only a few number of detected lacking trading days in the serie of tests.

The plots for BTC and the other assets/index are different. It is because AAPL, AMZN and DJIA are subject to financial markets, then there are not trading days during weekends (that explains the little regular jumps of the curve), whereas BTC is not submitted to financial market, hence there is a perfect linear relationship (and even equality) between real days line and clean data time line.

Finally, the linear aspect of all of the curves allows us to consider real time line and clean data time line as equivalent, that is convenient to produce plots that stick as more as possible to the "reality".

After performing all these analyses and modifications, the data contained in the aaplData, amznData, djiaData and btcData objects are now usable. They group together all the properties required to be quality HFFD.

A.3 Application : plots for share prices, asset values and market performance

Having quality HFFD means being able to use it in many applications. Before starting the work on the Markovianity test, it is already possible to display the price history (or values) for a given asset (or index), between given dates, for a given price type (open, high, low or close). This is the aim of the plot_prices file. The following figures represent some AMZN charts for different time scales.

Figure A.5: AMZN charts for a very large time scale

Figure A.6: AMZN charts for a large time scale

Figure A.7: AMZN charts for a little time scale

Figure A.8: AMZN charts for a very little time scale

This is a fairly common use of HFFD. The rest of the report focuses on the in-depth study of these new data, contained in the objects of the class CleanData, in order to derive volatility information. The study of volatility being strongly related to the study of the variation of the prices, it is capital to be aware of the pros and the cons of each model which aim is to represent the price (or value) of the asset (or index) as a stochastic process. For example, one can notice a big jump for AMZN about the end of January 2020. Shall we introduce a jump process to model such strong variations ? Or shall we only consider a volatility term linked to a standard Brownian Motion ? This is one of the main questions of this research paper.

APPENDIX B

Alternative methods in the case of micro-structure noise

This appendix is a relevant step in the study of Markovianity, since multiple interpretations can be given to J and AS's suggestion (which is the use of pre-averaging methods). The purpose of the first section is the presentation of Jacod's tools in his last conference (see [14]), while the second one presents other ways to compute the test statistic.

B.1 J. Jacod's suggestion

As stated in Section 4.2, J. Jacod constructs two new tools in order to build the $\hat{U}(x,p)$'s that are:

$$\bar{Y}_i^n := \frac{1}{k_n}\sum_{j=0}^{k_n-1} Y((i+j)\Delta_n) \text{ and } \hat{Y}_i^n := \frac{1}{k_n^2}\sum_{j=0}^{2k_n-1} (Y((i+j)\Delta_n) - \bar{Y}_i^n)^2$$

The first one represents the average of the observed process over a given window, while the second one is relative to its variance. Then applying the idea of the pre-average method to the definition of the estimator of $U(x,p)$ in the absence of MN (see (3.4)), J. Jacod gives explicit expressions for these quantities for all $p \in \{0,2,4\}$:

$$\hat{U}(x,0)_n = k_n\Delta_n \sum_{i=1}^{n} f_n(\bar{Y}^n_{2ik_n} - x)1_{\{|\bar{\Delta}_i^n| \le v_n\}}$$

$$\hat{U}(x,2)_n = 3\sum_{i=1}^{n} f_n(\bar{Y}^n_{2ik_n} - x)(\bar{\Delta}_i^n - \hat{Y}^n_{2ik_n})^2 1_{\{|\bar{\Delta}_i^n| \le v_n\}}$$

$$\hat{U}(x,4)_n = \frac{3}{k_n\Delta_n}\sum_{i=1}^{n} f_n(\bar{Y}^n_{2ik_n} - x)(|\bar{\Delta}_i^n|^4 - 6|\bar{\Delta}_i^n|^2\hat{Y}^n_{2ik_n} + 3|\hat{Y}^n_{2ik_n}|^2)1_{\{|\bar{\Delta}_i^n| \le v_n\}}$$

where $\bar{\Delta}_i^n := \bar{Y}^n_{2ik_n} - \bar{Y}^n_{(2i+1)k_n}$. These equations are, by definition, very much inspired by the definition of the estimator in the absence of MN (see (3.4)). However, it is regrettable not to have any mathematical proof from J. Jacod's work to explain this precise shape. For instance, there is no correspondence between the coefficient $1/m_p$ and the occurrences of the number 3. It also seems complicated to bring the $\hat{Y}^n_{2ik_n}$'s and the $\bar{Y}^n_{2ik_n}$'s into (3.4), whereas a simple use of $\bar{Y}^n_{ik_n}$ in the equation seems to satisfy the requirements of the pre-average method. This being said, this paper gives in Chapter 4 the results of the test using such an implementation.

Test results using such a framework seem disappointing. As explained in Section 4.2, the negativity of the values T_n reveals an inconsistency in the test process. For this reason, as well as the one mentioned above about the shape of the estimators, other methods for calculating the test statistic need to be put in place, still inspired by the pre-averaging method. Section B.1 then focuses on alternative methods to compute T_n. As previously, the results for each of them are detailed in Section 4.2.

B.2 Some alternative ways to compute the test statistic

In their first presentation ([15]), J and AS suggest to implement the pre-averaged log-price Y in the definition of $\hat{U}(x,p)_n$ to construct the estimators of U in the case of MN. A first intuitive method is then to replace the $X(i\Delta_n)$'s in (3.4) by the $\bar{Y}^n_{ik_n}$. Indeed, when i describes the discrete interval $[\![0, \lfloor \frac{n}{k_n} \rfloor]\!]$, all data records are processed once and only once (except the last ones between indexes $\lfloor \frac{n}{k_n} \rfloor k_n$ and n, which are not numerous, and are not decisive in view of the large sample of Y values). The underlaying formula for the $\hat{U}(x,p)_n$'s is then given by the following formula:

$$\hat{U}(x,p)_n := \Delta_n^{1-\frac{p}{2}} \frac{1}{m_p} \sum_{i=1}^{\lfloor \frac{n}{k_n} \rfloor} f_n(\bar{Y}^n_{(i-1)k_n} - x)|\bar{Y}^n_{ik_n} - \bar{Y}^n_{(i-1)k_n}|^p 1_{\{|\bar{Y}^n_{ik_n} - \bar{Y}^n_{(i-1)k_n}| \leq v_n\}} \tag{B.1}$$

Note: It is important to notice that even in the noisy case, the test statistic is computed as:

$$T_n = \sqrt{h_n/\Delta_n} \frac{\hat{\Phi}_n}{\sqrt{\hat{\Sigma}_n}} \tag{B.2}$$

This is due to the following: the pre-averaging step (construction of $\bar{Y}$) is here considered as sufficient to "get rid" of the MN. Indeed, the pre-averaging process vanishes the local irregularities. This first alternative method, as the remaining ones, then considers their datum $\hat{U}(x,p)_n$ as the equivalent of $\hat{U}(x,p)_n^{(nonnoisy)}$. The reason of (B.2) is that results can be biased due to the asymptotic definition of k_n. Multiplying k_n by a constant will obviously keep the asymptotic equivalence property (see definition of $\asymp$ in Section 3.3), but it will also divide T_n by the square root of such a constant, thus the value of the test statistic becomes irrelevant. This reasoning leads to the fact that the tuning parameters h_n and k_n can be defined with a multiplying factor, provided that the test statistic is defined as in (B.2).

The results of the test for this alternative, presented in Chapter 4, show that there is a difference between the results of Jacod's method and that one. It does not show that one of them is absolutely false, since the main result of the mathematical construction of the Markovianity test holds on a convergence property, but that the decision of the null or the random alternative is now relative to one method (see values of T_n for "AMZN"). This is the reason why other alternative methods must be considered, in order to provide several point of views for the conclusion of this paper (see Section 4.3).

***

The second alternative is inspired from the first one. The idea is to reduce the window k_n, as tried in the first part of this section, but with the intuitive approach of the first method. The idea is the following: consider a new pre-averaged price, but now relative to the lag $m = 20$:

$$\bar{Y}(i,m) := \frac{1}{m} \sum_{j=0}^{m-1} Y((i+j)\Delta_n) \tag{B.3}$$

The motivation of such a definition is to tend to the framework of Chapter 2, which gives a relevant value of the lag in order to "get rid" of MN effects. The method then consists in plugging (B.3) in the definition of $\hat{U}(x,p)_n^{(nonnoisy)}$, as above. The results of this method, that are given in Chapter 4, have also negative signs. However, their large absolute values tend to reject the null hypothesis. In order to bring more argument in this direction, it is relevant to search for another way to compute T_n.

***

The previous pre-averaging methods have been useful in that they provide information on the behavior of stochastic volatility. Unfortunately, they provide "big negative" values, which does not fit to the mathematical framework of Chapter 3. However, as a way to confirm the rejection of the null (or rather its non validation), using the results of a last method can be relevant. This is the purpose of what follows.

The idea of the last method is close to the pre-averaging method. In Chapter 2, the quadratic variation of the log-price X is estimated with a tool called V_m, for a given lag m, defined as :

$$V_m := \frac{1}{m} \sum_{k=0}^{m-1} \sum_{i=0}^{\lfloor \frac{n}{m} \rfloor - m} (X(t_{(i+1)m+k}) - X(t_{im+k}))^2 \tag{B.4}$$

In order to get rid of the effects of MN, which occur when the time between two data records is too short, the estimator of U can be defined as in (B.4), that is :

$$\hat{U}(x,p)_n := \frac{1}{m} \sum_{k=0}^{m-1} \hat{U}(x,p)_n^{(k)} \tag{B.5}$$

with $\hat{U}(x,p)_n^{(k)}$ defined as :

$$\hat{U}(x,p)_n^{(k)} := \Delta_n^{1-\frac{p}{2}} \frac{1}{m_p} \sum_{i=1}^{\lfloor \frac{n}{m} \rfloor - m} f_n(X(((i-1)m+k)\Delta_n) - x)|\Delta_i^{n,m}|^p 1_{\{|\Delta_i^{n,m}| \leq v_n\}}$$

where $\Delta_i^{n,m} := X(((i-1)m+k)\Delta_n) - X((im+k)\Delta_n)$. This is a way to consider the whole set of data records, without suffering from the effects of MN. This result holds since every $\hat{U}(x,p)_n^{(k)}$ converges in probability to $U(x,p)$, for all x and p. Hence, the definition of $\hat{U}(x,p)_n$ is relevant.

As previously, the results in Chapter 4 show a non validation of the null hypothesis. Using the reasoning in Section 4.3, it is then possible to reject the null hypothesis.

Bibliography

[1] Aït-Sahalia, Y. and Yu, J. 'High frequency market microstructure noise estimates and liquidity measures'. In: *The Annals of Applied Statistics* vol. 3, no. 1 (2009), pp. 422–457.

[2] Amazon. *Amazon financial report for Fourth Quarter 2019*. (2020).

[3] Bachelier, L. 'Théorie de la spéculation'. In: *Annales Scientifiques de l'École Normale Superieure* vol. 17 (1900), pp. 21–86.

[4] Black, F. and Scholes, M. 'The Pricing of Options and Corporate Liabilities'. In: *Journal of Political Economy* vol. 81, no. 3 (1973), pp. 637–654.

[5] Clark, P. K. 'A subordinated stochastic process model with finite variance for speculative prices'. In: *Econometrica: journal of the Econometric Society* vol. 41, no. 1 (1973), p. 135.

[6] Comte, F. and Renault, E. 'Long memory in continuous-time stochastic volatility models'. In: *Mathematical Finance. An International Journal of Mathematics, Statistcs and Financial Economics* vol. 8, no. 4 (1998), pp. 291–323.

[7] Cont, R. 'Long range dependence in financial markets'. In: *Fractals in Engineering* (2005), pp. 159–179.

[8] Cont, R. and Tankov, P. *Financial Modelling With Jump Processes*. Chapman Hall/CRC financial mathematics series, (2003).

[9] Cutler, D. and Summers, L. 'What moves stock prices?' In: *Journal of Portfolio Management* (1989), pp. 4–12.

[10] Fama, E. F. 'The Behavior of Stock-Market Prices'. In: *The Journal of Business* vol. 38, no. 1 (1965), pp. 34–105.

[11] Fan, J. and Wang, Y. 'Multi-scale Jump and Volatility Analysis for High-Frequency Financial Data'. In: *Journal of the American Statistical Association* vol. 102, no. 480 (2007), pp. 1349–1362.

[12] Gatheral, J. *The volatility surface: A practitioner's guide*. John Wiley Sons, (2006).

[13] Gatheral, J. and Rosenbaum, M. '"Volatility is rough'. In: *SSRN Electronic Journal* (2014).

[14] Jacod, J. and Aït-Sahalia, Y. 'Testing the Markov property in a high-frequency setting'. In: *International Conference "Theory of Probability and Its Applications: P.L. Chebyshev"* (2021).

[15] Jacod, J. and Aït-Sahalia, Y. 'Testing whether volatility can be written as a function of the asset price'. In: *University of Singapore courses* (2012).

[16] Latané, H. A. and J., R. 'Standard deviations of stock price ratios implied in option prices'. In: *The journal of finance* vol. 31, no. 2 (1986), pp. 369–381.

[17] Mandelbrot, B. 'The Variation of Certain Speculative Prices'. In: *The Journal of Business* vol. 36, no. 4 (1963), pp. 394–419.

[18] Merton, R. C. 'Theory of rational option pricing'. In: *The Bell journal of economics and management science* vol. 4, no. 1 (1973), p. 141.

[19] Samuelson, P. A. and Merton, R. C. 'A Complete Model of Warrant Pricing that Maximizes Utility'. In: *Industrial Management Review* vol. 10 (1972), pp. 17–46.