Analysis of Financial Data through Signal Processing Techniques

Konstantinos Drakakis¹

UCD CASL²
University College Dublin

Abstract

We show, using the Dow-Jones volume as a working example, that techniques borrowed from Signal Processing and Internet traffic modeling (such as the envelope, the Instantaneous Frequency, the Averaging function etc.) can help with the analysis of financial data. As an application, we successfully confirm the 4 year cycle of the Dow-Jones index.

Mathematics Subject Classification: 60G35, 91B70, 60G10, 42C40, 60H15

Keywords: Dow-Jones, finance, Internet traffic modeling, Signal Processing, Instantaneous Frequency, envelope, drift, multiscale, stochastic processes

1 Introduction

The recent mathematical formalism of Financial Mathematics, as represented by the Black-Scholes model (BSM) for the price evolution of financial instruments, is actually an abstract mathematical framework for the study of differential equations (DEs) with underlying stochastic processes as components, and could therefore be used, in principle, in other scientific fields where the relevant models call for similar stochastic DEs (SDEs). One such field in the broader area of Communications is Internet traffic modeling, where a lot of effort has been put recently into modeling Internet traffic on a particular network link as a stochastic process and into determining the properties of this process [8, 13, 12, 14]. Although from a mathematical point of view there is

¹Address: UCD CASL, University College Dublin, Belfield, Dublin 4, Ireland. Email: Konstantinos.Drakakis@ucd.ie. The author is also affiliated with the School of Mathematics, University College Dublin

²Complex & Adaptive Systems Laboratory (http://casl.ucd.ie)
hardly any difference, the terminology and the notation in the two fields are quite different, thus clouding at first sight the otherwise strong connection.

As then each field developed its own tools and methods for its own needs, it is conceivable that one could benefit from the accumulated knowledge in the other. In this work we intend to investigate just that: whether finance can clarify Internet traffic models, and especially whether tools developed for the study of Internet traffic can help with the analysis of financial data.

2 The tools

2.1 Finance

Let $S(t), t \in \mathbb{R}$ be a stochastic process in time that conforms with the BSM, or with one of its extensions [7, 11]; then there exist constants $r \in \mathbb{R}$ and $\sigma \in \mathbb{R}^+$ and a stochastic process $W(t), t \in \mathbb{R}$, such that $S(t)$ is the solution of the SDE:

$$\frac{dS(t)}{S(t)} = rdt + \sigma dW(t)$$

In the original model $W$ is a Brownian motion; the exact interpretation of the derivative $dW$ and the rules governing its behavior and handling are the subject of \(\hat{I}to's\) Calculus [7]; in the model extensions $r$ and $\sigma$ can be functions of $t$ or even stochastic processes themselves (note that $\sigma$ is, by convention, always taken non-negative), and $W$ may be a Fractional Brownian Motion or even have a stable component to simulate jumps. As $\frac{dS(t)}{S(t)} = d\ln(|S(t)|)$, the quantity $\ln(|S(t)|)$ appears quite important in itself: when $S(t)$ denotes as usual the return of an asset, it is known as the log-return.

2.2 Internet traffic

As Internet traffic appears, at least under some conditions, to have different properties at different time scales (namely, it is a multifractal [4, 3, 9]), the Internet traffic research community invented or adopted tools that analyze this property; the two most widely used among them are perhaps the Multifractal Spectrum [9] and Energy/Averaging function[4, 13, 2]. In this work we will focus on the latter.

Suppose we have a discretized version $s(n) = S(nT), n \in \mathbb{Z}$ of our original continuous and stationary process $S(t)$, which can be mapped to a piecewise constant approximation $S_T(t)$ of $S(t)$; in its original form [4, 13] the Energy function $E(j), j \in \mathbb{N}$ was the mean energy ($L^2$-norm) at scale $j$ of the Haar
wavelet transform of $S_T(t)$, which, because of randomness and stationarity, is equivalent to the mean square value of the Haar wavelet coefficients at that scale. As in all practical cases $S(nT)$ is non-zero only within a finite range of values for $n$ (say for $n = 1, \ldots, N$ without loss of generality, where typically $N = 2^m$ since the Haar wavelet has support of length 2), this definition suffered from an inherent deficiency: the coarser time scales, where the wavelets have large supports, yielded few coefficients, and therefore the relevant mean square value could not be estimated accurately; this was obvious by a “ripple” in the graphs at coarse scales.

The solution to this was to consider $s(n), n = 1, \ldots, N$ arranged on a ring, effectively periodizing it [2]. This allows us to obtain the same number of wavelet coefficients at all scales, just by wrapping the wavelets around the ring, thus estimating the mean square value at all scales with the same accuracy. For this to make sense, however, it is crucial that stationarity hold, as the statistical behavior of $s(n)$ should be the same small and large $n$.

The final improvement to the function came by normalizing the wavelets in a different way, thus forming the Averaging function $A(j), j \in \mathbb{N}$ so that the Central Limit Theorem (CLT) convergence exponent of $s$ can be read out of $A$'s graph [2]. To be explicit, we define:

$$A(j + 1) = E\left(\left\|\frac{1}{2^j} \sum_{i=1}^{2^j} s(i) - \frac{1}{2^j} \sum_{i=1}^{2^j} s(i + 2^j)\right\|^2\right), j \in \mathbb{N}$$

Then it can be shown that if:

$$\psi(n) \sqrt{E\left(\left\|\frac{1}{n} \sum_{i=1}^{n} s(i) - E(s(1))\right\|^2\right)} \xrightarrow{n \to \infty} 1$$

with $\psi(n) = Cn^a$ for some $C > 0$ and $a$, it follows that $\log_2(A(j + 1)) = K - aj$, so that $\Delta \log_2(A(j + 1)) = -a$. Multifractal processes have the property that $a$ is actually a function of $j$, so that the “difficulty of convergence” of the mean estimator $\frac{1}{n} \sum_{i=1}^{n} s(i)$ to the true mean varies with scale. Note that $a < 0.5$ corresponds to subcritical convergence where sample correlation is destructive and impedes convergence to the mean, while $a > 0.5$ correspond to supercritical convergence where sample correlation is constructive and accelerates convergence to the mean (compared to the usual CLT). It can be shown [1] that when a time scale is comparable to a time interval in which the statistical properties of the process change, then the Averaging function exhibits a change in slope there and usually has a local maximum.
2.3 Some statistics

2.3.1 Stationarity

There is in general no reason to assume that a BSM conforming process $dS$ is stationary; in fact we know that in the original BSM $S$ is of an exponential form with an exponential time drift, hence $dS$ in this case is certainly not stationary. $dS(t)/S(t) - rdt = \sigma dW$ can be stationary, however, and it is indeed in the original BSM, where $dW$ is white noise. Consequently, when applying the Averaging function, the correct process is $dS(t)/S(t) - rdt$, not $dS(t)$.

2.3.2 Drift and envelope

A practical difficulty arises when $r$ is time-varying: how can we remove this deterministic time-varying component (drift) and isolate the random process component? Broadly speaking, the drift is the local mean value each time, while the process is the oscillation around it. We can borrow then here the notion of the envelope of a signal from Telecommunications, closely tied to the idea of Instantaneous Frequency [6].

A simple way to form the envelope of a signal $x(t)$, $t \in \mathbb{R}$ is this: let the sequences $H = \{h_i\}_{i \in \mathbb{Z}}$ and $L = \{l_i\}_{i \in \mathbb{Z}}$ show the locations of the local maxima and minima of $x$, respectively, and let $\text{Lin}_{A,B}(t)$ be the continuous linearly interpolated function assuming the values $B = \{b_i\}_{i \in \mathbb{Z}}$ on $A = \{a_i\}_{i \in \mathbb{Z}}$ and changing slope on $A$ only. Let now $x^\uparrow$ be the pointwise maximum between $x$ and $\text{Lin}_{H,x(H)}(t)$ and $x^\downarrow$ the pointwise minimum between $x$ and $\text{Lin}_{L,x(L)}(t)$. The pair $x^\downarrow$, $x^\uparrow$ forms the envelope of $x$ (literally as $x^\downarrow(t) \leq x(t) \leq x^\uparrow(t)$ pointwise) and the drift is $\bar{x}(t) = \frac{x^\uparrow(t) + x^\downarrow(t)}{2}$; $x_1 = x - \bar{x}$ is the oscillating component, and we can keep applying this method to get $x_1 = \bar{x}_1 + x_2$ etc. We can even apply higher order splines to get smoother envelopes instead of piecewise linear [6].

2.3.3 Kolmogorov distance

Assume we have 2 probability density functions (PDFs) $f$ and $g$, and we would like to measure how similar they are to each other. A fast and reliable way to determine this is to measure their Kolmogorov distance [2] (KD). First, determine the cumulative density functions $F(x) = \int_{-\infty}^{x} f(t)dt$, $G(x) = \int_{-\infty}^{x} g(t)dt$; the Kolmogorov distance is $d_K(F,G) = \max_{x \in \mathbb{R}} |F(x) - G(x)|$, and is closely related to the Gini index [5] frequently used in Economics.
2.4 Data and acquisition

There is a big difference in the ease of data acquisition in the two fields. Financial data on stock prices, stock market indices, trade volume etc. tends to be a matter of public record and is readily available on the Internet today; all of the data we will subsequently use was obtained from the archive in Yahoo! Finance. In contrast, Internet traffic data is extremely hard to obtain, because of the potential privacy infringement. One therefore needs either to collect one’s own data, if one has the chance, or depend on the goodwill of others who make “sanitized” data publicly available.

We will consider below some measurements (specifically the volume) on the Dow-Jones index obtained from Yahoo! Finance; the values span the time from 1928 up to today.

3 Data analysis

3.1 Dow-Jones volume – Log derivative

Let $S(t)$ be the trade volume of the Dow-Jones index, as downloaded from the Yahoo! finance webpage; time is inverted, in the sense that the most recent values appear for small $t$. The plot of $dS(t)/S(t)$ shows that there is no drift ($r = 0$), so no pre-processing with the envelope is necessary; moreover, the graph (Figure 1a) appears to be a superposition of 2 processes: one that looks very much like a Gaussian (but further analysis is obviously required to determine whether it actually is) (Figure 1b), and another one that is bursty and responsible for the appearance of “spikes” (Figure 1c). This situation is reminiscent of the $\alpha$-$\beta$ Internet traffic model [12], where Internet traffic is represented as the superposition of 2 totally dissimilar processes: the bursty $\alpha$-process, and the Gaussian $\beta$-process. We can write then $dS(t)/S(t) = dW_\alpha(t) + dW_\beta(t)$.

Clearly $dW_\alpha(t)$ can be modeled as a train of spikes: $dW_\alpha(t) = \sum_{i \in \mathbb{Z}} c_i \delta(t - t_i)dt$; the distribution and the correlation of the distances $t_{i+1} - t_i$ is obviously very important, and in the original $\alpha$-$\beta$ traffic model they were usually taken to be i.i.d. exponential, although it was known that such an approximation was rather poor [10]. In our case also the autocorrelation does not corroborate the i.i.d. assumption (Figure 2b), and further study is necessary.

The mean and standard deviation of $dW_\beta(t)$ are found to be $\mu = 0.0226$ (almost 0) and $\sigma = 0.2315$. The KD between the marginal distribution of this process and a perfect Gaussian with the same mean and variance (see Figure 2d) is approximately 0.08, proving that the approximation of the marginal by Gaussian is reasonable, at least as a first approach. Further, the autocorrelation of the process shows poor long range dependence (Figure 2a), hence we
can reasonably assume that the samples are independent. Overall, \( dW_\beta(t) \) can be approximated by a Gaussian noise.

The only significant deviation in the approximation above comes from multifractality: the Averaging function (Figure 2c) shows that the process is a weak multifractal, as the slope is supercritical on the one hand, and a significant change of behavior around scale 12 (from 10 to 14) occurs on the other. A quick glance at the graph of \( dW_\beta(t) \) (Figure 1b) can explain that: the last quarter or so of the process seems to exhibit slightly different statistical behavior, with larger standard deviation (which is found to be about 0.32), so there is some activity at the time scale 2 orders of magnitude below the total length.

Incidentally, note that the analysis above can be used to confirm that the Dow-Jones volume conforms with the BS model.
Figure 2: Graphs of the correlations of $dW_\beta(t)$ and the distances of $dW_\alpha(t)$, the Averaging function of $dW_\beta(t)$, and the Kolmogorov Distance between $dW_\beta(t)$ and a Gaussian.
3.2 Dow-Jones volume – Log

The analysis of $dS(t)/S(t)$ above yielded straightforward and simple results in terms of modeling, but it concerned really the relative increments of volume rather than volume itself; indeed, a nontrivial computation would still be needed from a model on $dS(t)/S(t)$ to obtain $\ln(S(t))$. We will now present a model for $\ln(S(t))$ itself.

We see that $\ln(S(t))$ exhibits a nontrivial drift in time (Figure 3a); in order to isolate it efficiently from the oscillatory component, we apply the envelope technique repeatedly, so that what we compute to be the drift is reasonably free of oscillation: that is, setting $x(t) = \ln(S(t))$, we compute $x = x_1 + \bar{x}_1 + \bar{x}_2 + x_2 = \ldots = \sum_{i=1}^{n} \bar{x}_i + x_n$ (see section 2.3.2) (we used $n = 6$ here). The resulting drift (Figure 3b) can be well approximated by 2 straight lines, and, unless we are interested for very old values, by only one. The random process appearing this time, though, (Figure 3c) looks “rougher” than the previous one; this is confirmed by the autocorrelation (Figure 4b), which exhibits extremely long range dependence, and the Averaging function (Figure 4a) whose obvious curvature suggests strong multifractality. The mean and standard deviation are $\mu = 0.0141$ and $\sigma = 0.3425$, and the KD of the marginal from the corresponding Gaussian is 0.0589, so the process can be considered to be Gaussian in a very good approximation.

To sum up, we can write $\ln(S(t)) = d(t) + \sigma W(t)$, where $d(t) \approx r(T - t)$, for some constants $r > 0$, $T > 0$. It should be emphasized, of course, that $W(t)$ and $dW_\beta(t)$ have very different properties because what they represent is completely different to begin with: the latter is a process of increments, while the former is the sum of such increments (hence the long range dependence); it can be checked directly that $dW(t)$ has no long range dependence, just like $dW_\beta(t)$.

3.3 Primary cycle of the Dow-Jones volume

We can use the highly nonlinear concept of Instantaneous Frequency [6] (IF) to detect periodicity in the Dow-Jones volume fluctuations. Consider $x(t) = |dW_\beta(t)|$ and find its envelope $x^\dagger(t)$ ($x^\dagger(t) = 0$ in this case); it follows that $0 \leq \frac{dx(t)}{x^\dagger(t)} \leq 1$, hence we can write $\frac{dx(t)}{x^\dagger(t)} = \sin(\phi(t))$ for some function $\phi$.

The frequency can be expressed as $\nu(t) = \frac{1}{2\pi} \left| \frac{d\phi(t)}{dt} \right|$, so we need to isolate $\phi$ by applying $\sin^{-1}$; as this function is multivalued $\phi$ is not unique, and some care must be taken so that the $\phi$ we choose is actually continuous: this is known in Signal Processing as the phase unwrapping problem. Failure to unwrap the phase correctly will certainly cause $\nu(t)$ to have artificial “spikes” at the discontinuities.
Figure 3: Graphs of \( \ln(S(t)) \), its drift and its oscillatory component

Figure 4: Averaging function and autocorrelation of the oscillatory component of \( \ln(S(t)) \)
Figure 5: Instantaneous frequency of $dW_\beta(t)$ and its drift

The resulting $\nu(t)$ is totally uninformative, and looks just like noise (Figure 5a); but we can use again the envelope to separate the drift from the oscillation, just as we did above in section 3.2 (we used $n = 7$ this time), and study exclusively the former (the mean frequency, that is): now we can clearly see (Figure 5b) an oscillation that looks like being of a single frequency, just like an ordinary sine wave, albeit deformed. The period can be measured as the mean distance between consecutive zero crossings or consecutive extrema; following either method we find oscillation periods roughly between 2 and 5 years; the former gives a mean of 3.45 years approximately, while the latter a mean of 3.25 years approximately. Both measurements are consistent with what is known in the finance community as the 4 years cycle of the Dow-Jones ([15], but also widely referred to on the web). This success is reason enough to investigate further the potential of Signal Processing methods in financial predictions; for example, using more refined implementations of the IF framework, we might be able to carry out an analysis detailed enough to isolate secondary oscillation cycles of the Dow-Jones volume, to start with. The possibilities are really endless.

4 Conclusion

Analysis and modeling of financial data can benefit from techniques used in Internet traffic modeling and more generally in Signal Processing. Although we just touched upon the subject in this work by working out the specific example of the Dow-Jones volume, the results seems promising and point to a new direction that opens many possibilities, as the mathematical tools used can handle a wide range of signals and random processes, and can provably yield accurate predictions.
References


Received: January 29, 2008