Econophysics review: I. Empirical facts

Anirban Chakraborti, Ioane Muni Toke, Marco Patriarca, Frédéric Abergel

Quantitative Finance
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/rquf20

Econophysics review: I. Empirical facts

Anirban Chakraborti [a], Ioane Muni Toke [a], Marco Patriarca [b c] & Fréédéric Abergel [a]

[a] Laboratoire de Mathématiques Appliquées aux Systèmes, Ecole Centrale Paris, 92290 Châtenay-Malabry, France

[b] National Institute of Chemical Physics and Biophysics, Räävala 10, 15042 Tallinn, Estonia

[c] Instituto de Fisica Interdisciplinaire y Sistemas Complejos (CSIC-UIB), E-07122 Palma de Mallorca, Spain

Available online: 24 Jun 2011

To cite this article: Anirban Chakraborti, Ioane Muni Toke, Marco Patriarca & Fréédéric Abergel (2011): Econophysics review: I. Empirical facts, Quantitative Finance, 11:7, 991-1012

To link to this article: http://dx.doi.org/10.1080/14697688.2010.539248

Quantitative Finance, Vol. 11, No. 7, July 2011, 991–1012

Econophysics review: I. Empirical facts

ANIRBAN CHAKRABORTI*†, IOANE MUNI TOKE†, MARCO PATRIARCA‡§ and FRÉDÉRIC ABERGEL†

†Laboratoire de Mathématiques Appliquées aux Systèmes, Ecole Centrale Paris, 92290 Châtenay-Malabry, France
‡National Institute of Chemical Physics and Biophysics, Rävala 10, 15042 Tallinn, Estonia
§Instituto de Fisica Interdisciplinaire y Sistemas Complejos (CSIC-UIB), E-07122 Palma de Mallorca, Spain

(*Received 21 September 2009; in final form 5 November 2010*)

This article and the companion paper aim at reviewing recent empirical and theoretical developments usually grouped under the term *Econophysics*. Since the name was coined in 1995 by merging the words 'Economics' and 'Physics', this new interdisciplinary field has grown in various directions: theoretical macroeconomics (wealth distribution), microstructure of financial markets (order book modeling), econometrics of financial bubbles and crashes, etc. We discuss the interactions between Physics, Mathematics, Economics and Finance that led to the emergence of Econophysics. We then present empirical studies revealing the statistical properties of financial time series. We begin the presentation with the widely acknowledged 'stylized facts', which describe the returns of financial assets—fat tails, volatility clustering, autocorrelation, etc.—and recall that some of these properties are directly linked to the way 'time' is taken into account. We continue with the statistical properties observed on order books in financial markets. For the sake of illustrating this review, (nearly) all the stated facts are reproduced using our own high-frequency financial database. Finally, contributions to the study of correlations of assets such as random matrix theory and graph theory are presented. The companion paper will review models in Econophysics from the point of view of agent-based modeling.

Keywords: Computational finance; Correlation; Econophysics; Empirical finance

1. Introduction

What is Econophysics? Fifteen years after the word 'Econophysics' was coined by H.E. Stanley by merging the words 'Economics' and 'Physics' at an international conference on Statistical Physics held in Kolkata in 1995, this is still a commonly asked question. Many still wonder how theories aimed at explaining the physical world in terms of particles could be applied to understand complex structures, such as those found in the social and economic behavior of human beings. In fact, physics as a natural science is supposed to be precise or specific; its predictive powers are based on the use of a few but universal properties of matter that are sufficient to explain many physical phenomena. But in social sciences, are there analogous precise universal properties known for human beings, who, in contrast to fundamental particles, are certainly not identical to each other in any respect? And what little amount of information would be sufficient to infer some of their complex behavior? There exists a positive drive to answer these questions. In the 1940s, Majorana took scientific interest in financial and economic systems. He wrote a pioneering paper on the essential analogy between statistical laws in physics and in social sciences (di Ettore Majorana 1942, Mantegna 2005, 2006). However, during the following decades, only a few physicists, such as Kadanoff (1971) and Montroll and Badger (1974), had an interest in research into social or economic systems. It was not until the 1990s that physicists started turning to this interdisciplinary subject, and in the last few years they have made many successful attempts at approaching problems in various fields of social sciences (e.g., de Oliveira *et al*. (1999), Chakrabarti *et al*. (2006) and Stauffer *et al*. (2006)). In particular, in Quantitative Economics and Finance, physics research

*Corresponding author. Email: anirban.chakraborti@ecp.fr

Quantitative Finance
ISSN 1469–7688 print/ISSN 1469–7696 online © 2011 Taylor & Francis
http://www.informaworld.com
DOI: 10.1080/14697688.2010.539248

has begun to be complementary to the most traditional approaches such as mathematical (stochastic) finance. These various investigations, based on methods imported from (or also used in) physics, are the subject of the present paper.

1.1. Bridging physics and economics

Economics deals with how societies efficiently use their resources to produce valuable commodities and distribute them among different people or economic agents (Keynes 1973, Samuelson 1998). It is a discipline related to almost everything around us, starting from the marketplace through the environment to the fate of nations. At first sight this may seem a very different situation from that of physics, whose birth as a well-defined scientific theory is usually associated with the study of particular mechanical objects moving with negligible friction, such as falling bodies and planets. However, a deeper comparison shows many more analogies than differences. On a general level, both economics and physics deal with 'everything around us', but from different perspectives. On a practical level, the goals of both disciplines can be either purely theoretical in nature or strongly oriented towards the improvement of the quality of life. On a more technical side, analogies often become equivalences. Let us give some examples.

Statistical mechanics has been defined as the

> ''branch of physics that combines the principles and procedures of statistics with the laws of both classical and quantum mechanics, particularly with respect to the field of thermodynamics. It aims to predict and explain the measurable properties of macroscopic systems on the basis of the properties and behaviour of the microscopic constituents of those systems.''†

The tools of statistical mechanics or statistical physics (Landau 1965, Reif 1985, Pathria 1996), which include extracting the average properties of a macroscopic system from the microscopic dynamics of the system, are believed will prove useful for an economic system. Indeed, even though it is difficult or almost impossible to write down the 'microscopic equations of motion' for an economic system with all the interacting entities, economic systems may be investigated on various size scales. Therefore, an understanding of the global behavior of economic systems seems to need concepts such as stochastic dynamics, correlation effects, self-organization, self-similarity and scaling, and for their application we do not have to go into the detailed 'microscopic' description of the economic system.

Chaos theory has had some impact on Economics modeling (e.g., in the work of Brock and Hommes (1998) and Chiarella *et al*. (2006)). The theory of disordered systems has also played a key role in Econophysics and the study of 'complex systems'. The term 'complex systems' was coined to cover the great variety of such systems that include examples from physics, chemistry, biology and also social sciences. The concepts and methods of statistical physics turned out to be extremely useful when applied to these diverse complex systems, including economic systems. Many complex systems in natural and social environments share the characteristics of competition among interacting agents for resources and their adaptation to a dynamically changing environment (Arthur 1999, Parisi 1999). Hence, the concept of disordered systems helps, for instance, to go beyond the concept of a representative agent, an approach prevailing in much of (macro)economics and criticized by many economists (see, e.g., Kirman (1992) and Gallegati and Kirman (1999)). Minority games and their physical formulations have been exemplary.

Physics models have also helped to develop new theories to explain older observations in Economics. The Italian social economist Pareto investigated a century ago the wealth of individuals in a stable economy (Pareto 1897) by modeling them with the distribution $P(>x) \sim x^{-\alpha}$, where $P(>x)$ is the number of people having an income greater than or equal to x and α is an exponent (now known as the Pareto exponent) which he estimated to be 1.5. To explain such empirical findings, physicists have come up with some very elegant and intriguing kinetic exchange models in recent times, and we review these developments in the companion article. Although the economic activities of the agents are driven by various considerations such as 'utility maximization', the eventual exchanges of money in any trade can be simply viewed as money/wealth conserving two-body scattering, as in the entropy maximization based kinetic theory of gases. This qualitative analogy seems to be quite old and both economists and natural scientists have already noted it in various contexts (Saha and Srivastava 1950). Recently, the equivalence between the two maximization principles has been established quantitatively (Chakrabarti and Chakrabarti 2010).

Let us discuss another example of the similarities of interest and tools in Physics and Economics. The frictionless systems that mark the early history of physics were soon recognized to be rare cases: not only on the microscopic scale—where they obviously represent an exception due to the unavoidable interactions with the environment—but also on the macroscopic scale, where fluctuations of internal or external origin make prediction of their time evolution impossible. Thus equilibrium and non-equilibrium statistical mechanics, the theory of stochastic processes, and the theory of chaos became the main tools for studying real systems as well as an important part of the theoretical framework of modern physics. Very interestingly, the same mathematical tools have presided over the growth of classic modeling in Economics and more particularly in modern Finance. Following the work of Mandelbrot and Fama in the 1960s, physicists from 1990 onwards have studied the

†*Encyclopædia Britannica*. Retrieved June 11, 2010, from Encyclopædia Britannica Online.

fluctuation of prices and universalities in the context of scaling theories, etc. These links open the way for the use of a physics approach in Finance, complementary to the widespread mathematical approach.

1.2. Econophysics and finance

Mathematical finance has benefited considerably in the past 30 years from modern probability theory—Brownian motion, martingale theory, etc. Financial mathematicians are often proud to recall the most well-known source of the interactions between Mathematics and Finance: five years before Einstein's seminal work, the theory of Brownian motion was first formulated by the French mathematician Bachelier in his doctoral thesis (Bachelier 1900, Boness 1964, Haberman and Sibbett 1995), in which he used this model to describe price fluctuations at the Paris Bourse. Bachelier had even given a course as a 'free professor' at the Sorbonne University with the title: 'Probability calculus with applications to financial operations and analogies with certain questions from physics' (see the historical articles of Courtault *et al.* (2000), Taqqu (2001) and Forfar (2002)).

Then Itō, following the works of Bachelier, Wiener, and Kolmogorov, among many others, formulated the presently known Itō calculus (Itō and McKean 1996). Geometric Brownian motion, belonging to the class of Itō processes, later became an important ingredient of models in Economics (Osborne 1959, Samuelson 1965), and in the well-known theory of option pricing (Black and Scholes 1973, Merton 1973). In fact, stochastic calculus of diffusion processes combined with classical hypotheses in Economics led to the development of the *arbitrage pricing theory* (Duffie 1996, Follmer and Schied 2004). The deregulation of financial markets at the end of the 1980s led to the exponential growth of the financial industry. Mathematical finance followed the trend: stochastic finance with diffusion processes and exponential growth of financial derivatives have had intertwined developments. Finally, this relationship was carved in stone when the Nobel prize was presented to M.S. Scholes and R.C. Merton in 1997 (F. Black died in 1995) for their contribution to the theory of option pricing and their celebrated 'Black–Scholes' formula.

However, this whole theory is closely linked to classical economics hypotheses and has not been sufficiently grounded with empirical studies of financial time series. The Black–Scholes hypothesis of Gaussian log-returns of prices is in strong disagreement with empirical evidence. Mandelbrot (1960, 1963) was one of the first to observe a clear departure from Gaussian behavior for these fluctuations. It is true that, within the framework of stochastic finance and martingale modeling, more complex processes have been considered in order to take into account certain empirical observations: jump processes (see, e.g., Cont and Tankov (2004) for a textbook treatment) and stochastic volatility (e.g., Heston (1993) and Gatheral (2006)), in particular. But recent events in financial markets and the succession of financial crashes (see, e.g., Kindleberger and Aliber (2005) for a historical perspective) should lead scientists to rethink the basic concepts of modeling. This is where Econophysics is expected to come into play. During the past decades, the financial landscape has been changing dramatically: deregulation of markets and the growing complexity of products. From a technical point of view, the ever increasing speed and decreasing cost of computational power and networks have lead to the emergence of huge databases that record all transactions and order book movements up to the millisecond. The availability of these data should lead to models that are better founded empirically. Statistical facts and empirical models will be reviewed in this article and the companion paper. The recent turmoil on financial markets and the 2008 crash seem to plead for new models and approaches. The Econophysics community thus has an important role to play in future financial market modeling, as suggested by contributions from Bouchaud (2008), Farmer and Foley (2009), and Lux and Westerhoff (2009).

1.3. A growing interdisciplinary field

The chronological development of Econophysics has been well covered in the book of Roehner (2002). Here it is worth mentioning a few landmarks. The first article on the analysis of finance data that appeared in a physics journal was that of Mantegna (1991). The first conference on Econophysics was held in Budapest in 1997 and has since been followed by numerous schools, workshops and regular series of meetings: APFA (Application of Physics to Financial Analysis), WEHIA (Workshop on Economic Heterogeneous Interacting Agents), and Econophys-Kolkata, amongst others. In recent years the number of papers has increased dramatically; the community has grown rapidly and several new directions of research have opened up. Renowned physics journals such as *The Reviews of Modern Physics*, *Physical Review Letters*, *Physical Review* E, *Physica* A, *Europhysics Letters*, *European Physical Journal* B, *International Journal of Modern Physics* C, etc. now publish papers in this interdisciplinary area. Economics and mathematical finance journals, especially *Quantitative Finance*, receive contributions from many physicists. The interested reader can also follow the developments quite well from the preprint server www.arxiv.org. In fact, recently a new section called quantitative finance has been added to it. One could also visit the web sites of the *Econophysics Forum* (www.unifr.ch/econophysics) and *Econophysics.Org* (www.econophysics.org). Previous texts addressing Econophysics issues, such as those of Bouchaud and Potters (2000, Mantegna and Stanley (2007) and Gabaix (2009), may be complementary to the present review. The first textbook in Econophysics (Sinha *et al.* 2010) has been published.

1.4. Organization of the review

This article reviews recent empirical and theoretical developments that use tools from Physics in the fields of Economics and Finance. In section 2, empirical studies revealing the statistical properties of financial time series

are reviewed. We present the widely acknowledged 'stylized facts' describing the distribution of the returns of financial assets. In section 3 we continue with the statistical properties observed on order books in financial markets. We reproduce most of the stated facts using our own high-frequency financial database. In the last part of this article (section 4), we review contributions on correlation on financial markets, among which are the computation of correlations using high-frequency data, analyses based on random matrix theory and the use of correlations to build economics taxonomies. In the companion paper, Econophysics models are reviewed from the point of view of agent-based modeling. Using previous work originally presented in the fields of behavioral finance and market microstructure theory, econophysicists have developed agent-based models of order-driven markets that have been extensively reviewed. We then turn to models of wealth distribution where an agent-based approach also prevails. As mentioned above, Econophysics models help bring a new look at certain Economics observations, and advances based on kinetic theory models are presented. Finally, a detailed review of game theory models and the now classic minority games composes the final part.

2. Statistics of financial time series: Price, returns, volumes, volatility

Recording a sequence of prices of commodities or assets produces what is called a time series. The analysis of financial time series has been of great interest, not only to practitioners (an empirical discipline), but also to theoreticians for making inferences and predictions. The inherent uncertainty in financial time series and the theory makes it especially interesting to economists, statisticians and physicists (Tsay 2005).

Different kinds of financial time series have been recorded and studied for decades, but the scale changed 20 years ago. The computerization of stock exchanges that took place all over the world in the mid 1980s and early 1990s has lead to an explosion in the amount of data recorded. Nowadays, all transactions on financial markets are recorded *tick-by-tick*, i.e. every event on a stock is recorded with a time stamp defined up to the millisecond, leading to huge amounts of data. For example, as of today (2010), the Reuters Datascope Tick History (RDTH) database records roughly 25 gigabytes of data *every trading day*.

Prior to this improvement in recording market activity, statistics could only be computed using daily data at best. Now scientists can compute intra-day statistics with high frequency. This allows us to check known properties on new time scales (see, e.g., section 2.2), but also implies that special care needs to be taken in the treatment (see, for example, the computation of correlation at high frequency in section 4.1).

It is a formidable task to present an exhaustive review on this topic, but we try to give a flavor of some of the aspects in this section.

2.1. 'Stylized facts' of financial time series

The concept of 'stylized facts' was introduced in macroeconomics by Kaldor (1961), who advocated that a scientist studying a phenomenon "should be free to start off with a stylized view of the facts". In his work, Kaldor isolated several statistical facts characterizing macroeconomic growth over long periods and in several countries, and took these robust patterns as a starting point for theoretical modeling.

This expression has thus been adopted to describe the empirical facts that arose in statistical studies of financial time series and that seem to be persistent across various time periods, places, markets, assets, etc. One can find many different lists of these facts in several reviews (e.g., Bollerslev *et al.* (1994), Pagan (1996), Guillaume *et al.* (1997) and Cont (2001)). We choose in this article to present a minimum set of facts now widely acknowledged, at least for the prices of equities.

2.1.1. Fat-tailed empirical distribution of returns. Let p_t be the price of a financial asset at time t. We define its return over a period of time τ to be

$$r_\tau(t) = \frac{p(t+\tau) - p(t)}{p(t)} \approx \log(p(t+\tau)) - \log(p(t)). \quad (1)$$

It has been largely observed—starting with Mandelbrot (1963) (see, e.g., Gopikrishnan *et al.* (1999) for tests on more recent data)—and it is the first stylized fact, that the empirical distributions of financial returns and log-returns are fat-tailed. In figure 1 we reproduce the empirical density function of normalized log-returns from Gopikrishnan *et al.* (1999) computed on the S&P 500 index. In addition, we plot similar distributions for normalized returns on a liquid French stock (BNP Paribas) with $\tau = 5$ minutes. This graph is computed by sampling a set of tick-by-tick data from 9:05 a.m. to 5:20 p.m. between 1 January 2007 and 30 May 2008, i.e. 356 days of trading. Except where mentioned otherwise, this data set will be used for all empirical graphs in this section. Figure 2 reproduces the cumulative distribution on a log–log scale from Gopikrishnan *et al.* (1999). We also show the same distribution on a linear–log scale computed on our data for a larger time scale of $\tau = 1$ day, showing similar behavior.

Many studies report similar observations on different sets of data. For example, using two years of data on more than 1000 US stocks, Gopikrishnan *et al.* (1998) find that the cumulative distribution of returns asymptotically follows the power law $F(r_\tau) \sim |rt|^{-\alpha}$ with $\alpha > 2$ ($\alpha \approx 2.8$–3). With $\alpha > 2$, the second moment (the variance) is well-defined, excluding stable laws with infinite variance. There have been various suggestions for the form of the distribution: generalized hyperbolic Student-t, normal inverse Gaussian, exponentially truncated stable, and others, but no general consensus exists on the exact form of the tails. Although being the most widely acknowledged and the most elementary, this stylized fact is not easily met by all financial modeling. Gabaix *et al.* (2006) and Wyart and Bouchaud (2007) recall that

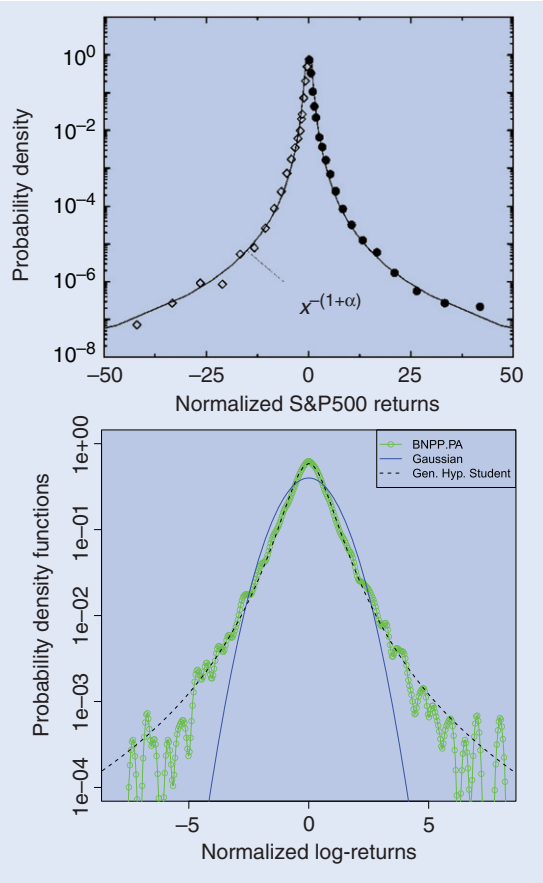

Figure 1. (Top) Empirical probability density function of the normalized 1-min S&P 500 returns between 1984 and 1996. Reproduced from Gopikrishnan *et al.* (1999). (Bottom) Empirical probability density function of BNP Paribas normalized log-returns over a period of $\tau = 5$ min.

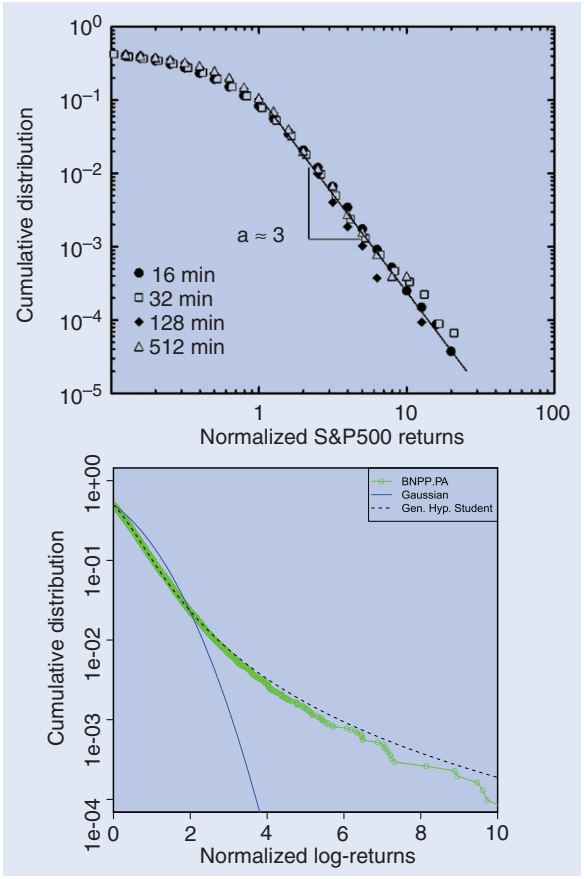

Figure 2. Empirical cumulative distributions of S&P 500 daily returns. (Top) Reproduced from Gopikrishnan *et al.* (1999), on a log–log scale. (Bottom) Computed using official daily close price between 1 January 1950 and 15 June 2009, i.e. 14,956 values, on a linear–log scale.

efficient market theory has difficulties in explaining fat tails. Lux and Sornette (2002) have shown that models known as 'rational expectation bubbles', popular in economics, produce very fat-tailed distributions ($\alpha < 1$) that are in disagreement with the statistical evidence.

2.1.2. Absence of autocorrelations of returns. Figure 3 plots the autocorrelation of log-returns defined as $\rho(T) \sim \langle r_\tau(t+T) r_\tau(t) \rangle$ with $\tau = 1$ minute and 5 minutes. We observe here, as is widely known (see, e.g., Pagan (1996) and Cont *et al.* (1997)), that there is no evidence of a correlation between successive returns, which is the second 'stylized fact'. The autocorrelation function decays very rapidly to zero, even for a few lags of 1 minute.

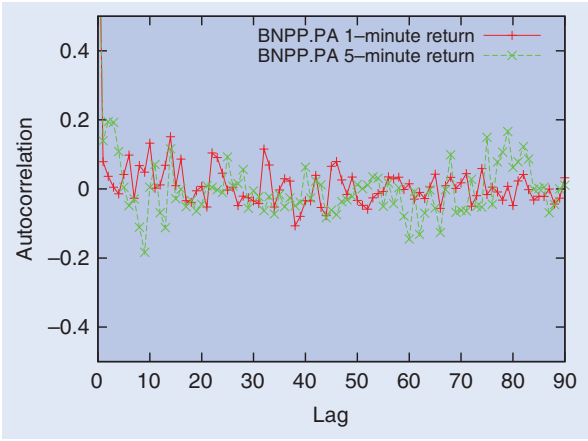

Figure 3. Autocorrelation function of BNPP.PA returns.

2.1.3. Volatility clustering. The third 'stylized fact' that we present here is of primary importance. The absence of a correlation between returns must not be mistaken for a property of independence and identical distribution: price fluctuations are not identically distributed and the properties of the distribution change with time. In particular, absolute returns or squared returns exhibit a long-range slowly decaying autocorrelation function. This phenomenon is widely known as 'volatility clustering', and was formulated by Mandelbrot (1963) as "large changes tend to be followed by large changes—of either sign—and small changes tend to be followed by small changes".

Figure 4 plots the autocorrelation function of absolute returns for $\tau = 1$ minute and 5 minutes. The levels of autocorrelation at the first lags vary wildly with the parameter τ. For our data, it is found to be maximum (more than 70% at the first lag) for returns sampled every

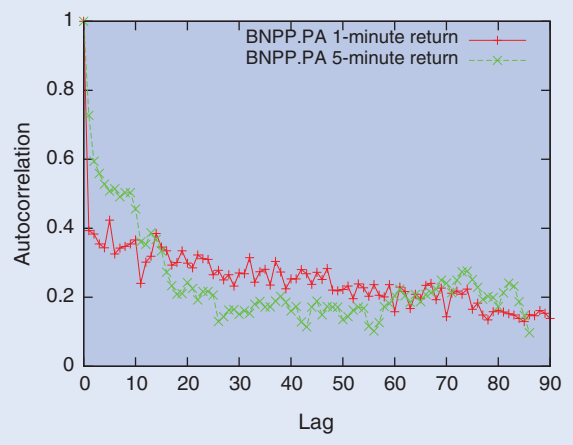

Figure 4. Autocorrelation function of BNPP.PA absolute returns.

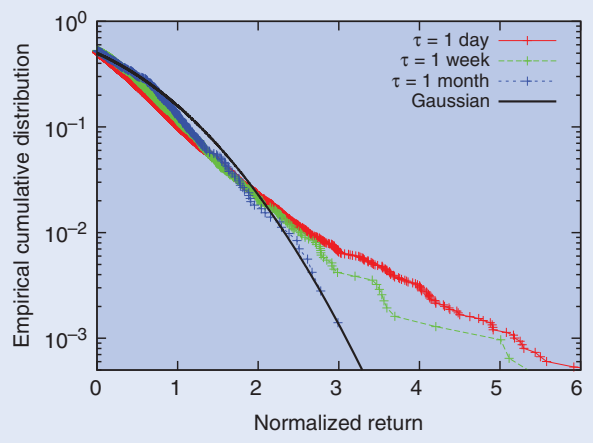

Figure 5. Distribution of log-returns of S&P 500 daily, weekly and monthly returns. Same data set as figure 2 (bottom).

five minutes. However, whatever the sampling frequency, autocorrelation is still above 10% after several hours of trading. For these data, we can grossly fit a power law decay with exponent 0.4. Other empirical tests report exponents between 0.1 and 0.3 (Cizeau *et al.* 1997, Cont *et al.* 1997, Liu *et al.* 1997).

2.1.4. Aggregational normality. It has been observed that as one increases the time scale over which the returns are calculated, the fat-tail property becomes less pronounced, and the distribution approaches the Gaussian form, which is the fourth 'stylized-fact'. This cross-over phenomenon is documented by Kullmann *et al.* (1999), who study the evolution of the Pareto exponent of the distribution with the time scale. Figure 5 plots these standardized distributions for the S&P 500 index between 1 January 1950 and 15 June 2009. It is clear that the more the time scale increases, the more Gaussian the distribution becomes. The fact that the shape of the distribution changes with τ makes it clear that the random process underlying prices must have a non-trivial temporal structure.

2.2. Getting the right 'time'

2.2.1. Four ways to measure 'time'. In the previous section, all the 'stylized facts' have been presented in *physical time*, or *calendar time*, i.e. the time series were indexed, as we expect them to be, in hours, minutes, seconds, milliseconds. Let us recall here that the tick-by-tick data available on financial markets all over the world is time-stamped up to the millisecond, but the order of magnitude of the guaranteed precision is much larger, usually one second or a few hundred milliseconds.

Calendar time is the time usually used to compute the statistical properties of financial time series. This means that computing these statistics involves sampling, which might be a delicate thing to do when dealing, for example, with several stocks with different liquidity. Therefore, three other ways to keep track of time may be used.

Let us first introduce *event time*. Using this count, time is increased by one unit each time one order is submitted to the observed market. This framework is natural when dealing with the simulation of financial markets, as will be shown in the companion paper. The main outcome of event time is its 'smoothing' of the data. In event time, intra-day seasonality (lunch break) or an outburst of activity consequent to some news are smoothed in the time series, since we always have one event per time unit.

Now, when dealing with time series of prices, another count of time might be relevant, and we call it *trade time* or *transaction time*. Using this count, time is increased by one unit each time a transaction occurs. The advantage of this count is that limit orders submitted far away in the order book, and that may thus be of lesser importance with respect to the price series, do not increase the clock by one unit.

Finally, proceeding with focusing on important events to increase the clock, we can use *tick time*. Using this count, time is increased by one unit each time the price changes. Thus consecutive market orders that progressively 'eat' liquidity until the first best limit is removed in an order book are counted as one unit time.

Let us finish by noting that, with these definitions, when dealing with mid prices, or bid and ask prices, a time series in event time can easily be extracted from a time series in calendar time. Furthermore, one can always extract a time series in trade time or in price time from a time series in event time. However, one cannot extract a series in price time from a series in trade time, as the latter ignores limit orders that are submitted inside the spread, and thus change mid, bid or ask prices without any transaction taking place.

2.2.2. Revisiting 'stylized facts' with a new clock. Using the right clock might be of primary importance when dealing with statistical properties and estimators. For example, Griffin and Oomen (2008) investigate the standard realized variance estimator (see section 4.1) in trade time and tick time. Muni Toke (2010) also recalls that the differences observed on a spread distribution in trade time and physical time are meaningful. In this

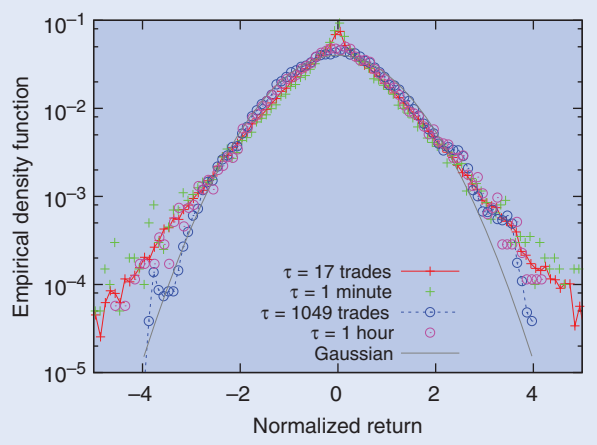

Figure 6. Distribution of log-returns of stock BNPP.PA. This empirical distribution is computed using data from 1 April 2007 to 31 May 2008.

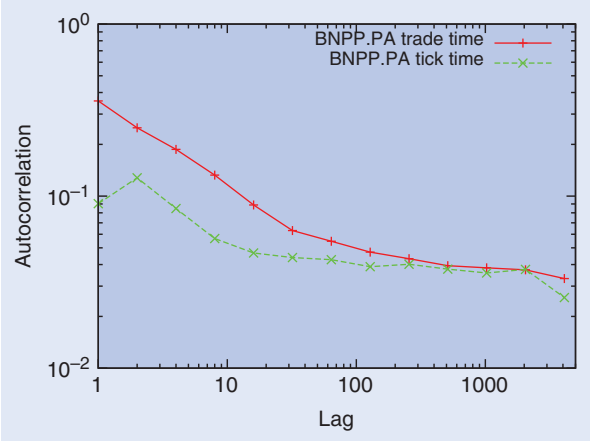

Figure 7. Autocorrelation of trade signs for stock BNPP.PA.

section we compute some statistics complementary to those presented in the previous section 2.1 and show the role of the clock in the studied properties.

Aggregational normality in trade time. We have seen above that when the sampling size increases, the distribution of the log-returns tends to be more Gaussian. This property is much better seen using trade time. Figure 6 plots the distributions of the log-returns for BNP Paribas stock using 2-month-long data in calendar time and trade time. Over this period, the average number of trades per day is 8562, so that 17 trades (respectively 1049 trades) corresponds to an average calendar time step of 1 minute (respectively 1 hour). We observe that the distribution of returns sampled every 1049 trades is much more Gaussian than that sampled every 17 trades (aggregational normality), and that it is also more Gaussian that that sampled every 1 hour (faster convergence in trade time). Note that this property appears to be valid in a multi-dimensional setting (Huth and Abergel 2009).

Autocorrelation of trade signs in tick time. It is well known that the series of signs of the trades on a given stock (usual convention: $+1$ for a transaction at the ask price, -1 for a transaction at the bid price) exhibit a large autocorrelation. It was observed by Lillo and Farmer (2004), for example, that the autocorrelation function of the signs of trades (ϵ_n) was a slowly decaying function in $n^{-\alpha}$, with $\alpha \approx 0.5$. We compute this statistics for the trades on BNP Paribas stock from 1 January 2007 to 31 May 2008. We plot the result in figure 7. We find that the first values for short lags are about 0.3, and that the log–log plot clearly shows some power-law decay with $\alpha \approx 0.7$.

A very plausible explanation for this phenomenon relies on the execution strategies of certain major brokers in a given market. These brokers have large transactions to execute for certain clients. In order to avoid the market moving because of a large order (see section 3.6 on market impact), they tend to split large orders into small orders. We believe that this strategy explains, at least partly, the large autocorrelation observed. Using data on markets where orders are publicly identified and linked to a given broker, it can be shown that the autocorrelation function of the order signs *of a given broker* is even higher. See Bouchaud *et al.* (2009) for a review of these facts and associated theories.

We present here further evidence supporting this explanation. We compute the autocorrelation function of order signs *in tick time*, i.e. taking into account only transactions that make the price change. The results are plotted in figure 7. We find that the first values for short lags are about 0.10, which is much smaller than the values observed with the previous time series. This supports the idea that many small transactions progressively 'eat' the available liquidity at the best quotes. Note however that, even in tick time, the correlation also remains positive for large lags.

2.2.3. Correlation between volume and volatility. Investigating time series of cotton prices, Clark (1973) noted that "trading volume and price change variance seem to have a curvilinear relationship". *Trade time* allows us to obtain a better view of this property: Plerou *et al.* (2000) and Silva and Yakovenko (2007), among others, show that the variance of log-returns after N trades, i.e. over a time period of N in trade time, is proportional to N. We confirm this observation by plotting the second moment of the distribution of log-returns after N trades as a function of N for our data, as well as the average number of trades and the average volatility in a given time interval. The results are shown in figures 8 and 9.

These results should be placed in relation to those presented by Gopikrishnan *et al.* (2000b), who studied the statistical properties of the number of shares traded $Q_{\Delta t}$ for a given stock in a fixed time interval Δt. They analysed transaction data for the largest 1000 stocks for the two-year period 1994–95, using a database that recorded every transaction for all securities in three major US stock markets. They found that the distribution $P(Q_{\Delta t})$ displayed a power-law decay, as shown in figure 10, and that the time correlations in $Q_{\Delta t}$ displayed long-range persistence. Further, they investigated the relation between $Q_{\Delta t}$

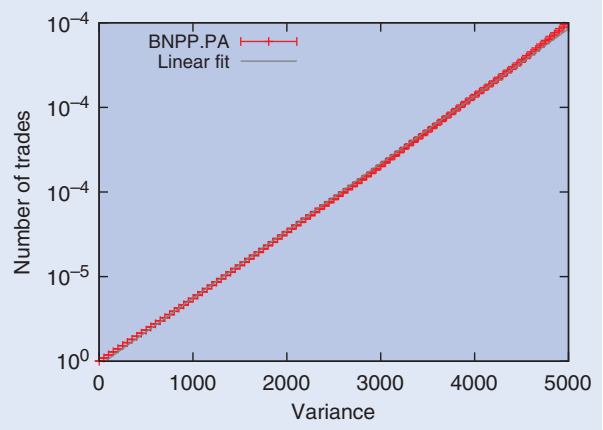

Figure 8. Second moment of the distribution of returns over N trades for the stock BNPP.PA.

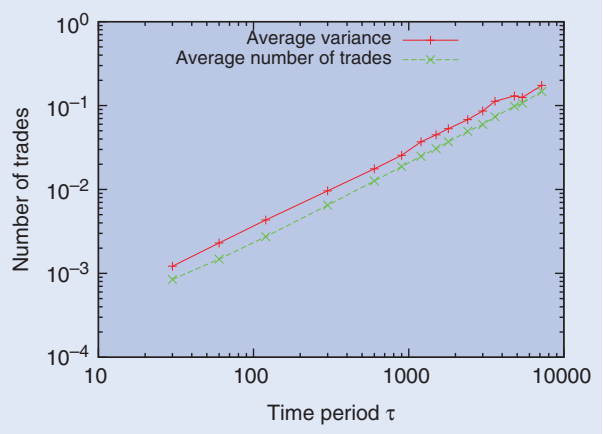

Figure 9. Average number of trades and average volatility in time period τ for the stock BNPP.PA.

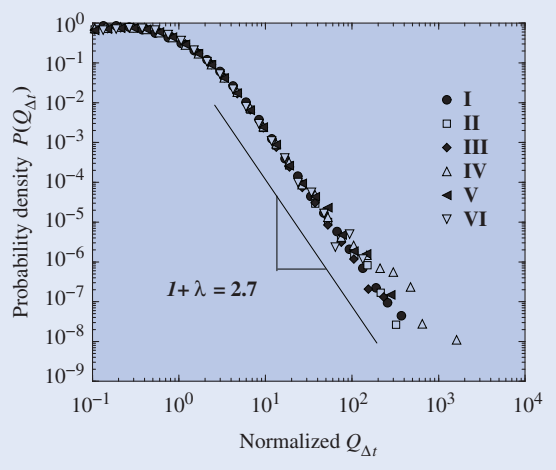

Figure 10. Distribution of the number of shares traded $Q_{\Delta t}$. Adapted from Gopikrishnan *et al.* (2000b).

and the number of transactions $N_{\Delta t}$ in time interval Δt and found that the long-range correlations in $Q_{\Delta t}$ were largely due to those of $N_{\Delta t}$. Their results are consistent with the interpretation that the large equal-time correlations previously found between $Q_{\Delta t}$ and the absolute value of the price change $|G_{\Delta t}|$ (related to volatility) are largely due to $N_{\Delta t}$. Therefore, studying the variance of price changes in *trade time* suggests that the number of trades is a good proxy for the unobserved volatility.

2.2.4. A link with stochastic processes: Subordination. These empirical facts (aggregational normality in trade time, relationship between volume and volatility) reinforce the interest in models based on the subordination of stochastic processes, which was introduced to financial modeling by Clark (1973).

Let us introduce it here. Assuming proportionality between the variance $\langle x \rangle_\tau^2$ of the centered returns x and the number of trades N_τ over time period τ, we can write

$$\langle x \rangle_\tau^2 = \alpha N_\tau. \tag{2}$$

Therefore, assuming normality in trade time, we can write the density function of the log-returns after N trades as

$$f_N(x) = \frac{e^{-x^2/2\alpha N}}{\sqrt{2\pi\alpha N}}. \tag{3}$$

Finally, denoting $K_\tau(N)$ as the probability density function of having N trades in time period τ, the distribution of log-returns in calendar time can be written as

$$P_\tau(x) = \int_0^\infty \frac{e^{-x^2/2\alpha N}}{\sqrt{2\pi\alpha N}} K_\tau(N) dN. \tag{4}$$

This is subordination of the Gaussian process x_N using the number of trades N_τ as the *directing process*, i.e. as the new clock. With this kind of modelization, it is expected, since P_N is Gaussian, that the observed non-Gaussian behavior will come from $K_\tau(N)$. For example, specific choices of the directing processes may lead to a symmetric stable distribution (Feller 1968). Clark (1973) tested empirically a log-normal subordination with time series of cotton prices. In a similar way, Silva and Yakovenko (2007) find that exponential subordination with a kernel,

$$K_\tau(N) = \frac{1}{\eta\tau} e^{-N/\eta\tau}, \tag{5}$$

is in good agreement with empirical data. If the orders were submitted to the market in an independent way and at a constant rate η, then the distribution of the number of trades per time period τ should be a Poisson process with intensity $\eta\tau$. Therefore, the empirical fit of equation (5) is inconsistent with such a simplistic hypothesis of the distribution of the time of arrivals of orders. We will suggest in the next section possible distributions that fit our empirical data.

3. Statistics of order books

The computerization of financial markets in the second half of the 1980s provided empirical scientists with easier access to extensive data on order books. Biais *et al.* (1995) is an early study of the new data flows on the newly (at that time) computerized Paris Bourse.

Variables crucial to a fine modeling of order flows and dynamics of order books are studied: time of arrival of orders, placement of orders, size of orders, shape of order book, etc. Many subsequent papers offer complementary empirical findings and modeling (e.g., Gopikrishnan *et al.* (2000a), Challet and Stinchcombe (2001), Maslov and Mills (2001), Bouchaud *et al.* (2002) and Potters and Bouchaud (2003)). Before going further into our review of available models, we summarize some of these empirical facts.

For each of the enumerated properties, we present new empirical plots. We use Reuters tick-by-tick data on the Paris Bourse. We select four stocks: France Telecom (FTE.PA), BNP Paribas (BNPP.PA), Societe Générale (SOGN.PA) and Renault (RENA.PA). For any given stock, the data display time-stamps, traded quantities, traded prices, the first five best-bid limits and the first five best-ask limits. From now on, we will denote by $a_i(t)$ (respectively $(b_j(t))$ the price of the ith limit at the ask (respectively the jth limit at the bid). Except where stated otherwise, all statistics are computed using all trading days from 1 October 2007 to 30 May 2008, i.e. 168 trading days. On a given day, orders submitted between 9:05 a.m. and 5:20 p.m. are taken into account, i.e. the first and last minutes of each trading day are removed.

Note that we do not deal in this section with the correlations of the signs of trades, since statistical results on this have already been treated in section 2.2.2. Note also that although most of these facts are widely acknowledged, we will not describe them as new 'stylized facts for order books', since their ranges of validity are still to be checked among various products/stocks, markets and epochs, and strong properties need to be properly extracted and formalized from these observations. However, we will keep them in mind as we go through the new trend of 'empirical modeling' of order books.

Finally, let us recall that the markets we are dealing with are electronic order books with no official market maker, in which orders are submitted in a double auction and executions follow price/time priority. This type of exchange has now been adopted nearly all over the world, but this was not obvious when computerization was incomplete. Different market mechanisms have been widely studied in the microstructure literature (see, e.g., Garman (1976), Kyle (1985), Glosten (1994), Biais *et al.* (1997), O'Hara (1997) and Hasbrouck (2007)). We will not review this literature here (except for Garman (1976) in our companion paper), as this would be too large a digression. However, such literature is linked in many respects to the problems reviewed in this paper.

3.1. Time of arrival of orders

As explained in the previous section, the choice of the time count might be of prime importance when dealing with 'stylized facts' of empirical financial time series. When reviewing the subordination of stochastic processes (Clark 1973, Silva and Yakovenko 2007), we have seen

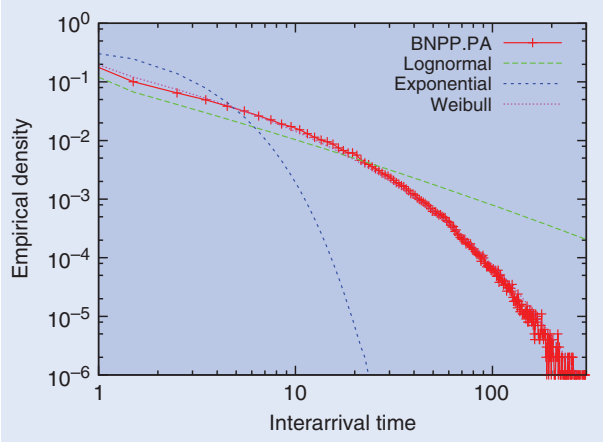

Figure 11. Distribution of inter-arrival times for stock BNPP.PA on a log scale.

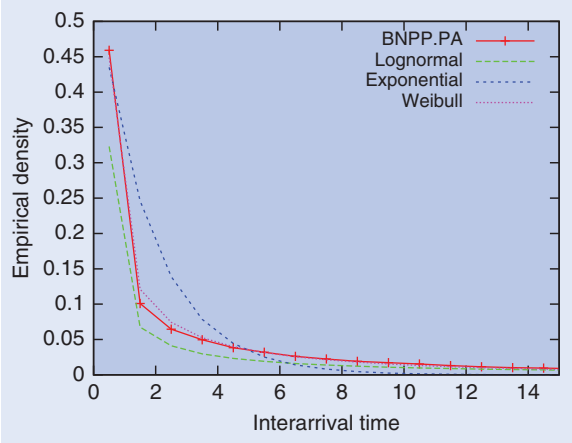

Figure 12. Distribution of inter-arrival times for stock BNPP.PA (main body, linear scale).

that the Poisson hypothesis for the arrival times of orders is not empirically verified.

We compute the empirical distribution for inter-arrival times—or durations—of market orders on the stock BNP Paribas using the data set described in the previous section. The results are plotted in figures 11 and 12, both on linear and log scales. It is clearly observed that the exponential fit is not a good one. We checked, however, that the Weibull distribution fit is potentially a very good one. Weibull distributions were suggested, for example, by Ivanov *et al.* (2004). Politi and Scalas (2008) also obtain good fits with q-exponential distributions.

In the Econometrics literature, these observations of non-Poissonian arrival times have given rise to a large trend of modeling irregular financial data. Engle and Russell (1997) and Engle (2000) introduced autoregressive condition duration or intensity models that may help model these processes of order submission (see Hautsch (2004) for a textbook treatment).

Using the same data, we compute the empirical distribution of the number of transactions in a given time period τ. The results are plotted in figure 13. It seems that the log-normal and gamma distributions are both good candidates, however neither really describes the

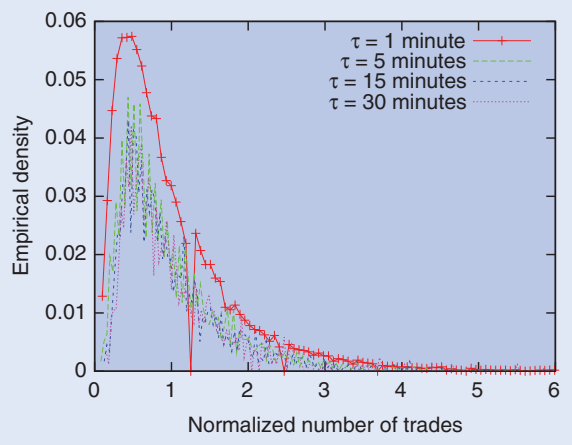

Figure 13. Distribution of the number of trades in a given time period τ for stock BNPP.PA. This empirical distribution is computed using data from 1 October 2007 to 31 May 2008.

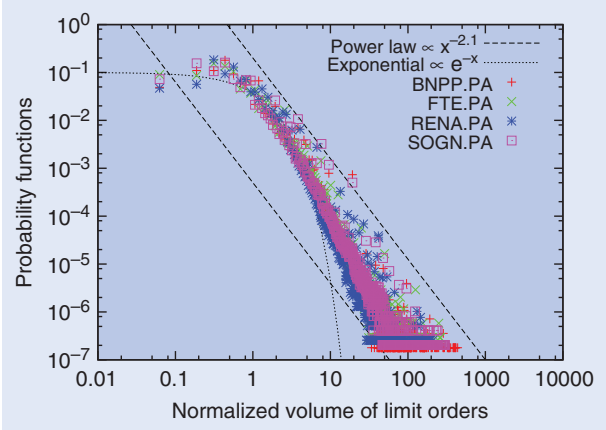

Figure 15. Distribution of normalized volumes of limit orders. Quantities are normalized by their mean.

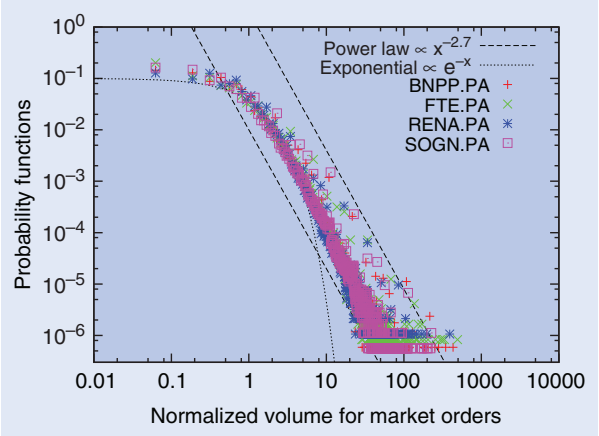

Figure 14. Distribution of volumes of market orders. Quantities are normalized by their mean.

empirical result, suggesting a complex structure for the arrival of orders. A similar result for Russian stocks was presented by Dremin and Leonidov (2005).

3.2. *Volume of orders*

Empirical studies show that the unconditional distribution of order size is very complex to characterize. Gopikrishnan *et al.* (2000a) and Maslov and Mills (2001) observe a power-law decay with exponent $1+\mu \approx 2.3$–2.7 for market orders and $1+\mu \approx 2.0$ for limit orders. Challet and Stinchcombe (2001) report a clustering property: orders tend to have a 'round' size in packages of shares, and clusters are observed around the 100s and 1000s. At the time of writing, no consensus has emerged with respect to the proposed models, and it is plausible that such a distribution varies very widely with the product and market.

Figure 14 plots the distribution of the volume of market orders for the four stocks composing our benchmark. Quantities are normalized by their mean. A power-law coefficient is estimated by a Hill estimator (see, e.g., Hill (1975) and de Haan *et al.* (2000)). We find a power law with exponent $1+\mu \approx 2.7$, which confirms the studies previously cited. Figure 15 displays the same distribution for limit orders (of all available limits). We find an average value of $1+\mu \approx 2.1$, consistent with previous studies. However, we note that the power law is a poorer fit in the case of limit orders: data normalized by their mean collapse badly on a single curve, and computed coefficients vary with the stock.

3.3. *Placement of orders*

3.3.1. Placement of arriving limit orders. Bouchaud *et al.* (2002) observe a broad power-law placement around the best quotes on French stocks, confirmed by Potters and Bouchaud (2003) for US stocks. The observed exponents are quite stable across stocks, but exchange dependent: $1+\mu \approx 1.6$ on the Paris Bourse, $1+\mu \approx 2.0$ on the New York Stock Exchange, and $1+\mu \approx 2.5$ on the London Stock Exchange. Mike and Farmer (2008) propose fitting the empirical distribution with a Student distribution with 1.3 degree of freedom.

We plot the distribution of the following quantity computed on our data set, i.e. using only the first five limits of the order book: $\Delta p = b_0(t-) - b(t)$ (respectively $\{a(t) - a_0(t-)\}$ if a bid (respectively an ask) order arrives at price $b(t)$ (respectively $a(t)$), where $b_0(t-)$ (respectively $a_0(t-)$) is the best bid (respectively ask) before the arrival of this order. The results are plotted in figures 16 (on a semi-log scale) and 17 (on a linear scale).

These graphs are computed with incomplete data (five best limits), therefore we do not observe a placement as broad as in Bouchaud *et al.* (2002). However, our data make it clear that fat tails are observed. We also observe an asymmetry in the empirical distribution: the left side is less broad than the right side. Since the left side represents limit orders submitted *inside* the spread, this is expected. Thus, the empirical distribution of the placement of arriving limit orders is maximum at zero

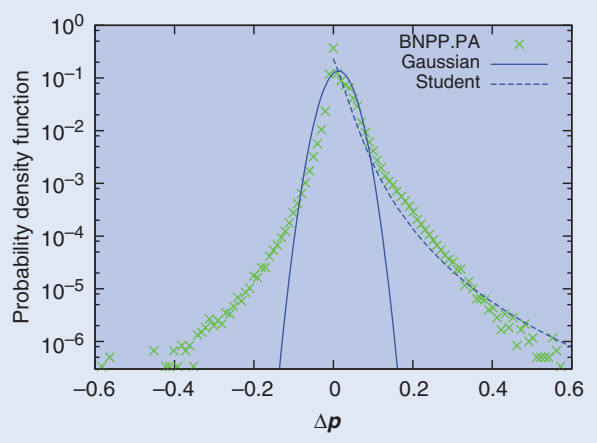

Figure 16. Placement of limit orders using the same best quote reference on a semi-log scale. The data used for this computation are from the BNP Paribas order book from 1 September 2007 to 31 May 2008.

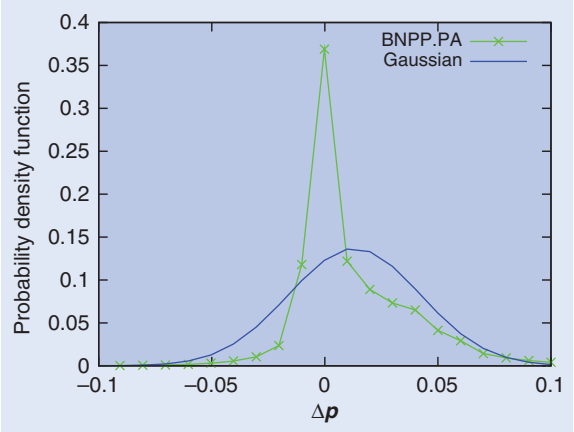

Figure 17. Placement of limit orders using the same best quote reference on a linear scale. The data used for this computation are from the BNP Paribas order book from 1 September 2007 to 31 May 2008.

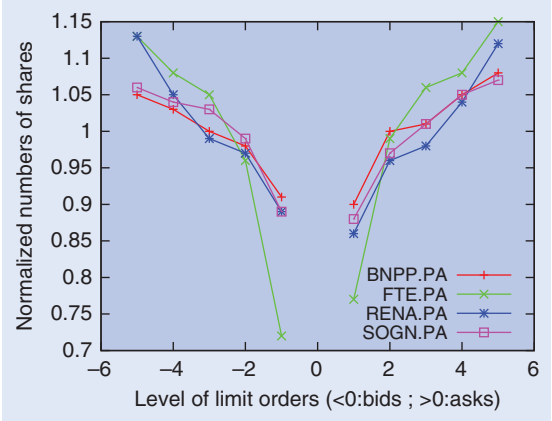

Figure 18. Average quantity offered in the limit order book.

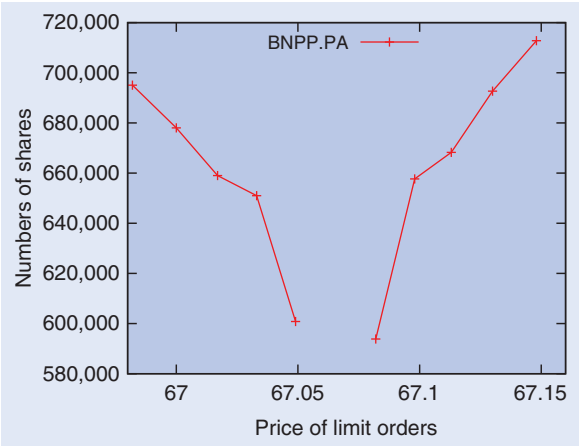

Figure 19. Average limit order book: Price and depth.

(same best quote). We then ask the question: How is this translated in terms of the shape of the order book?

3.3.2. Average shape of the order book. Contrary to what one might expect, it seems that the maximum of the average offered volume in an order book is located away from the best quotes (see, e.g., Bouchaud *et al.* (2002)). Our data confirm this observation: the average quantity offered on the five best quotes increases with the level. This result is presented in figure 18. We also compute the average price of these levels in order to plot a cross-sectional graph similar to those presented by Biais *et al.* (1995). Our result is presented for stock BNP.PA in figure 19 and displays the expected shape. The results for the other stocks are similar. We find that the average gap between two levels is constant among the five best bids and asks (less than one tick for FTE.PA, 1.5 ticks for BNPP.PA, 2.0 ticks for SOGN.PA, and 2.5 ticks for RENA.PA). We also find that the average spread is roughly twice as big as the average gap (a factor of 1.5 for FTE.PA, 2 for BNPP.PA, 2.2 for SOGN.PA, and 2.4 for RENA.PA).

3.4. *Cancelation of orders*

Challet and Stinchcombe (2001) show that the distribution of the average lifetime of limit orders fits a power law with exponent $1+\mu \approx 2.1$ for canceled limit orders and $1+\mu \approx 1.5$ for executed limit orders. Mike and Farmer (2008) find that, in either case, the exponential hypothesis (Poisson process) is not satisfied in the market.

We compute the average lifetime of canceled and executed orders on our data set. Since our data does not include a unique identifier of a given order, we reconstruct lifetime orders as follows: each time a cancelation is detected, we go back through the history of the limit order submissions and look for a matching order with the same price and same quantity. If an order is not matched, we discard the cancelation from our lifetime data. The results are presented in figures 20 and 21. We observe a power-law decay with coefficients $1+\mu \approx 1.3$–1.6 for both canceled and executed limit orders, with little variation among stocks. These results are slightly different from those presented in previous studies: similar for

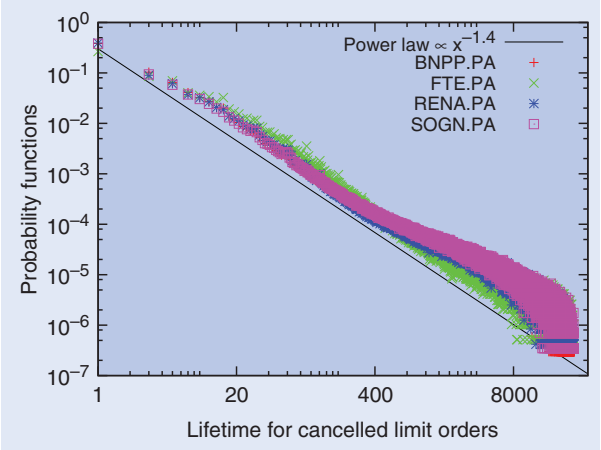

Figure 20. Distribution of the estimated lifetime of canceled limit orders.

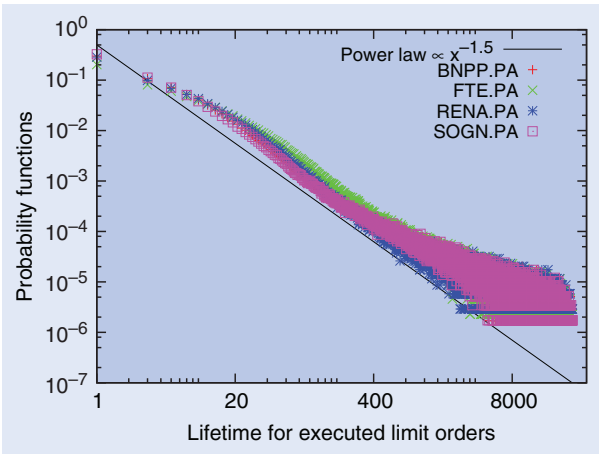

Figure 21. Distribution of the estimated lifetime of executed limit orders.

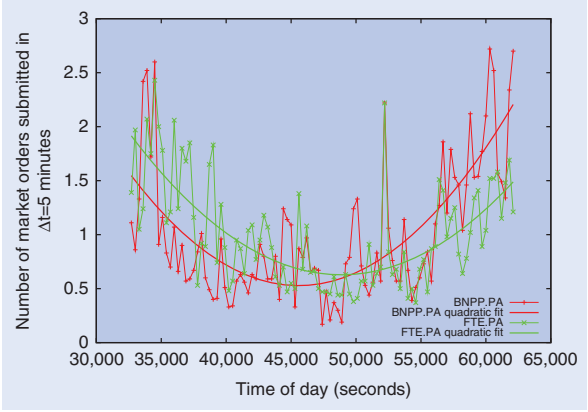

Figure 22. Normalized average number of market orders in a 5-min interval.

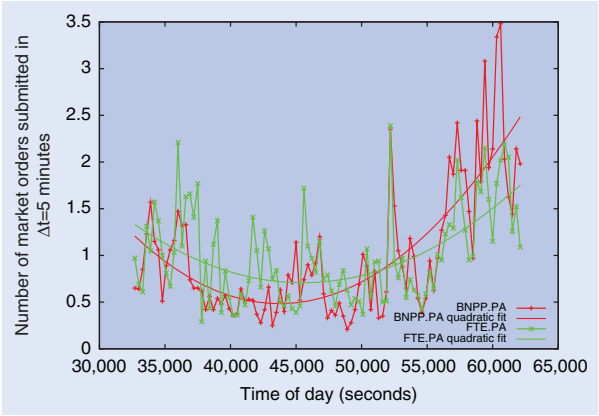

Figure 23. Normalized average number of limit orders in a 5-min interval.

executed limit orders, but our data exhibit a lower decay as for canceled orders. Note that the observed cut-off in the distribution for lifetimes greater than 20,000 seconds is due to the fact that we do not take into account the execution or cancelation of orders submitted on the previous day.

3.5. *Intra-day seasonality*

The activity on financial markets is of course not constant throughout the day. Figure 22 (respectively figure 23) plots the (normalized) number of market (respectively limit) orders arriving in a five-minute interval. It is clear that a U-shape is observed (an ordinary least-square quadratic fit is plotted): the observed market activity is larger at the beginning and at the end of the day, and quieter around mid-day. Such a U-shaped curve is well-known (see Biais *et al.* (1995), for example). For our data, we observe that the number of orders in a five-minute interval can vary by a factor of 10 throughout the day.

Challet and Stinchcombe (2001) note that the average number of orders submitted to the market in a period ΔT varies widely during the day. They also observe that these quantities for market orders and limit orders are highly correlated. Such a type of intra-day variation of global market activity is a well-known fact, already observed by Biais *et al.* (1995), for example.

3.6. *Market impact*

The statistics we have presented may help one to understand a phenomenon of primary importance for any financial market practitioner: the market impact, i.e. the relationship between the volume traded and the expected price shift once the order has been executed. To a first approximation, one can understand that it is closely linked to many of the items described above: the volume of market orders submitted, the shape of the order book (how many pending limit orders are hit by one large market order), the correlation of trade signs (one may assume that large orders are split in order to avoid a large market impact), etc.

Many empirical studies are available. An empirical study of the price impact of individual transactions on 1000 stocks on the NYSE was conducted by Lillo *et al.* (2003). It was found that proper re-scaling makes all the

curves collapse onto a single concave master curve. This function increases as a power that is of the order 1/2 for small volumes, but then increases more slowly for large volumes. They obtain similar results in each year for the period 1995 to 1998.

We will not review the large literature on market impact any further, but rather refer the reader to the recent exhaustive synthesis of Bouchaud *et al.* (2009), where different types of impacts, as well as some theoretical models, are discussed.

4. Correlations of assets

The word 'correlation' is defined as "a relation existing between phenomena or things or between mathematical or statistical variables which tend to vary, be associated, or occur together in a way not expected on the basis of chance alone".† When we talk about correlations in stock prices, what we are really interested in are relations between variables such as stock prices, order signs, transaction volumes, etc. and, more importantly, how these relations affect the nature of the statistical distributions and laws that govern the price time series. This section deals with several topics concerning the linear correlation observed in financial data. The first part deals with the important issue of computing correlations in high-frequency data. As mentioned earlier, the computerization of financial exchanges has lead to the availability of a huge amount of tick-by-tick data, and computing correlations using these intra-day data raises many issues concerning the usual estimators. The second and third parts deal with the use of correlation in order to cluster assets with potential applications in risk management problems.

4.1. Estimating covariance on high-frequency data

Let us assume that we observe d time series of prices or log-prices p_i, $i=1,\ldots,d$, at times t_m, $m=0,\ldots,M$. The usual estimator of the covariance of prices i and j is the *realized covariance estimator*, which is computed as

$$\hat{\Sigma}_{ij}^{RV}(t)=\sum_{m=1}^{M}(p_i(t_m)-p_i(t_{m-1}))(p_j(t_m)-p_j(t_{m-1})). \quad (6)$$

The problem is that high-frequency tick-by-tick data record changes in prices when they occur, i.e. at random times. Tick-by-tick data are thus asynchronous, contrary to daily close prices, for example, that are recorded at the same time for all the assets on a given exchange. Using standard estimators without caution could be one cause for the 'Epps effect', first observed by Epps (1979), who stated that "[c]orrelations among price changes in common stocks of companies in one industry are found to decrease with the length of the interval for which the price changes are measured". This has since largely been verified by, for example, Bonanno *et al.* (2001) and Reno (2003). Hayashi and Yoshida (2005) showed that the non-synchronicity of tick-by-tick data and the necessary sampling of time series in order to compute the usual realized covariance estimator partially explain this phenomenon. We very briefly review here two covariance estimators that do not need any synchronicity (hence, sampling) in order to be computed.

4.1.1. The Fourier estimator. The Fourier estimator was introduced by Malliavin and Mancino (2002). Let us assume that we have d time series of log-prices that are observations of Brownian semi-martingales p_i:

$$\mathrm{d}p_i=\sum_{j=1}^{K}\sigma_{ij}\mathrm{d}W_j+\mu_i\mathrm{d}t, \quad i=1,\ldots,d. \quad (7)$$

The coefficients of the covariance matrix are then written as $\Sigma_{ij}(t)=\sum_{k=1}^{K}\sigma_{ik}(t)\sigma_{jk}(t)$. Malliavin and Mancino (2002) show that the Fourier coefficients of $\Sigma_{ij}(t)$ are, with n_0 a given integer,

$$a_k(\Sigma_{ij})=\lim_{N\to\infty}\frac{\pi}{N+1-n_0}\sum_{s=n_0}^{N}\frac{1}{2}[a_s(\mathrm{d}p_i)a_{s+k}(\mathrm{d}p_j)+b_{s+k}(\mathrm{d}p_i)b_s(\mathrm{d}p_j)], \quad (8)$$

$$b_k(\Sigma_{ij})=\lim_{N\to\infty}\frac{\pi}{N+1-n_0}\sum_{s=n_0}^{N}\frac{1}{2}[a_s(\mathrm{d}p_i)b_{s+k}(\mathrm{d}p_j)-b_s(\mathrm{d}p_i)a_{s+k}(\mathrm{d}p_j)], \quad (9)$$

where the Fourier coefficients $a_k(\mathrm{d}p_i)$ and $b_k(\mathrm{d}p_i)$ of $\mathrm{d}p_i$ can be directly computed on the time series. Indeed, re-scaling the time window on $[0,2\pi]$ and using integration by parts, we have

$$a_k(\mathrm{d}p_i)=\frac{p(2\pi)-p(0)}{\pi}-\frac{k}{\pi}\int_0^{2\pi}\sin(kt)\,p_i(t)\mathrm{d}t. \quad (10)$$

This latter integral can be discretized and computed approximately using the times t_m^i of observations of the process p_i. Therefore, fixing a sufficiently large N, one can compute an estimator Σ_{ij}^{F} of the covariance of the processes i and j (see Reno (2003) and Iori and Precup (2007) for examples of empirical studies using this estimator).

4.1.2. The Hayashi–Yoshida estimator. Hayashi and Yoshida (2005) proposed a simple estimator in order to compute covariance/correlation without any need for synchronicity of time series. As for the Fourier estimator, it is assumed that the observed process is a Brownian semi-martingale. The time window of observation is easily partitioned into d families of intervals $\Pi^i=(U_m^i)$, $i=1,\ldots,d$, where $t_m^i=\inf\{U_{m+1}^i\}$ is the time of the mth observation of process i. Let $\Delta p_i(U_m^i)=p_i(t_m^i)-p_i(t_{m-1}^i)$. The *cumulative covariance estimator*, as the authors called

†*Merriam–Webster Online Dictionary*. Retrieved 14 June 2010 from http://www.merriam-webster.com/dictionary/correlations

it, or the *Hayashi–Yoshida estimator*, as it is largely referred to, is then built as follows:

$$\hat{\Sigma}_{ij}^{\mathrm{HY}}(t) = \sum_{m,n} \Delta p_i(U_m^i)\Delta p_j(U_n^j)1_{\{U_m^i \cap U_n^j \neq \emptyset\}}. \qquad (11)$$

There is an extensive literature in Econometrics that tackles the new challenges posed by high-frequency data. For readers wishing to go beyond this brief presentation, we refer to the econometrics reviews of Barndorff-Nielsen and Shephard (2007) and McAleer and Medeiros (2008), for example.

4.2. Correlation matrix and random matrix theory

With stock market data being essentially *multivariate* time series data, we construct a correlation matrix to study the spectra and contrast them with the random multivariate data from a coupled map lattice. It is known from previous studies that the empirical spectra of correlation matrices drawn from time series data for most part follow random matrix theory (RMT; see, e.g., Gopikrishnan *et al.* (2001)).

4.2.1. Correlation matrix and eigenvalue density. *Correlation matrix*. If there are N assets with price $P_i(t)$ for asset i at time t, then the logarithmic return of stock i is $r_i(t) = \ln P_i(t) - \ln P_i(t-1)$, which for a certain consecutive sequence of trading days forms the return vector $r\}_i$. In order to characterize the synchronous time evolution of stocks, the equal time correlation coefficients between stocks i and j is defined as

$$\rho_{ij} = \frac{\langle r_i r_j\rangle - \langle r_i\rangle\langle r_j\rangle}{\sqrt{[\langle r_i^2\rangle - \langle r_i\rangle^2][\langle r_j^2\rangle - \langle r_j\rangle^2]}}, \qquad (12)$$

where $\langle\ldots\rangle$ indicates a time average over the trading days included in the return vectors. These correlation coefficients form an $N \times N$ matrix with $-1 \le \rho_{ij} \le 1$. If $\rho_{ij} = 1$, the stock price changes are completely correlated; if $\rho_{ij} = 0$, the stock price changes are uncorrelated; and if $\rho_{ij} = -1$, then the stock price changes are completely anti-correlated.

Correlation matrix of the spatio-temporal series from coupled map lattices. Consider a time series of the form $z'(x,t)$, where $x = 1, 2, \ldots, n$ and $t = 1, 2, \ldots, p$ denote the discrete space and time, respectively. Here, the time series at every spatial point is treated as a different variable. We define the normalized variable as

$$z(x,t) = \frac{z'(x,t) - \langle z'(x)\rangle}{\sigma(x)}, \qquad (13)$$

where $\langle\ldots\rangle$ represent temporal averages and $\sigma(x)$ the standard deviation of z' at position x. Then, the equal-time cross-correlation matrix that represents the spatial correlations can be written as

$$S_{x,x'} = \langle z(x,t)z(x',t)\rangle, \quad x, x' = 1, 2, \ldots, n. \qquad (14)$$

The correlation matrix is symmetric by construction. In addition, a large class of processes are translation invariant and the correlation matrix can also contain that additional symmetry. We will use this property for our correlation models in the context of the coupled map lattice. In time series analysis, the averages $\langle\ldots\rangle$ have to be replaced by estimates obtained from finite samples. As usual, we will use the maximum likelihood estimates, $\langle a(t)\rangle \approx (1/p)\sum_{t=1}^{p} a(t)$. These estimates contain statistical uncertainties that disappear for $p \to \infty$. Ideally, one requires $p \gg n$ in order to obtain reasonably correct correlation estimates (see Chakraborti *et al.* (2007) for details of the parameters).

Eigenvalue density. The interpretation of the spectra of empirical correlation matrices should be done carefully if one wants to be able to distinguish between system-specific signatures and universal features. The former express themselves in the smoothed level density, whereas the latter are usually represented by the fluctuations on top of this smooth curve. In time series analysis, the matrix elements are not only prone to uncertainty such as measurement noise on the time series data, but also statistical fluctuations due to finite-sample effects. When characterizing time series data in terms of random matrix theory, one is not interested in these trivial sources of fluctuations that are present in every data set, but one would like to identify the significant features that would be shared, in principle, by an 'infinite' amount of data without measurement noise. The eigenfunctions of the correlation matrices constructed from such empirical time series carry the information contained in the original time series data in a 'graded' manner and they also provide a compact representation for it. Thus, by applying an approach based on random matrix theory, one tries to identify non-random components of the correlation matrix spectra as deviations from random matrix theory predictions (Gopikrishnan *et al.* 2001).

We will look at the eigenvalue density that has been studied in the context of applying random matrix theory methods to time series correlations. Let $\mathcal{N}(\lambda)$ be the integrated eigenvalue density that gives the number of eigenvalues less than a given value λ. Then, the eigenvalue or level density is given by $\rho(\lambda) = \mathrm{d}\mathcal{N}(\lambda)/\mathrm{d}\lambda$. This can be obtained by assuming a random correlation matrix and is found to be in good agreement with the empirical time series data from stock market fluctuations. From random matrix theory considerations, the eigenvalue density for random correlations is given by

$$\rho_{\mathrm{rmt}}(\lambda) = \frac{Q}{2\pi\lambda}\sqrt{(\lambda_{\max} - \lambda)(\lambda - \lambda_{\min})}, \qquad (15)$$

where $Q = N/T$ is the ratio of the number of variables to the length of each time series. Here, $\lambda_{\max}$ and $\lambda_{\min}$, representing the maximum and minimum eigenvalues of the random correlation matrix, respectively, are given by $\lambda_{\max,\min} = 1 + 1/Q \pm 2\sqrt{1/Q}$. However, due to the presence of correlations in the empirical correlation matrix, this eigenvalue density is often violated for a certain number of dominant eigenvalues. They often correspond to system-specific information in the data. Figure 24 shows the eigenvalue density for S&P 500 data and also for the chaotic data from the coupled map lattice.

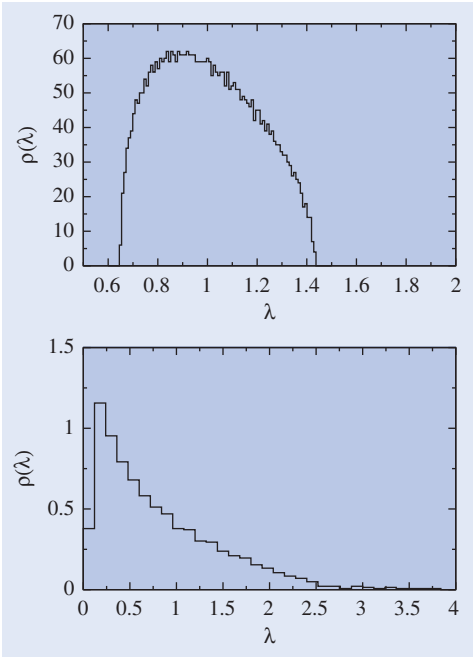

Figure 24. (Upper panel) The spectral density for multivariate spatio-temporal time series drawn from coupled map lattices. (Lower panel) The eigenvalue density for the return time series of the S&P 500 stock market data (8938 time steps). Reproduced from Chakraborti *et al.* (2007).

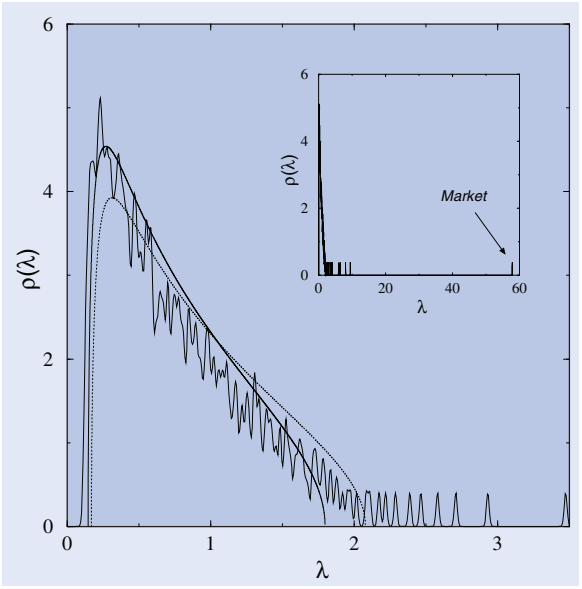

Figure 25. Eigenvalue spectrum of the correlation matrices. Adapted from Laloux *et al.* (1999).

Clearly, both curves are qualitatively different. Thus, the presence or absence of correlations in the data is manifest in the spectrum of the corresponding correlation matrices.

4.2.2. Earlier estimates and studies using random matrix theory. Laloux *et al.* (1999) showed that results from random matrix theory were useful for understanding the statistical structure of the empirical correlation matrices appearing in the study of price fluctuations. The empirical determination of a correlation matrix is a difficult task. If one considers N assets, the correlation matrix contains $N(N-1)/2$ mathematically independent elements that must be determined from N time series of length T. If T is not very large compared with N, then, generally, the determination of the covariances is noisy, and therefore the empirical correlation matrix is, to a large extent, random. The smallest eigenvalues of the matrix are the most sensitive to this 'noise'. But the eigenvectors corresponding to these smallest eigenvalues determine the minimum risk portfolios in Markowitz theory. It is thus important to distinguish 'signal' from 'noise' or, in other words, to extract the eigenvectors and eigenvalues of the correlation matrix containing real information (those important for risk control) from those that do not contain any useful information and are unstable in time. It is useful to compare the properties of an empirical correlation matrix with a 'null hypothesis'—a random matrix that arises, for example, from a finite time series of strictly uncorrelated assets. Deviations from the random matrix case might then suggest the presence of true information. The main result of their study was the remarkable agreement between the theoretical prediction (based on the assumption that the correlation matrix is random) and empirical data concerning the density of the eigenvalues (shown in figure 25) associated with the time series of the different stocks of the S&P 500 (or other stock markets).

Cross-correlations in financial data were also studied by Plerou *et al.* (1999, 2002). They analysed cross-correlations between price fluctuations of different stocks using the methods of RMT. Using two large databases, they calculated cross-correlation matrices of returns constructed from (i) 30-min returns of 1000 US stocks for the two-year period 1994–95, (ii) 30-min returns of 881 US stocks for the two-year period 1996–97, and (iii) one-day returns of 422 US stocks for the 35-year period 1962–96. They also tested the statistics of the eigenvalues λ_i of cross-correlation matrices against a 'null hypothesis'. They found that a majority of the eigenvalues of the cross-correlation matrices were within the RMT bounds $[\lambda_{\min}, \lambda_{\max}]$, as defined above, for the eigenvalues of random correlation matrices. They also tested the eigenvalues of the cross-correlation matrices within the RMT bounds for universal properties of random matrices and found good agreement with the results for the Gaussian orthogonal ensemble (GOE) of random matrices—implying a large degree of randomness in the measured cross-correlation coefficients. Furthermore, they found that the distribution of eigenvector components for the eigenvectors corresponding to the eigenvalues outside the RMT bounds displayed systematic deviations from the RMT prediction and that these 'deviating eigenvectors' were stable in time. They analysed the components of the deviating eigenvectors and found that the largest eigenvalue corresponded to an influence common to all stocks. Their analysis of the remaining deviating eigenvectors showed distinct groups, the identities of which corresponded to conventionally identified business sectors.

4.3. Analyses of correlations and economic taxonomy

4.3.1. Models and theoretical studies of financial correlations. Podobnik *et al*. (2000) studied how the presence of correlations in physical variables contributes to the form of the probability distributions. They investigated a process with correlations in the variance generated by a Gaussian or a truncated Levy distribution. For both Gaussian and truncated Levy distributions, they found that, due to the correlations in the variance, the process 'dynamically' generated power-law tails in the distributions, the exponents of which could be controlled through the way the correlations in the variance were introduced. For a truncated Levy distribution, the process could extend a truncated distribution beyond the *truncation cut-off*, leading to a crossover between a Levy stable power law and their 'dynamically generated' power law. It was also shown that the process could explain the crossover behavior observed in the S&P 500 stock index.

Noh (2000) proposed a model for correlations in stock markets in which the markets were composed of several groups, within which the stock price fluctuations were correlated. The spectral properties of empirical correlation matrices (Laloux *et al*. 1999, Plerou *et al*. 1999) were studied in relation to this model and the connection between the spectral properties of the empirical correlation matrix and the structure of correlations in stock markets was established.

The correlation structure of extreme stock returns were studied by Cizeau *et al*. (2001). It was commonly believed that the correlations between stock returns increased in high-volatility periods. They investigated how much of these correlations could be explained within a simple non-Gaussian one-factor description with time-independent correlations. Using surrogate data with the true market return as the dominant factor, it was shown that most of these correlations, measured using a variety of different indicators, could be accounted for. In particular, their one-factor model could explain the level and asymmetry of empirical exceeding correlations. However, more subtle effects required an extension of the one-factor model, where the variance and skewness of the residuals also depended on the market return.

Burda *et al*. (2001) provided a statistical analysis of three S&P 500 covariances with evidence for raw tail distributions. They studied the stability of these tails with respect to reshuffling for the S&P 500 data and showed that the covariance with the strongest tails was robust, with a spectral density in remarkable agreement with random Levy matrix theory. They also studied the inverse participation ratio for the three covariances. The strong localization observed at both ends of the spectral density was analogous to the localization exhibited in the random Levy matrix ensemble. They showed that the stocks with the largest scattering were the least susceptible to correlations and were the likely candidates for the localized states.

4.3.2. Analyses using graph theory and economic taxonomy. Mantegna (1999) introduced a method for finding a hierarchical arrangement of stocks traded in financial markets by studying the clustering of companies using correlations of asset returns. With an appropriate metric—based on the earlier explained correlation matrix coefficients ρ_{ij} between all pairs of stocks i and j of the portfolio, computed using equation (12) by considering the synchronous time evolution of the difference of the logarithm of the daily stock price—a fully connected graph was defined in which the nodes are companies, or stocks, and the 'distances' between them are obtained from the corresponding correlation coefficients. The minimum spanning tree (MST) was generated from the graph by selecting the most important correlations and was used to identify clusters of companies. The hierarchical tree of the sub-dominant ultrametric space associated with the graph provided information useful for investigating the number and nature of the common economic factors affecting the time evolution of the logarithm of price of well-defined groups of stocks. Several other attempts have been made to obtain clustering from a huge correlation matrix.

Bonanno *et al*. (2001) studied the high-frequency cross-correlation existing between pairs of stocks traded in a financial market in a set of 100 stocks traded in US equity markets. A hierarchical organization of the investigated stocks was obtained by determining the metric distance between stocks and by investigating the properties of the sub-dominant ultrametric associated with it. A clear modification of the hierarchical organization of the set of stocks investigated was detected when the time horizon used to determine stock returns was changed. The hierarchical location of stocks of the energy sector was investigated as a function of the time horizon. The hierarchical structure explored by the minimum spanning tree also seemed to give information about the influential power of the companies.

It also turned out that the hierarchical structure of the financial market could be identified in accordance with the results obtained by an independent clustering method, based on Potts super-paramagnetic transitions as studied by Kullmann *et al*. (2000), where the spins correspond to companies and the interactions are functions of the correlation coefficients determined from the time dependence of the companies' individual stock prices. The method is a generalization of the clustering algorithm of Blatt *et al*. (1996) to the case of anti-ferromagnetic interactions corresponding to anti-correlations. For the Dow Jones Industrial Average, no anti-correlations were observed in the investigated time period and the previous results obtained by different tools were well reproduced. For the S&P 500, where anti-correlations occur, repulsion between stocks modified the cluster structure of the $N = 443$ companies studied, as shown in figure 26. The efficiency of the method is represented by the fact that the figure matches well with the corresponding result obtained by the minimal spanning tree method, including the specific composition of the clusters. For example, at the lowest level of the hierarchy (highest temperature

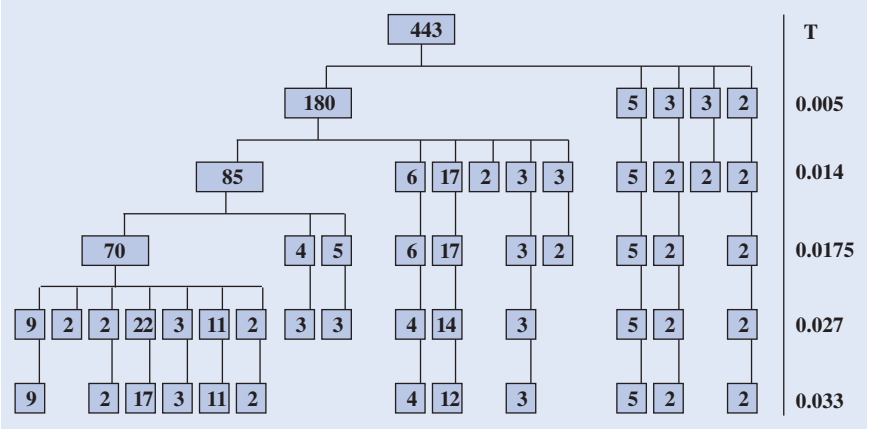

Figure 26. The hierarchical structure of clusters of S&P 500 companies in the ferromagnetic case. The number of elements of the cluster are indicated in boxes. The clusters consisting of single companies are not indicated. Adapted from Kullmann *et al.* (2000).

in the super-paramagnetic phase) the different industrial branches can clearly be identified: oil, electricity, gold mining companies, etc. build separate clusters.

The network of influence was investigated by means of the time-dependent correlation method of Kullmann *et al.* (2000). They studied the correlations as a function of the time shift between pairs of stock return time series of tick-by-tick data of the NYSE. They investigated whether or not any 'pulling effect' between stocks existed, i.e. whether or not, at any given time, the return value of one stock influenced that of another stock at a different time. They found that, in general, two types of mechanisms generated significant correlation between any two given stocks. One was some kind of external effect (say, economic or political news) that influenced both stock prices simultaneously, and the change for both prices appeared at the same time, such that the maximum of the correlation was at zero time shift. The second effect was that one of the companies had an influence on the other company, indicating that one company's operation depended on the other, so that the price change of the influenced stock appeared later because it required some time to react to the price change of the first stock, displaying a 'pulling effect'. A weak but significant effect with the real data set was found, showing that, in many cases, the maximum correlation was at non-zero time shift, indicating the direction of influence between the companies, and the characteristic time was of the order of a few minutes, which was compatible with the efficient market hypothesis. In the pulling effect, they found that, in general, more important companies (which were traded more) pulled the relatively smaller companies.

The time-dependent properties of the minimum spanning tree (introduced by Mantegna), called a 'dynamic asset tree', were studied by Onnela *et al.* (2003b). The nodes of the tree were identified with stocks and the distance between them was a unique function of the corresponding element of the correlation matrix. Using the concept of a central vertex, chosen as the most strongly connected node of the tree, the mean occupation layer was defined, which was an important characteristic of the tree. During crashes, the strong global correlation in the market manifested itself by a low value of the mean occupation layer. The tree seemed to have a scale-free structure, where the scaling exponent of the degree distribution was different for 'business as usual' and 'crash' periods. The basic structure of the tree topology was very robust with respect to time.

Financial correlation matrix and constructing asset trees. Two different sets of financial data were used. The first set was from the Standard & Poor's 500 index (S&P 500) of the New York Stock Exchange (NYSE) from 2 July 1962 to 31 December 1997 and contained 8939 daily closing values. The second set recorded the split-adjusted daily closure prices for a total of $N = 477$ stocks traded on the New York Stock Exchange (NYSE) over the period of 20 years from 2 January 1980 to 31 December 1999. This amounted to a total of 5056 prices per stock, indexed by time variable $\tau = 1, 2, \ldots, 5056$. For analysis and smoothing purposes, the data were divided time-wise into M *windows* $t = 1, 2, \ldots, M$ of width T, where T corresponds to the number of daily returns included in the window. Note that several consecutive windows overlap each other, the extent of which is dictated by the window step length parameter δT, which describes the displacement of the window and is also measured in trading days. The choice of window width is a trade-off between too noisy and too smoothed data for small and large window widths, respectively. The results presented here were calculated from monthly stepped four-year windows, i.e. $\delta T = 250/12 \approx 21$ days and $T = 1000$ days. A large scale of different values for both parameters were explored, and the cited values were found optimal (Onnela 2000). With these choices, the overall number of windows is $M = 195$.

The above definition of a correlation matrix given by equation (12) is used. These correlation coefficients form an $N \times N$ correlation matrix $\mathbf{C}'$, which serves as the basis for the trees discussed below. An asset tree is then constructed according to the methodology of Mantegna (1999). For the purpose of constructing asset trees, a distance is defined between a pair of stocks. This distance is associated with the edge connecting the stocks and it is expected to reflect the level at which the stocks

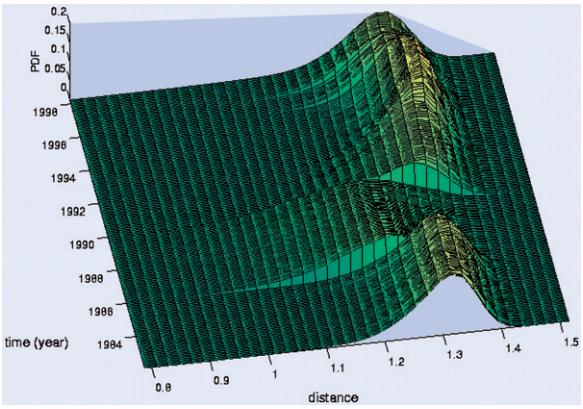

Figure 27. Distribution of all $N(N-1)/2$ distance elements d_{ij} contained in the distance matrix $\mathbf{D}^t$ as a function of time. Reproduced from Onnela *et al.* (2003c).

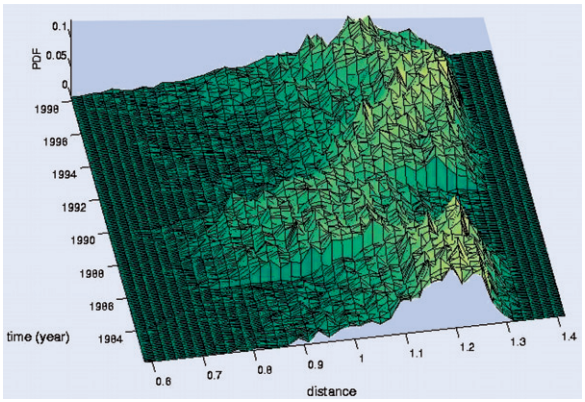

Figure 28. Distribution of the $(N-1)$ distance elements d_{ij} contained in the asset (minimum spanning) tree $\mathbf{T}^t$ as a function of time. Reproduced from Onnela *et al.* (2003c).

are correlated. A simple nonlinear transformation $d_{ij}^t = [2(1-\rho_{ij}^t)]^{1/2}$ is used to obtain distances with the property $2 \geq d_{ij} \geq 0$, forming an $N \times N$ symmetric distance matrix $\mathbf{D}^t$. Therefore, if $d_{ij}=0$, the stock price changes are completely correlated, and if $d_{ij}=2$, the stock price changes are completely anti-uncorrelated. The trees for different time windows are not independent of each other, but form a series through time. Consequently, this multitude of trees is interpreted as a sequence of evolutionary steps of a single *dynamic asset tree*. An additional hypothesis is required concerning the topology of the metric space: the ultrametricity hypothesis. In practice, this leads to determining the minimum spanning tree (MST) of the distances, denoted $\mathbf{T}^t$. The spanning tree is a simply connected acyclic (no cycles) graph that connects all N nodes (stocks) with $N-1$ edges such that the sum of all edge weights, $\sum_{d_{ij}^t \in \mathbf{T}^t} d_{ij}^t$, is minimum. We refer to the minimum spanning tree at time t by the notation $\mathbf{T}^t = (V, E^t)$, where V is a set of vertices and E^t is the corresponding set of unordered pairs of vertices, or edges. Since the spanning tree criterion requires all N nodes always to be present, the set of vertices V is time independent, which is why the time superscript has been dropped from the notation. The set of edges E^t, however, does depend on time, as it is expected that edge lengths in the matrix $\mathbf{D}^t$ evolve over time, and thus different edges are selected in the tree at different times.

Market characterization. We plot the distribution of (i) distance elements d_{ij}^t contained in the distance matrix $\mathbf{D}^t$ (figure 27), and (ii) distance elements d_{ij} contained in the asset (minimum spanning) tree $\mathbf{T}^t$ (figure 28). In both plots, but most prominently in figure 27, there appears to be a discontinuity in the distribution between roughly 1986 and 1990. The part that has been cut out, pushed to the left and made flatter is a manifestation of Black Monday (19 October 1987), and its length along the time axis is related to the choice of window width T (Onnela *et al.* 2003a, b).

Also, note that in the distribution of tree edges in figure 28, most edges included in the tree seem to come from the area to the right of the value 1.1 in figure 27, and the largest distance element is $d_{\max} = 1.3549$.

Tree occupation and central vertex. Let us focus on characterizing the spread of nodes on the tree by introducing the quantity the *mean occupation layer*,

$$l(t, v_c) = \frac{1}{N}\sum_{i=1}^{N} \mathrm{lev}(v_i^t), \tag{16}$$

where $\mathrm{lev}(v_i)$ denotes the level of vertex v_i. The levels, not to be confused with the distances d_{ij} between nodes, are measured in natural numbers in relation to the *central vertex* v_c, the level of which is taken to be zero. Here the mean occupation layer indicates the layer on which the mass of the tree, on average, is conceived to be located. The central vertex is considered to be the parent of all other nodes in the tree, and is also known as the root of the tree. It is used as the *reference* point in the tree, against which the locations of all other nodes are relative. Thus all other nodes in the tree are children of the central vertex. Although there is an *arbitrariness* in the choice of the central vertex, it is proposed that the vertex is central, in the sense that any change in its price strongly affects the course of events in the market as a whole. Three alternative definitions for the central vertex have been proposed in studies, all yielding similar and, in most cases, identical outcomes. The idea is to find the node that is most strongly connected to its nearest neighbors. For example, according to one definition, the central node is the one with the highest *vertex degree*, i.e. the number of edges that are incident with (neighbors of) the vertex. Also, one may have either (i) a static (fixed at all times) or (ii) a dynamic (updated at each time step) central vertex, but again the results do not seem to vary significantly. Studies of the variation of the topological properties and nature of the trees with time have been performed.

Economic taxonomy. Mantegna's idea of linking stocks in an ultrametric space was motivated *a posteriori* by the property of such a space to provide a meaningful economic taxonomy (Onnela *et al.* 2002). Mantegna examined the meaningfulness of the taxonomy by comparing the grouping of stocks in the tree with a third-party reference grouping of stocks, e.g. by their industry

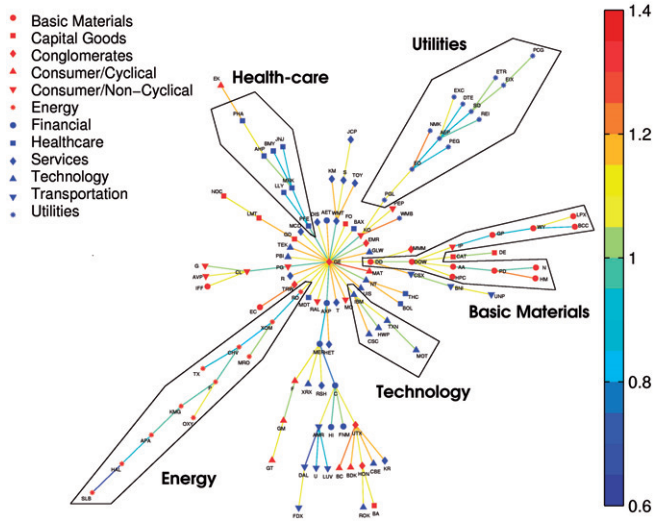

Figure 29. Snapshot of the dynamic asset tree connecting the examined 116 stocks of the S&P 500 index. The tree was produced using a four-year window width and is centered on 1 January 1998. Business sectors are indicated according to Forbes (www.forbes.com). In this tree, General Electric (GE) was used as the central vertex and eight layers can be identified. Reproduced from Onnela *et al.* (2003b).

classifications (Mantegna 1999). In this case, the reference was provided by Forbes (www.forbes.com), which uses its own classification system, assigning each stock a sector (higher level) and industry (lower level) category. In order to visualize the grouping of stocks, a sample asset tree was constructed for a smaller dataset (shown in figure 29), which consists of 116 S&P 500 stocks, extending from the beginning of 1982 to the end of 2000, resulting in a total of 4787 price quotes per stock (Onnela *et al.* 2003b). The window width was set at $T = 1000$, and the shown sample tree is located time-wise at $t = t^*$, corresponding to 1.1.1998. The stocks in this dataset fall into 12 *sectors*, which are Basic Materials, Capital Goods, Conglomerates, Consumer/Cyclical, Consumer/Non-Cyclical, Energy, Financial, Healthcare, Services, Technology, Transportation and Utilities. The sectors are indicated in the tree (see figure 29) by different markers, while the industry classifications are omitted for reasons of clarity. The term sector is used exclusively to refer to the given third-party classification system of stocks. The term *branch* refers to a subset of the tree, to all the nodes that share the specified common parent. In addition to the parent, it is required to have a reference point to indicate the generational direction (i.e. who is who's parent) in order for a branch to be well-defined. Without this reference there is absolutely no way to determine where one branch ends and another begins. In this case, the reference is the central node. There are some branches in the tree in which most of the stocks belong to just one sector, indicating that the branch is fairly homogeneous with respect to business sectors. This finding is in accordance with that of Mantegna (1999), although there are branches that are fairly heterogeneous, such as that extending directly downwards from the central vertex (see figure 29).

5. Partial conclusion

This first part of our review has reported the statistical properties of financial data (time series of prices, order book structure, asset correlations). Some of these properties, such as fat tails of returns and volatility clustering, are widely known and acknowledged as 'financial stylized facts'. They are now widely cited in order to compare financial models, and reveal the insufficiency of many classical stochastic models of financial assets. Some other properties are newer findings that were obtained by studying high-frequency data of the whole order book structure. The volume of orders, interval time between orders, intra-day seasonality, etc. are essential phenomena to be understood when working in financial modeling. The important role of studies of correlations has been emphasized. Besides the technical challenges raised by high-frequency data, many studies based, for example, on random matrix theory or clustering algorithms can help to obtain a better grasp of certain economic problems. It is our belief that future modeling in finance will have to be partly based on Econophysics studies of agent-based models in order to incorporate these 'stylized facts' in a comprehensive way. Agent-based reasoning for order book models, wealth exchange models and game theoretic models will be reviewed in the following part of the review (see companion paper).

Acknowledgements

The authors would like to thank their collaborators and two anonymous reviewers whose comments greatly helped improve the review. A.C. is grateful to B.K. Chakrabarti, K. Kaski, J. Kertesz, T. Lux, M. Marsili, D. Stauffer and V.M. Yakovenko for invaluable suggestions and criticisms.

References

Arthur, W., Complexity and the economy. *Science*, 1999, **284**, 107–109.

Bachelier, L., Theorie de la speculation. *Annales Scientifiques de l'Ecole Normale Superieure*, 1900, **III-17**, 21–86.

Barndorff-Nielsen, O.E. and Shephard, N., Variation, jumps and high frequency data in financial econometrics. In *Advances in Economics and Econometrics: Theory and Applications, Ninth World Congress*, Econometric Society Monographs, edited by R. Blundell, T. Persson and W.K. Newey, pp. 328–372, 2007 (Cambridge University Press: Cambridge).

Biais, B., Foucault, T. and Hillion, P., *Microstructure des Marches Financiers: Institutions, Modeles et Tests Empiriques*, 1997 (Presses Universitaires de France: Paris).

Biais, B., Hillion, P. and Spatt, C., An empirical analysis of the limit order book and the order flow in the Paris Bourse. *J. Finance*, 1995, 1655–1689.

Black, F. and Scholes, M., The pricing of options and corporate liabilities. *J. Polit. Econ.*, 1973, **81**, 637–654.

Blatt, M., Wiseman, S. and Domany, E., Superparamagnetic clustering of data. *Phys. Rev. Lett.*, 1996, **76**, 3251–3254.

Bollerslev, T., Engle, R.F. and Nelson, D.B., Arch models. In *Handbook of Econometrics 4*, edited by R.F. Engle and D.L. McFadden, pp. 2959–3038, 1994 (Elsevier: Amsterdam).

Bonanno, G., Lillo, F. and Mantegna, R.N., High-frequency cross-correlation in a set of stocks. *Quant. Finance*, 2001, **1**, 96–104.

Boness, A.J., English translation of Théorie de la spéculation. In *The Random Character of Stock Market Prices*, edited by P.H. Cootner, pp. 17–75, 1964 (MIT Press: Cambridge, MA).

Bouchaud, J.-P., Economics needs a scientific revolution. *Nature*, 2008, **455**, 1181.

Bouchaud, J.-P., Farmer, J.D. and Lillo, F., How markets slowly digest changes in supply and demand. In *Handbook of Financial Markets: Dynamics and Evolution*, edited by T. Hens and K.R. Schenk-Hopp, pp. 57–160, 2009 (North-Holland: San Diego).

Bouchaud, J.-P. and Potters, M., *Theory of Financial Risks: From Statistical Physics to Risk Management*, 2000 (Cambridge University Press: Cambridge).

Bouchaud, J.-P., Mézard, M. and Potters, M., Statistical properties of stock order books: Empirical results and models. *Quant. Finance*, 2002, **2**, 251–256.

Brock, W.A. and Hommes, C.H., Heterogeneous beliefs and routes to chaos in a simple asset pricing model. *J. Econ. Dynam. Control*, 1998, **22**, 1235–1274.

Burda, Z., Jurkiewicz, J., Nowak, M.A., Papp, G. and Zahed, I., Levy matrices and financial covariances. cond-mat/0103108, 2001.

Chakrabarti, A.S. and Chakrabarti, B.K., Statistical theories of income and wealth distribution. *Economics E-Journal* (open access), 2010, **4**, 1–32.

Chakrabarti, B.K., Chakraborti, A. and Chatterjee, A., eds, *Econophysics and Sociophysics: Trends and Perspectives*, 1st ed., 2006 (Wiley–VCH: Berlin).

Chakraborti, A., Patriarca, M. and Santhanam, M.S., Financial time-series analysis: A brief overview. In *Econophysics of Markets and Business Networks*, 2007 (Springer: Milan).

Challet, D. and Stinchcombe, R., Analyzing and modeling 1 + 1d markets. *Physica A*, 2001, **300**, 285–299.

Chiarella, C., He, X. and Hommes, C., A dynamic analysis of moving average rules. *J. Econ. Dynam. Control*, 2006, **30**, 1729–1753.

Cizeau, P., Liu, Y., Meyer, M., Peng, C.K. and Stanley, H.E., Volatility distribution in the S&P500 stock index. *Physica A*, 1997, **245**, 441–445.

Cizeau, P., Potters, M. and Bouchaud, J., Correlation structure of extreme stock returns. *Quant. Finance*, 2001, **1**, 217–222.

Clark, P.K., A subordinated stochastic process model with finite variance for speculative prices. *Econometrica*, 1973, **41**, 135–155.

Cont, R., Empirical properties of asset returns: Stylized facts and statistical issues. *Quant. Finance*, 2001, **1**, 223–236.

Cont, R., Potters, M. and Bouchaud, J.-P., Scale invariance and beyond. In *Proceedings of the CNRS Workshop on Scale Invariance*, edited by F.G.B. Dubrulle and D. Sornette, 1997 (Springer: Berlin). Available online at: http://ssrn.com/abstract = 39420 or doi:10.2139/ssrn.39420.

Cont, R. and Tankov, P., *Financial Modelling with Jump Processes*, 2004 (Chapman & Hall/CRC: London).

Courtault, J., Kabanov, Y., Bru, B., Crepel, P., Lebon, I. and Le Marchand, A., Louis Bachelier on the centenary of Theorie de la Speculation. *Math. Finance*, 2000, **10**, 339–353.

de Haan, L., Resnick, S. and Drees, H., How to make a Hill plot. *Ann. Statist.*, 2000, **28**, 254–274.

de Oliveira, S.M., de Oliveira, P.M.C. and Stauffer, D., *Evolution, Money, War and Computers*, 1999 (B. G. Teubner: Stuttgart).

di Ettore Majorana, N., Il valore delle leggi statistiche nella fisica e nelle scienze sociali. *Scientia*, 1942, **36**, 58–66.

Dremin, I. and Leonidov, A., On distribution of number of trades in different time windows in the stock market. *Physica A*, 2005, **353**, 388–402.

Duffie, D., *Dynamic Asset Pricing Theory*, 1996 (Princeton University Press: Princeton, NJ).

Engle, R.F., The econometrics of ultra-high-frequency data. *Econometrica*, 2000, **68**, 1–22.

Engle, R.F. and Russell, J.R., Forecasting the frequency of changes in quoted foreign exchange prices with the autoregressive conditional duration model. *J. Empir. Finance*, 1997, **4**, 187–212.

Epps, T.W., Comovements in stock prices in the very short run. *J. Am. Statist. Assoc.*, 1979, **74**, 291–298.

Farmer, J.D. and Foley, D., The economy needs agent-based modelling. *Nature*, 2009, **460**, 685–686.

Feller, W., *Introduction to the Theory of Probability and its Applications*, Vol. 2, 1968 (Wiley: New York).

Follmer, H. and Schied, A., *Stochastic Finance: An Introduction In Discrete Time 2*, 2nd revised ed., 2004 (Walter de Gruyter).

Forfar, D., Louis Bachelier. In *Louis Bachelier The MacTutor History of Mathematics Archive* (published online), edited by J. O'Connor and E.F. Robertson, 2002.

Gabaix, X., Power laws in economics and finance. *A. Rev. Econ.*, 2009, **1**, 255–294.

Gabaix, X., Gopikrishnan, P., Plerou, V. and Stanley, H.E., Institutional investors and stock market volatility. *Q. J. Econ.*, 2006, **121**, 461–504.

Gallegati, M. and Kirman, A.P., editors, *Beyond the Representative Agent*, 1st ed., 1999 (Edward Elgar Publishing: Cheltenham).

Garman, M.B., Market microstructure. *J. Financial Econ.*, 1976, **3**, 257–275.

Gatheral, J., *The Volatility Surface: A Practitioner's Guide*, 2006 (Wiley: New York).

Glosten, L.R., Is the electronic open limit order book inevitable? *J. Finance*, 1994, **49**, 1127–1161.

Gopikrishnan, P., Meyer, M., Amaral, L.A. and Stanley, H.E., Inverse cubic law for the probability distribution of stock price variations. *Eur. Phys. J. B*, 1998, **3**, 139–140.

Gopikrishnan, P., Plerou, V., Amaral, L.A., Meyer, M. and Stanley, H.E., Scaling of the distribution of fluctuations of financial market indices. *Phys. Rev. E*, 1999, **60**, 5305–5316.

Gopikrishnan, P., Plerou, V., Gabaix, X. and Stanley, H.E., Statistical properties of share volume traded in financial markets. *Phys. Rev. E*, 2000a, **62**, 4493–4496.

Gopikrishnan, P., Plerou, V., Gabaix, X. and Stanley, H.E., Statistical properties of share volume traded in financial markets. *Phys. Rev. E*, 2000b, **62**, R4493–R4496.

Gopikrishnan, P., Rosenow, B., Plerou, V. and Stanley, H.E., Quantifying and interpreting collective behavior in financial markets. *Phys. Rev. E*, 2001, **64**, 035106-1–035106-4.

Griffin, J.E. and Oomen, R.C.A., Sampling returns for realized variance calculations: Tick time or transaction time? *Econometr. Rev.*, 2008, **27**, 230–253.

Guillaume, D., Dacorogna, M., Davé, R., Müller, U., Olsen, R. and Pictet, O., From the bird's eye to the microscope: A survey of new stylized facts of the intra-daily foreign exchange markets. *Finance Stochast.*, 1997, **1**, 95–129.

Haberman, S. and Sibbett, T.A., editors, English translation of: Louis bachelier, *Théorie de la spéculation*, *Annales scientifiques de l'Ecole Normale Superieure*. In *History of Actuarial Science* 7, 1995 (Pickering and Chatto: London).

Hasbrouck, J., *Empirical Market Microstructure: The Institutions, Economics, and Econometrics of Securities Trading*, 2007 (Oxford University Press: New York).

Hautsch, N., *Modelling Irregularly Spaced Financial Data*, 2004 (Springer: Berlin).

Hayashi, T. and Yoshida, N., On covariance estimation of non-synchronously observed diffusion processes. *Bernoulli*, 2005, **11**, 359–379.

Heston, S., A closed-form solution for options with stochastic volatility with applications to bond and currency options. *Rev. Financial Stud.*, 1993, **6**, 327–343.

Hill, B.M., A simple general approach to inference about the tail of a distribution. *Ann. Statist.*, 1975, **3**, 1163–1174.

Huth, N. and Abergel, F., The times change: Multivariate subordination, empirical facts. SSR NeLibrary, 2009.

Iori, G. and Precup, O.V., Weighted network analysis of high-frequency cross-correlation measures. *Phys. Rev. E*, 2007, **75**, 036110–7.

Itō, K. and McKean, H., *Diffusion Processes and Their Sample Paths*, 1996 (Springer: Berlin).

Ivanov, P.C., Yuen, A., Podobnik, B. and Lee, Y., Common scaling patterns in intertrade times of U. S. stocks. *Phys. Rev. E*, 2004, **69**, 056107-1–056107-7.

Kadanoff, L., From simulation model to public policy: An examination of Forrester's 'Urban Dynamics'. *Simulation*, 1971, **16**, 261–268.

Kaldor, N., Capital accumulation and economic growth. In *The Theory of Capital*, edited by F.A. Lutz and D.C. Hague, pp. 177–222, 1961 (Macmillan: London).

Keynes, J.M., *The General Theory of Employment, Interest and Money*, 1973 (The Royal Economic Society, Macmillan Press: London).

Kindleberger, C.P. and Aliber, R.Z., *Manias, Panics, and Crashes: A History of Financial Crises*, 5th ed., 2005 (Wiley: New York).

Kirman, A., Whom or what does the representative individual represent? *J. Econ. Perspect.*, 1992, 117–136.

Kullmann, L., Kertesz, J. and Mantegna, R.N., Identification of clusters of companies in stock indices via Potts super-paramagnetic transitions. *Physica A*, 2000, **287**, 412–419.

Kullmann, L., Toyli, J., Kertesz, J., Kanto, A. and Kaski, K., Characteristic times in stock market indices. *Physica A*, 1999, **269**, 98–110.

Kyle, A.S., Continuous auctions and insider trading. *Econometrica*, 1985, **53**, 1315–1335.

Laloux, L., Cizeau, P., Bouchaud, J. and Potters, M., Noise dressing of financial correlation matrices. *Phys. Rev. Lett.*, 1999, **83**, 1467–1470.

Landau, L.D., *Statistical Physics*, vol. 5 of *Theoretical Physics*, 1965 (Pergamon Press: Oxford).

Lillo, F. and Farmer, J.D., The long memory of the efficient market. *Stud. Nonlinear Dynam. Econometr.*, 2004, **8**, 1–33.

Lillo, F., Farmer, D. and Mantegna, R., Econophysics: Master curve for price-impact function. *Nature*, 2003, **421**, 129–130.

Liu, Y., Cizeau, P., Meyer, M., Peng, C.K. and Stanley, H.E., Correlations in economic time series. *Physica A*, 1997, **245**, 437–440.

Lux, T. and Sornette, D., On rational bubbles and fat tails. *J. Money, Credit, Bank.*, 2002, **34**, 589–610.

Lux, T. and Westerhoff, F., Economics crisis. *Nature Phys.*, 2009, **5**, 2–3.

Malliavin, P. and Mancino, M.E., Fourier series method for measurement of multivariate volatilities. *Finance Stochast.*, 2002, **6**, 49–61.

Mandelbrot, B., The Pareto–Levy law and the distribution of income. *Int. Econ. Rev.*, 1960, 79–106.

Mandelbrot, B., The variation of certain speculative prices. *J. Business*, 1963, **36**, 394–419.

Mantegna, R., Levy walks and enhanced diffusion in Milan stock exchange. *Physica A*, 1991, **179**, 232–242.

Mantegna, R., Hierarchical structure in financial markets. *Eur. Phys. J. B*, 1999, **11**, 193–197.

Mantegna, R., Presentation of the English translation of Ettore Majorana's paper: The value of statistical laws in physics and social sciences. *Quant. Finance*, 2005, **5**, 133–140.

Mantegna, R., The tenth article of Ettore Majorana. *Europhys. News*, 2006, **37**, 15–17.

Mantegna, R. and Stanley, H.E., *Introduction to Econophysics: Correlations and Complexity in Finance*, 2007 (Cambridge University Press: Cambridge).

Maslov, S. and Mills, M., Price fluctuations from the order book perspective – Empirical facts and a simple model. *Physica A*, 2001, **299**, 234–246.

McAleer, M. and Medeiros, M.C., Realized volatility: A review. *Econometr. Rev.*, 2008, **27**, 10–45.

Merton, R., Theory of rational option pricing. *Bell J. Econ. Mgmt Sci.*, 1973, 141–183.

Mike, S. and Farmer, J.D., An empirical behavioral model of liquidity and volatility. *J. Econ. Dynam. Control*, 2008, **32**, 200–234.

Montroll, E. and Badger, W., *Introduction to Quantitative Aspects of Social Phenomena*, 1974 (Gordon and Breach: New York).

Muni Toke, I., 'Market making' in an order book model and its impact on the bid–ask spread. In *Econophysics of Order-Driven Markets*, 2010 (Springer: Milan).

Noh, J.D., Model for correlations in stock markets. *Phys. Rev. E*, 2000, **61**, 5981–5982.

O'Hara, M., *Market Microstructure Theory*, 2nd ed., 1997 (Blackwell: Oxford).

Onnela, J.-P., Taxonomy of financial assets. Master's thesis, Helsinki University of Technology, 2000.

Onnela, J.-P., Chakraborti, A., Kaski, K. and Kertesz, J., Dynamic asset trees and portfolio analysis. *Eur. Phys. J. B*, 2002, **30**, 285–288.

Onnela, J.-P., Chakraborti, A., Kaski, K. and Kertesz, J., Dynamic asset trees and Black Monday. *Physica A*, 2003a, **324**, 247–252.

Onnela, J.-P., Chakraborti, A., Kaski, K., Kertesz, J. and Kanto, A., Dynamics of market correlations: Taxonomy and portfolio analysis. *Phys. Rev. E*, 2003b, **68**, 056110-1–056110-12.

Onnela, J.-P., Chakraborti, A., Kaski, K., Kertez, J. and Kanto, A., Asset trees and asset graphs in financial markets. *Phys. Scripta*, 2003c, **T106**, 48–54.

Osborne, M.F.M., Brownian motion in the stock market. *Oper. Res.*, 1959, **7**, 145–173.

Pagan, A., The econometrics of financial markets. *J. Empir. Finance*, 1996, **3**, 15–102.

Pareto, V., *Cours d'Economie Politique*, 1897 (Rouge: Lausanne). Reprinted as a volume of Oeuvres Completes, edited by G. Bousquet and G. Busino, 1964 (Droz: Geneve).

Parisi, G., Complex systems: A physicist's viewpoint. *Physica A*, 1999, **263**, 557–564.

Pathria, R.K., *Statistical Mechanics*, 2nd ed., 1996 (Butterworth-Heinemann: Oxford).

Plerou, V., Gopikrishnan, P., Nunes Amaral, L.A., Gabaix, X. and Eugene Stanley, H., Economic fluctuations and anomalous diffusion. *Phys. Rev. E*, 2000, **62**, R3023–R3026.

Plerou, V., Gopikrishnan, P., Rosenow, B., Amaral, L.A.N., Guhr, T. and Stanley, H.E., Random matrix approach to cross correlations in financial data. *Phys. Rev. E*, 2002, **65**, 066126-1–066126-18.

Plerou, V., Gopikrishnan, P., Rosenow, B., Amaral, L.A.N. and Stanley, H.E., Universal and nonuniversal properties of cross correlations in financial time series. *Phys. Rev. Lett.*, 1999, **83**, 1471–1474.

Podobnik, B., Ivanov, P.C., Lee, Y., Chessa, A. and Stanley, H.E., Systems with correlations in the variance: Generating power law tails in probability distributions. *Europhys. Lett.*, 2000, **50**, 711–717.

Politi, M. and Scalas, E., Fitting the empirical distribution of intertrade durations. *Physica A*, 2008, **387**, 2025–2034.

Potters, M. and Bouchaud, J.P., More statistical properties of order books and price impact. *Physica A*, 2003, **324**, 133–140.

Reif, F., *Fundamentals of Statistical and Thermal Physics*, 1985 (McGraw-Hill: Singapore).

Reno, R., A closer look at the Epps effect. *Int. J. Theor. Appl. Finance*, 2003, **6**, 87–102.

Roehner, B., *Patterns of Speculation: A Study in Observational Econophysics*, 2002 (Cambridge University Press: Cambridge).

Saha, M.N. and Srivastava, B.N., *A Treatise on Heat*, 3 ed., 1950 (The Indian Press: Allahabad).

Samuelson, P., Proof that properly anticipated prices fluctuate randomly. *Ind. Mgmt Rev*., 1965, **6**, 41–49.
Samuelson, P., *Economics*, 1998 (McGraw-Hill: Auckland).
Silva, A.C. and Yakovenko, V.M., Stochastic volatility of financial markets as the fluctuating rate of trading: An empirical study. *Physica A*, 2007, **382**, 278–285.
Sinha, S., Chatterjee, A., Chakraborti, A. and Chakrabarti, B.K., *Econophysics: An Introduction*, 2010 (Wiley–VCH: Weinheim).
Stauffer, D., de Oliveira, S.M., de Oliveira, P.M.C. and de Sa Martins, J.S., *Biology, Sociology, Geology by Computational Physicists*, 2006 (Elsevier: Amsterdam).
Taqqu, M., Bachelier and his times: A conversation with Bernard Bru. *Finance Stochast*., 2001, **5**, 3–32.
Tsay, R., *Analysis of Financial Time Series*, 2005 (Wiley–Interscience: New York).
Wyart, M. and Bouchaud, J.-P., Self-referential behaviour, overreaction and conventions in financial markets. *J. Econ. Behav. Organiz*., 2007, **63**, 1–24.