A PROCESS-RECONSTRUCTION ANALYSIS OF MARKET FLUCTUATIONS

R. VILELA MENDES

Laboratório de Mecatrónica, DEEC, Instituto Superior Técnico,
Av. Rovisco Pais, 1096 Lisboa Codex, Portugal
*Zentrum für interdisziplinäre Forschung, Universität Bielefeld,
Wellenberg 1, 33615 Bielefeld, Germany
vilela@cii.fc.ul.pt

R. LIMA*

Centre de Physique Théorique CNRS, Luminy, Case 907,
F13288 Marseille Cedex 9, France
lima@cpt.univ-mrs.fr

T. ARAÚJO*

Dept. Economia, ISEG, R. Miguel Lupi 20,
1200 Lisboa, Portugal
tanya@iseg.utl.pt

The statistical properties of a stochastic process may be described (1) by the expectation values of the observables, (2) by the probability distribution functions or (3) by probability measures on path space. Here an analysis of level (3) is carried out for market fluctuation processes. Gibbs measures and chains with complete connections are considered. Some other topics are also discussed, in particular the asymptotic stationarity of the processes and the behavior of statistical indicators of level (1) and (2). We end up with some remarks concerning the nature and origin of the market fluctuation process and its relation to the efficient market hypothesis.

Keywords: Market fluctuations; Gibbs measures; Chains with complete connections.

1. Introduction

When a physical phenomenon is measured with a set of instruments, what we register is a sequence of values of some variable X

$$\cdots X_{-2} X_{-1} X_0 X_1 X_2 \cdots$$

which takes values in a space Y. We will call Y the *state space* and the space of sequences $Y^{\mathbb{Z}}$ the *path space*. Statistical properties of the phenomenon may be described at three different levels:

(1) By the expectation values of the observables;
(2) By the probability measures on the state space Y;
(3) By the probability measures on path space Y^Z.

One obtains three different characterizations of the phenomenon which represent successively finer levels of description of the statistical properties. Borrowing a terminology used in large deviation theory [1, 2], we will call these three types of description, respectively, *level 1, 2 and 3-statistical indicators.*

To obtain expectation values and probability measures we would require infinite samples and a law of large numbers. For any finite sample we obtain finite versions of the expectation values, of the probability on state space and of the probability on path space which are called the *mean partial sums*, the *empirical measures* (or empirical probability distribution functions — pdf's) and the measures on the *empirical process*.

Level-1 and level-2 analysis are the most common ones and their statistical indicators the most commonly quoted when a stochastic process is analyzed. However to the same expectation values for the observables or to the same pdf's, different processes may be associated. Therefore full understanding of the process requires the determination of the level-3 indicators. Recent advances have been obtained on the identification of processes, especially in connection with the analysis of hydrodynamic turbulence data [3–8]. In particular it has been clarified that analysis and reconstruction of the process involves two different but related steps. One is the identification of the *grammar* of the process, that is, the allowed transitions in the state space or the subspace in path space that corresponds to actual orbits of the system. The second step is the identification of the *measure*, which concerns the occurrence frequency of each orbit in typical samples. Although largely independent from each other, this two features have a related effect on the constraints they impose on the statistical indicators.

Identification of grammars and measures (in particular Gibbs measures) has been dealt with recently, in particular in the context of hydrodynamic turbulence and other dynamical systems. Market fluctuations is an interesting stochastic process. Some analogies have been found between this process and some of the features of turbulence data [9, 10]. However, when statistical indicators are computed, it turns out that the two processes are different [11–13]. Nevertheless the statistical tools that have been developed for turbulence are mathematical devices which are not process-dependent and they may be applied to any stochastic process. Of course, underlying this approach is the working hypothesis that statistical methods, by themselves, are an appropriate tool to describe and reconstruct the market fluctuation process. This hypothesis underlies the modern view of the *efficient market*, namely the idea that overreactions and underreactions of the market are also a random process [14]. In this view, the expected value of abnormal returns would be zero. Contrariwise, if a well defined deterministic pattern of over- and underreaction is ever found then, in addition to chance, a behavioral component [15, 16] must be

included in any description of the market. Behavioral trends, however, may turn out not inconsistent with a pure statistical description if the different reaction times of the market agents are taken into account, as well as the secondary reactions of the agents to each other moves [17].

The emphasis on this paper will be on level-3 analysis and on the reconstruction of the processes. Nevertheless we have also dedicated some time to the computation, for market fluctuations, of the level-1 and level-2 statistical indicators used in the past for turbulence data. In particular the behavior of some of these indicators already provides information on the nature of the grammars. This analysis is carried out in Sec. 3. Section 4 is dedicated to the search for a Gibbs measure and, once the long-memory features of the market process are exhibited, Sec. 5 attempts to describe the process in the framework of chains with complete connections.

However, the first step in the analysis of any stochastic process is to inquire about the stationarity of the process and whether typical samples are available. This is the subject of the next section.

2. Is There an Asymptotically Stationary Market Fluctuation Phenomenon?

Large samples of high-frequency finance data are now available. However high-frequency data may not be the more appropriate data to begin understanding the stochastic process that underlies the market mechanism. This is because, when comparing minute to monthly variations for example, one is comparing systems with very different compositions, trading agents operating on the minute scale being in general different from those operating in longer time scales. This is evidenced, for example, by the different scaling laws for low and high-frequency data. In market data one faces a complexity versus statistics trade-off. The high frequency data certainly provides better statistics but it also involves the interplay of many more reaction time scales and market compositions in the trading process. For this reason, to "purify" as much as possible our samples, we have decided to concentrate on daily data. The price to be paid for this choice is the fact that, as compared for example with a large scale hydrodynamics experiment, the available amount of one-day market fluctuation data is relatively small. If, in addition, the data were non-stationary, the chances to obtain a reliable statistical analysis would be rather slim.

Reliable application of statistical mechanics tools to any kind of signal, presupposes that two conditions are fulfilled. First, that the process that generates the data has some kind of underlying stationarity or asymptotic stationarity. Second, that the time sequence that is presented to the analysis is a typical sample for the process. The second condition, of course, we can only hope that it is realized and to improve our belief in this condition several different signals of a similar nature should be analyzed (several different stocks, or currencies or markets). As to the first condition it requires some preprocessing of the data. We will concentrate in
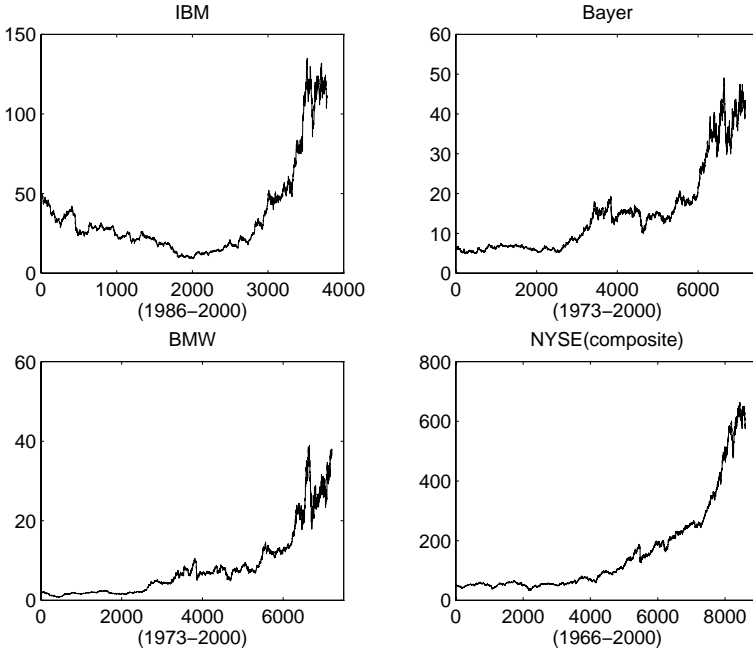
Fig. 1.   Daily price fluctuation data.

this paper in the daily fluctuation data of industrial stocks and indexes and the objective is to try to extract the features of the market process that acts on them. We look at each stock as an experimental probe that, while reacting to the market pressures, may reveal some of the mechanisms of the market process.

Market prices are by nature non-stationary entities. They fluctuate, they have general trends that depend on the general state of the economy, on the total amount of capital flowing to the market, on the general acceleration of the economy, on long and medium term political decisions and expectations, etc. Nevertheless, our hypothesis is that, if all these global factors are extracted from the data, there are still some invariant features that characterize this peculiar human phenomenon.

The type of data that will be analyzed is displayed in Fig. 1 that shows daily price data $p(t)$ for three stocks and for the NYSE composite index. Its non-stationary nature is very apparent. The first step is to extract the general trend. This is done, in a smooth way by a polynomial fit $q(t)$ (Fig. 2 shows an example, where a 7-degree polynomial is used). Figure 3 shows the difference $p(t) - q(t) + \langle p(t) \rangle$. Clearly the data is still very far from stationary because, due to the market volume acceleration, recent fluctuations carry a much larger weight. Therefore the last step is a rescaling of the data, that is

$$x(t) = p(t) \frac{\langle p(t) \rangle}{q(t)} \tag{1}$$
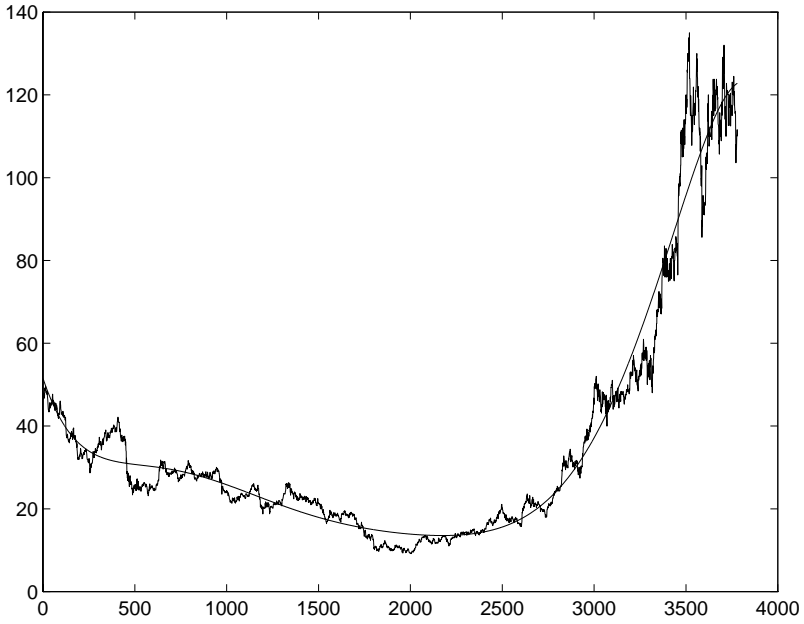
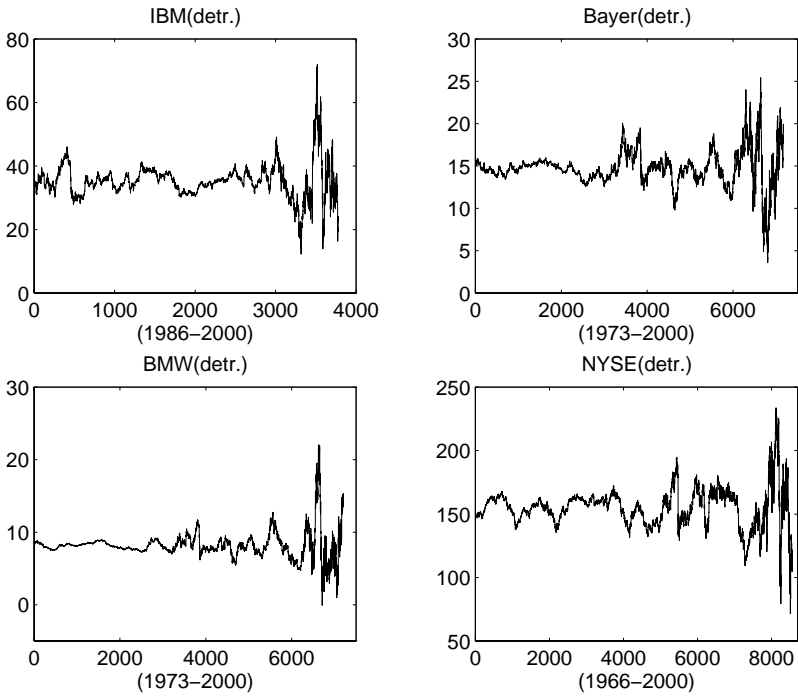Fig. 2. Detrending by a polynomial.
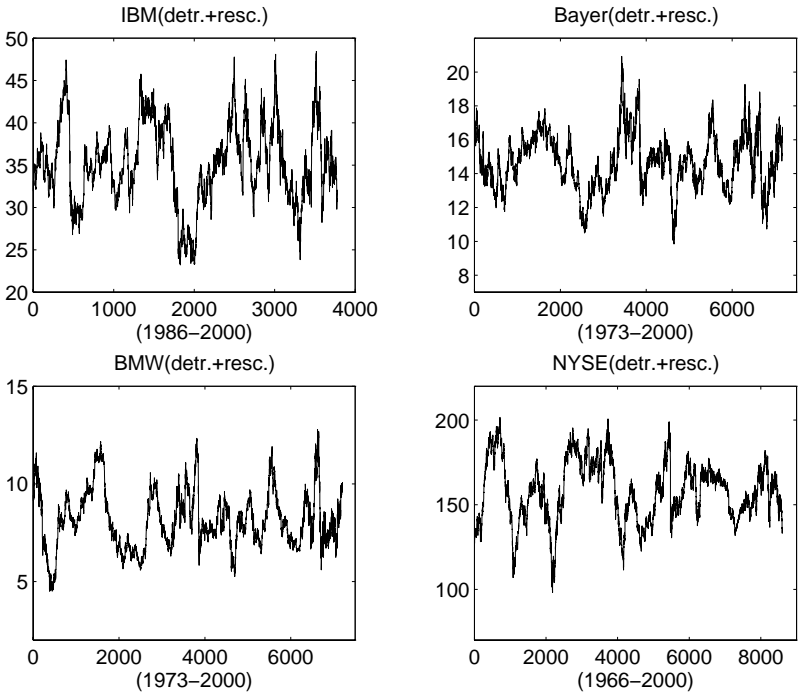

Fig. 3. Detrended data.

Fig. 4.   Detrended and rescaled data.

are the signals to be analyzed. They are shown in Fig. 4. To anyone used to examine
turbulence data, it looks as if the market signals are now somewhat stable. That
does not mean, of course, that they are stationary in the strict sense. However
it suggests that in spite of currency adjustments, increased number of players,
trade volumes and other macroeconomic indicators, there is something more or less
permanent in this human game.

Detrending and rescaling of the data is important because we will be analyzing
price differences over large time intervals. For one-day differences of log-price, the
results would be identical to those obtained from the raw data. Detrending and
rescaling the data, the overall amplitude of price fluctuations becomes reasonably
uniform over the time span of the data. However the process is not (locally) station-
ary, as seen in Figs. 5 and 6 that show the strong variation in time of the volatility
(here defined as the standard deviation of the price fluctuations). The two figures on
the left show the standard deviation computed on a sliding time window of 10 days.
On the right one compares the cumulative standard deviation for the rescaled (full
line) and the non-rescaled data (dashed line). It is quite apparent that only the
rescaled data has the chance to belong to an asymptotically stationary process.
Once the data is detrended and rescaled there is in fact no evidence [18] for an
abnormal increase, in recent times, of the volatility in the underlying process.
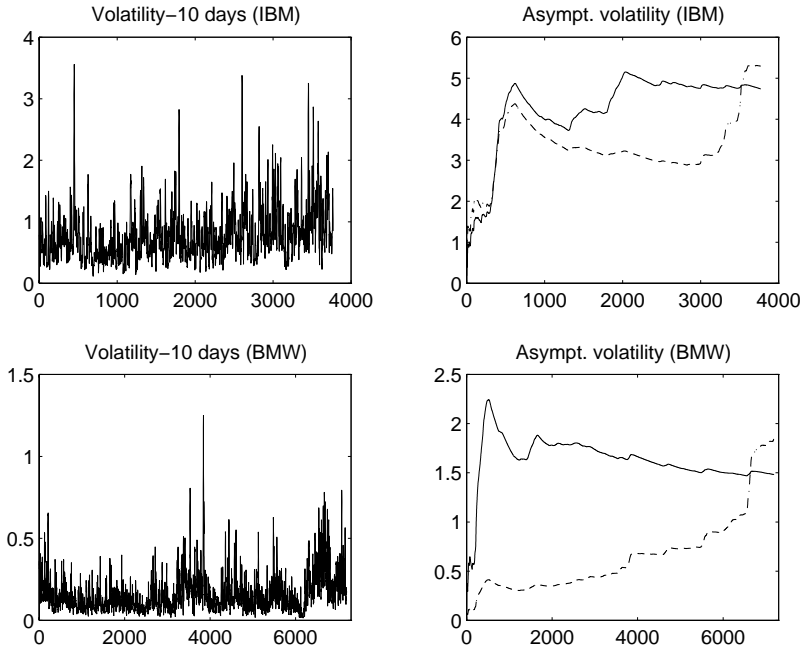
Fig. 5.   Ten-days window volatility and comparison of asymptotic volatility for the rescaled and non-rescaled data (IBM and BMW).
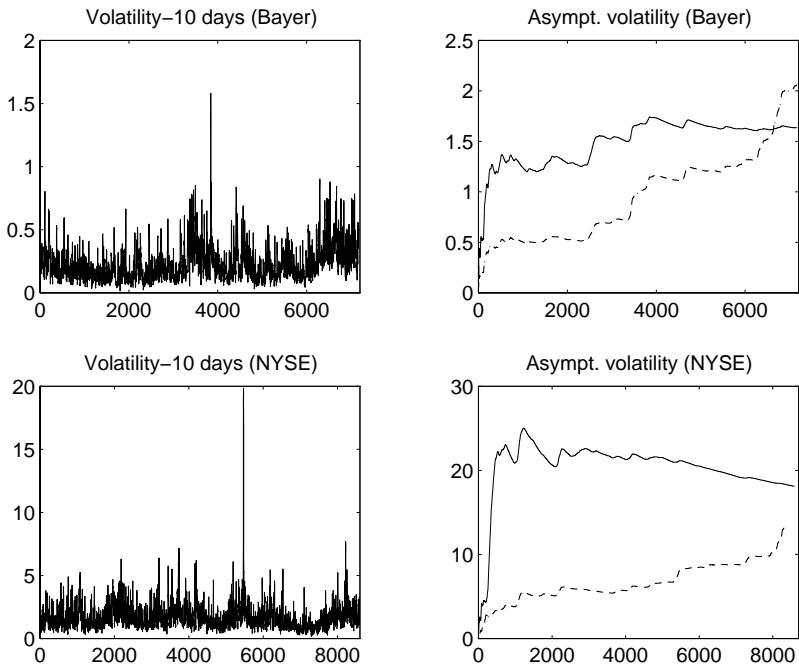

Fig. 6.   Same as in Fig. 5 for Bayer and NYSE.

A direct test of stationarity of the detrended and rescaled data was obtained by coding with a 5-symbols alphabet (as explained in Sec. 4). Then, computing the entropies of multi-symbol words, in the first and the second half of the samples, no significant difference is found.

3. Statistical Indicators for Typical Samples

Here we concentrate on level-1 and level-2 analysis of the regularized samples discussed in Sec. 2, that is, we compute quantities related to averages values and to probability distribution functions (pdf's). The level-3 analysis of the processes will be done in the latter sections.

The main variables that are used to construct the statistical indicators are the differences of log-prices

$$r(t, n) = \log p(t + n) - \log p(t) \tag{2}$$

sometimes called the n-days return. For each experimental sample, three main statistical indicators are computed:

(i) The maximum (over t) of $r(t, n)$

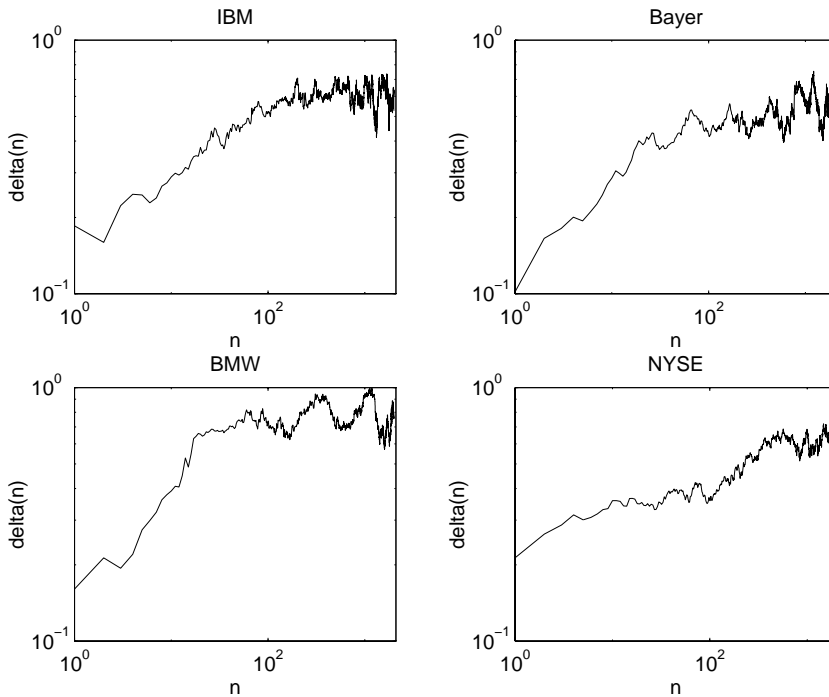$$\delta(n) = \max_t \{r(t, n)\} . \tag{3}$$

Fig. 7.   Maximum $\delta(n)$ of log-prices differences.

(ii) The moments of the distribution of $|r(t, n)|$

$$S_q(n) = \langle |r(t, n)|^q \rangle \tag{4}$$

with $\langle\ \rangle$ meaning the sample average.

(iii) If inside a certain range, the moments satisfy

$$S_q(n) \sim n^{\chi(q)}\,, \tag{5}$$

then the scaling exponent $\chi(q)$ is another important statistical indicator.

The results obtained from our detrended and rescaled samples are displayed in Figs. 7 to 9. Figure 7 refers to $\delta(n)$ and Fig. 8 shows $S_q(n)$ as a function of n for different values of q (from top to bottom $q = 1$ to $q = 8$). The large fluctuations in $\delta(n)$ for large values of n and in $S_q(n)$ for large q are quite natural given the size of the data samples.

In the range $n = 2$ to $n = 60$ the moments follow an approximate power law of the type of Eq. (5) and from the behavior in this region we have extracted the scaling exponent $\chi(q)$ shown in Fig. 9. The main conclusions from this analysis of the statistical indicators are:

(a) $\delta(n)$ is log-concave, that is, $\log \delta(n)$ is concave as a function of $\log n$, increasing and probably (with better statistics) being asymptotically constant for large r;
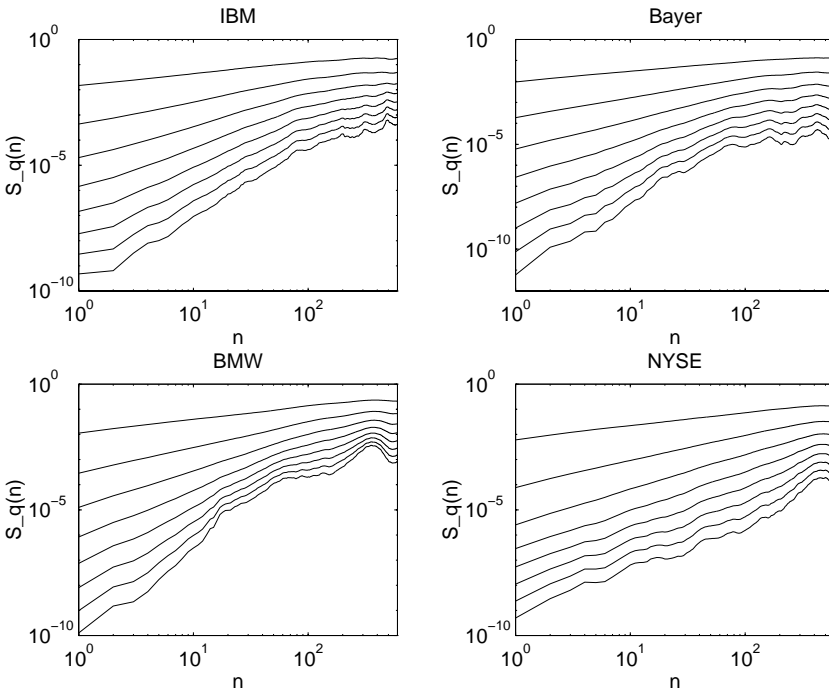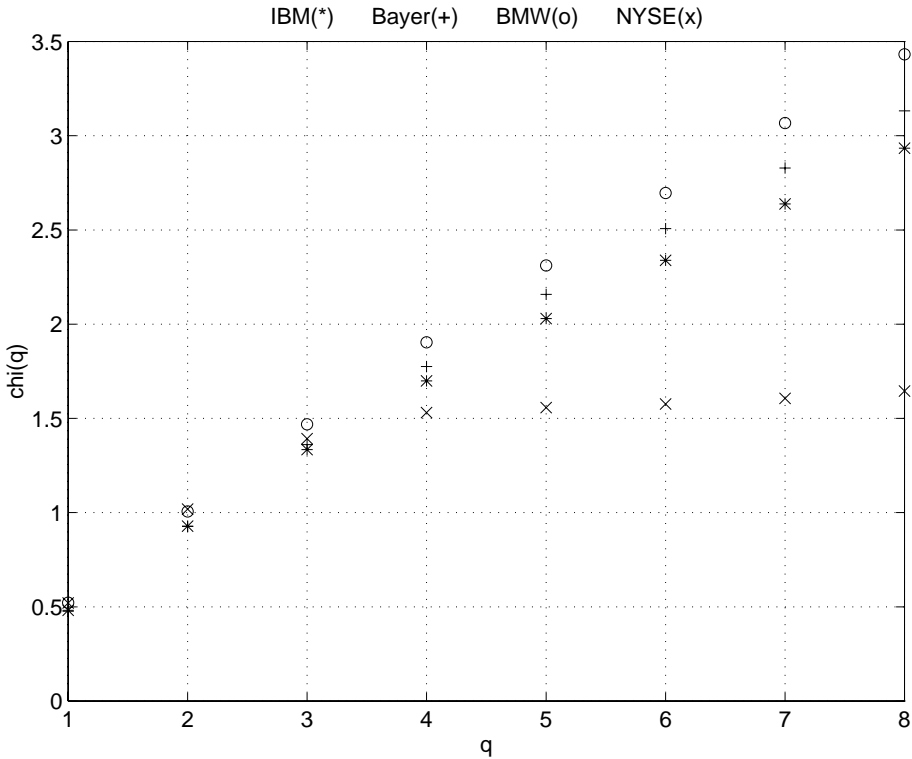

Fig. 8.   Moments of the $|r(t, n)|$ distribution.

Fig. 9.   Scaling exponent $\chi(q)$.

(b) $S_q(n)$ is also an increasing log-concave function of n, allowing a power law approximation in a limited range;

(c) The scaling law $\chi(q)$ is an increasing concave function of q;

(d) For all samples, $\chi(1)$ computed in the scaling region ($n = 2$ to $n = 60$) is very close to 0.5;

(e) The scaling properties of the NYSE index seem somewhat different from those of the other stocks. However this is only apparent for $q \geq 5$, where poor statistics effects may already be felt.

From this analysis one also obtains precise statements concerning the similarities and differences between hydrodynamic turbulence and the market fluctuation process. Properties (a) to (c) are shared by the turbulence data, although the numerical values of the statistical indicators are quite different. For example, for turbulence data $\chi(1) = \frac{1}{3}$ whereas here $\chi(1) \approx 0.5$, showing the essentially uncorrelated nature of the signal for $n \geq 2$. The correlation function of one-day returns and its absolute value

$$C(r(1), T) = \langle r(t + T, 1) r(t, 1) \rangle \tag{6}$$

and

$$C(|r(1)|, T) = \langle |r(t+T, 1)| \, |r(t, 1)| \rangle \tag{7}$$

are shown in Fig. 10. One sees that for $T \geq 2$ the returns are uncorrelated, their correlation function remaining at the noise level. In contrast the correlation for the absolute value remains non-negligible for a longer time (at least up to $T = 10$). This means that although the returns are linearly uncorrelated, non-linear functions of the returns remain correlated for longer periods.

The behavior of the statistical indicators $\delta(n)$, $S_q(n)$ and $\chi(q)$ already has some strong implications on the level-3 features of the process, namely on the structure of its grammar. In fact, without restrictions on the allowed transitions $\delta(n)$ and $S_q(n)$ would be independent of n and $\chi(q) = 0$ for all q. In particular, property (a) implies that if the process is a topological Markov chain the transitions allowed by the transition matrix T must lie inside a strictly convex domain around the diagonal of T [5].

Figure 11 illustrates the dynamics of one-day returns

$$r(t, 1) \rightarrow r(t+1, 1) \,. \tag{8}$$

It shows that the bulk of the data consists of a central core of small fluctuations with a few large flights away from this core. This structure of the data will have a strong influence on the results obtained in the next section.
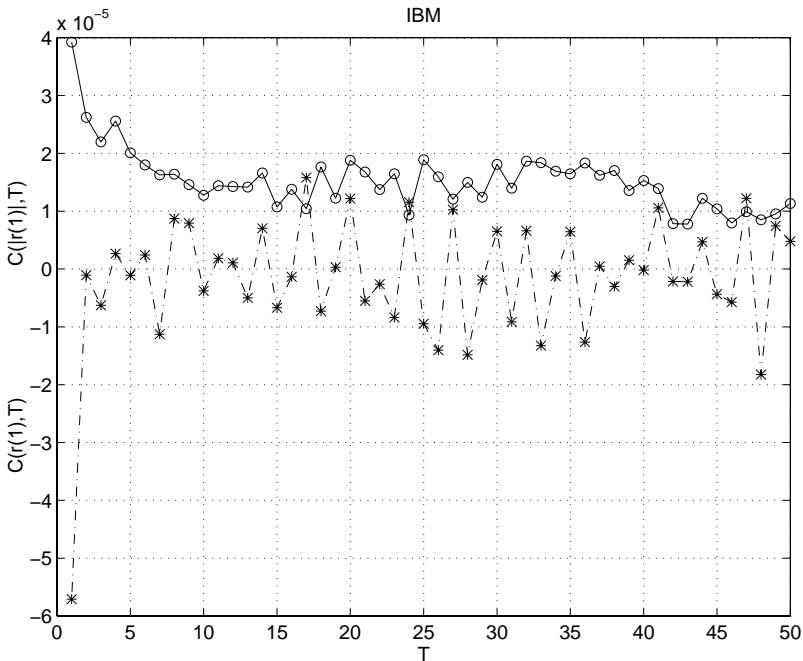

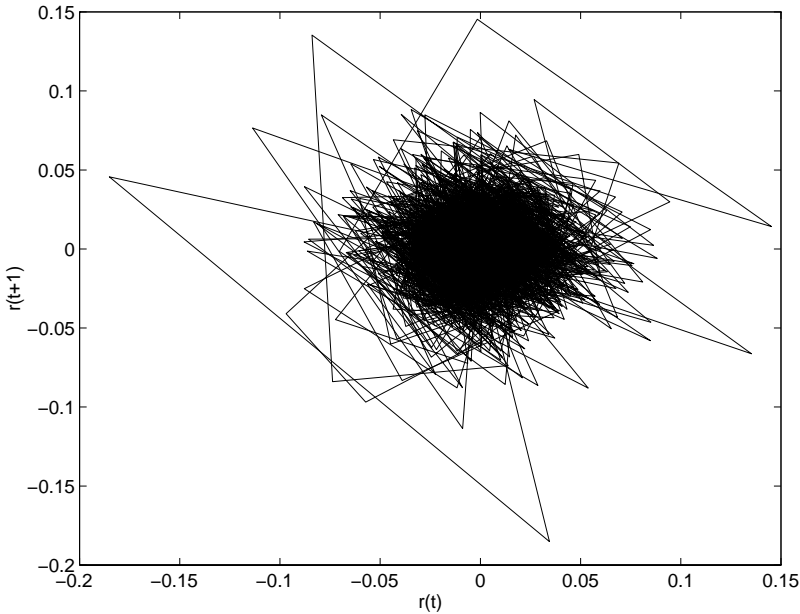Fig. 10.   Correlation functions of $r(t, 1)$ and $|r(t, 1)|$.

Fig. 11.   The dynamics of one-day returns.

4. Looking for a Gibbs Measure

Let us assume a coding of the dynamical system by a finite alphabet Σ. Then the space Ω of orbits of the system are infinite sequences $\omega = i_1 i_2 \cdots i_k \cdots$, $i_k \in \Sigma$, with the dynamical law being a shift σ on these symbol sequences.

$$\sigma\omega = i_2 \cdots i_k \cdots . \tag{9}$$

Depending on the dynamical law of the coded system, not all sequences will be allowed. The set of allowed sequences in Ω defines the *grammar* of the shift. The set of all sequences which coincide on the first n symbols is called a *n-cylinder* (or *n-block*) and is denoted $[i_1 i_2 \cdots i_n]$. The probability measures over the cylinders is the main tool that is used to characterize the dynamical properties and is a piece of information that may be inferred from the data.

A particularly important measure on the cylinders is the Gibbs measure defined by [19, 20]:

$$c_1 \le \frac{\mu([i_1(\omega)i_2(\omega)\cdots i_n(\omega)])}{\exp(-nP + (S_n\phi)(\omega))} \le c_2 \tag{10}$$

with $(S_n\phi)(\omega) = \sum_{k=0}^{n-1} \phi(\sigma^k\omega)$, ϕ being a Hölder continuous function on Ω called the *potential* and $P(\phi, G)$ a function depending on the potential and the grammar called the *pressure* of ϕ.

The (equilibrium) Gibbs measure and the pressure bear an important relationship to the entropy

$$h(\mu) = \lim_{n \to \infty} \frac{H_n}{n} = \lim_{n \to \infty} \frac{1}{n} \sum_{i_1 \cdots i_n} \mu([i_1 i_2 \cdots i_n]) \log \mu([i_1 i_2 \cdots i_n]). \tag{11}$$

This is the variational principle that states that, for each potential and grammar, the $\sup_\eta \{h(\eta) + \int \phi d\eta\}$ taken over all σ-invariant measures η is reached only for the Gibbs measure μ and equals the pressure

$$P(\phi, G) = h(\mu) + \int \phi d\mu. \tag{12}$$

The potential may be chosen in such a way that $P = 0$. Such potential is called a *normalized potential*. In this case we have the following result [7]

$$\phi(\omega) = \lim_{n \to \infty} \log \frac{\mu([i_1(\omega) \cdots i_n(\omega)])}{\mu([i_2(\omega) \cdots i_n(\omega)])}. \tag{13}$$

In principle this formula may be used to construct the potential using the empirical measures $\tilde{\mu}([i_1(\omega) \cdots i_n(\omega)])$ obtained from the experimental sample. The problem is that Eq. (13) requires the use of blocks of length n as large as possible but, for a finite sample, the statistics of such blocks suffers from large uncertainties.

For practical purposes the most important class of Gibbs measures is the one associated to finite range potentials, that is, functions on Ω that depend only on the first r symbols of a sequence $\omega \in \Omega$. The importance of finite range potentials lies in the fact that they may be used to uniformly approximate any Hölder continuous potential and, on the other hand, given a limited amount of experimental data, only finite-range potentials may be reliably inferred from experiment.

An important property of range-r potentials is that for all values $i_1 i_2 \cdots i_n$ with $n \geq r$ [7],

$$\mu([i_1 \cdots i_n]) = \frac{\mu([i_1 \cdots i_r])\mu([i_2 \cdots i_{r+1}]) \times \cdots \times \mu([i_{n-r+1} \cdots i_n])}{\mu([i_2 \cdots i_r])\mu([i_3 \cdots i_{r+1}]) \times \cdots \times \mu([i_{n-r+1} \cdots i_{n-1}])}. \tag{14}$$

We will make use of this important relation in our attempt to look for a Gibbs measure for the market fluctuation data. On the one hand the relation (14) allows to express the entropy in terms of measures of cylinders of finite length only, namely

$$h(\mu) = -\sum_{i_1 \cdots i_k} \mu([i_1 \cdots i_k]) \log \frac{\mu([i_1 \cdots i_k])}{\mu([i_1 \cdots i_{k-1}])} = H_k - H_{k-1} \tag{15}$$

for all $k \geq r$ if $r > 1$. If $r = 1 h(\mu) = H_1$. H_k is the entropy associated to cylinders of length k,

$$H_k = -\sum_{i_1 \cdots i_k} \mu([i_1 \cdots i_k]) \log \mu([i_1 \cdots i_k]). \tag{16}$$

This provides a criterium to find the range of the potential. Using the empirical cylinder probabilities one computes H_k for successively larger k. Then, the range of the potential is found when $H_k - H_{k-1}$ tends to a constant value. Once the range

is found, the potential may be constructed directly from the empirical weights $\tilde{\mu}([i_{1\cdots i_k}])$.

Another important consequence of Eq. (14) is that for $k > r$,

$$\mu([i_1\cdots i_{k+1}]) = \frac{\mu([i_1\cdots i_k])\mu([i_2\cdots i_{k+1}])}{\mu([i_2\cdots i_k])}\,. \tag{17}$$

We will use both the criterium following from Eqs. (15) and (17) to test for the possibility to construct a Gibbs measure for the market fluctuation data.

A five-symbols code Σ

$$\Sigma = \{-2, -1, 0, 1, 2\} \tag{18}$$

is used for the one-day return data $r(t)$

$$r(t) = \log p(t+1) - \log p + (t)\,. \tag{19}$$

The average $\overline{r(t)}$ and standard deviation $s = \sqrt{(r^2(t) - \overline{r(t)}^2)}$ of the returns are computed. Then,

$$(r(t) - \overline{r(t)}) > s \Longleftrightarrow 2\,,$$

$$s \geq (r(t) - \overline{r(t)}) > \frac{s}{3} \Longleftrightarrow 1\,,$$

$$\frac{s}{3} \geq (r(t) - \overline{r(t)}) > -\frac{s}{3} \Longleftrightarrow 0\,, \tag{20}$$

$$-\frac{s}{3} \geq \sigma(r(t) - \overline{r(t)}) > -s \Longleftrightarrow -1\,,$$

$$-s \geq (r(t) - \overline{r(t)}) \Longleftrightarrow -2\,.$$

This coding is used and the empirical frequencies $\tilde{\mu}([i_1\cdots i_n])$ for blocks of successively larger order n are found. Of course n cannot be arbitrarily large because of statistics. Results will not be reliable whenever 5^n is larger than the size N of the data sample. The statistical reliability may be directly tested either by comparing the number of different occurring blocks and 5^n or by observing the fall-off of the empirically computed $H_n - H_{n-1}$.

First we try to estimate a possible range for the potential using the criterium discussed above. The results are shown in Figs. 12 and 13 for the analyzed stocks and the NYSE index. The plots on the left show the quantity $H_k - H_{k-1}$ and the plots on the right compare the number $p(k)$ of occurring blocks of size k in the data with the maximum possible number, 5^k. Already for $k = 2$ the difference $H_k - H_{k-1}$ seems to stabilize, staying nearly constant until $k = 4$. After $k = 4$ it falls off, reflecting the lack of statistics also apparent in the comparison of $p(k)$ with 5^k in the right hand side plots. These results seem to suggest that the data is described by a very short range potential. Notice that for a similar analysis performed on hydrodynamic turbulence data the results are quite different with $H_k - H_{k-1}$ rising
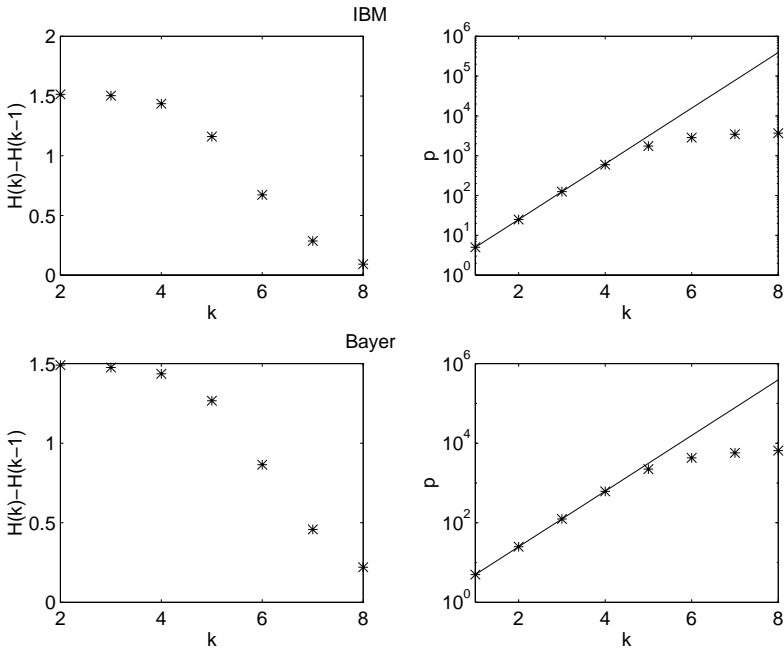
Fig. 12. $H_k - H_{k-1}$ and the number of occurring blocks of size k for $r(t)$ process with the 5-symbols alphabet (IBM and Bayer).
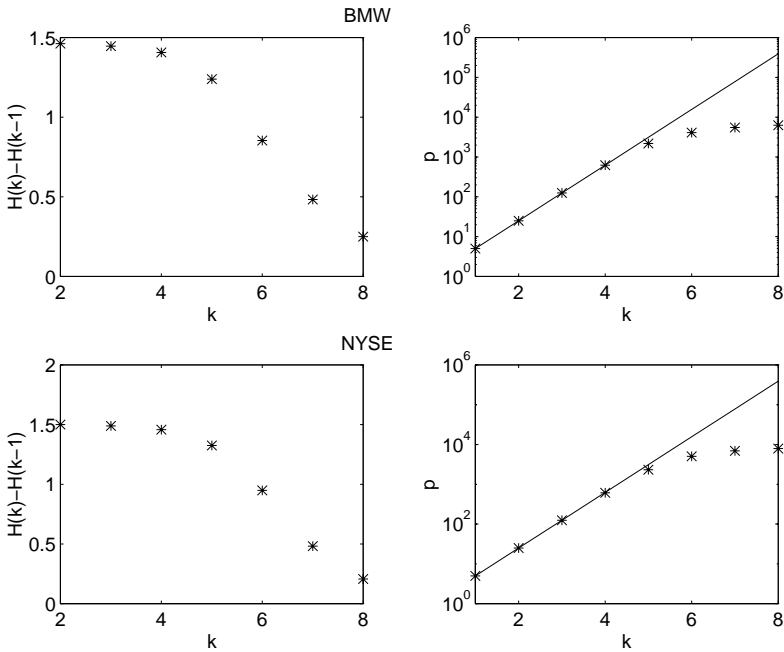

Fig. 13.   Same as Fig. 12 for BMW and NYSE.

smoothly up to a certain saturation level and then decreasing when one reaches the lack of statistics level.

To check whether the short-range potential suggested by this criterium is reliable or whether it simply results from some misleading feature of the data, we have performed the test following from Eq. (17). For successively higher k we estimate

$$\mu_e([i_1 \cdots i_{k+1}]) = \frac{\tilde{\mu}([i_1 \cdots i_k])\tilde{\mu}([i_2 \cdots i_{k+1}])}{\tilde{\mu}([i_2 \cdots i_k])}$$

from the empirical $\tilde{\mu}([i_1 \cdots i_k])$ and $\tilde{\mu}([i_1 \cdots i_{k-1}])$, which is then compared with the empirically observed $\tilde{\mu}([i_1 \cdots i_{k+1}])$. The standard deviation of the relative positive errors

$$\varepsilon_k = \max\left(0, \frac{\tilde{\mu}([i_1 \cdots i_{k+1}]) - \mu_e([i_1 \cdots i_{k+1}])}{\frac{1}{2}(\tilde{\mu}([i_1 \cdots i_{k+1}]) + \mu_e([i_1 \cdots i_{k+1}]))}\right) \tag{21}$$

is computed and the number of blocks for which this error is one and two standard deviations above the mean is computed. The result is plotted in Fig. 14 where the number of underestimation errors that are one (o) and two (∗) standard deviations away from the mean error are compared with the total number $p(k)$ of different observed blocks of each length k.

One sees that the number of large deviation errors is very large and, identifying the blocks for which these errors occur, one finds out that they all correspond to
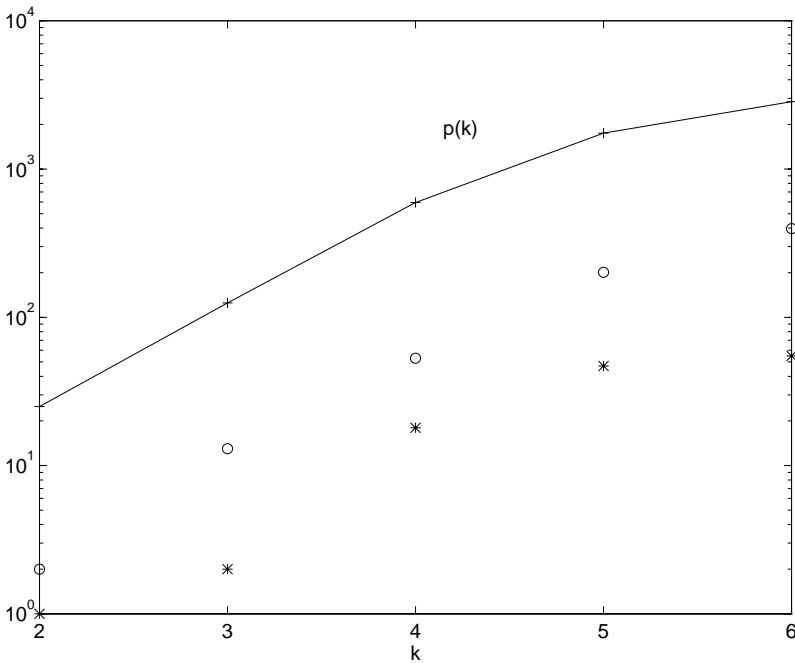

Fig. 14.   Underestimation errors one (o) and two (∗) standard deviations away from the mean and the total number $p(k)$ of observed blocks.

blocks involving large positive or negative $r(t)$'s (2 and -2). The conclusion is that a short-range potential would describe the small fluctuations in the data, the large fluctuations being badly described by it. The reason why the empirically found difference $H_k - H_{k-1}$ seems to saturate for a small k is because, as is apparent from Fig. 11, the bulk of the data consists mostly of small fluctuations plus a few large flights. The saturation of $H_k - H_{k-1}$ for small k is a reflection of the largely uncorrelated nature of the small fluctuations, whereas other features like the large deviations, persistence of non-linear correlations (volatility), etc. are not captured by a short-range potential.

Large deviations being misrepresented by an empirically constructed measure is typical of situations where the actual measure is non-Gibbsian [21, 22]. In our case, however, it may also occur that the measure is Gibbsian but with a long-range potential. This would correspond to a sharp rise of $H_k - H_{k-1}$ at $k = 2$ followed by a very slow increase above $k = 2$. In the empirical results a small increase may be hidden by the fact that, as the block length increases, the statistics becomes poorer. A large deviation analysis applied to the calculation of H_k, using a standard technique [23] to construct the free energy and the deviation function from the data, is consistent with this hypothesis.
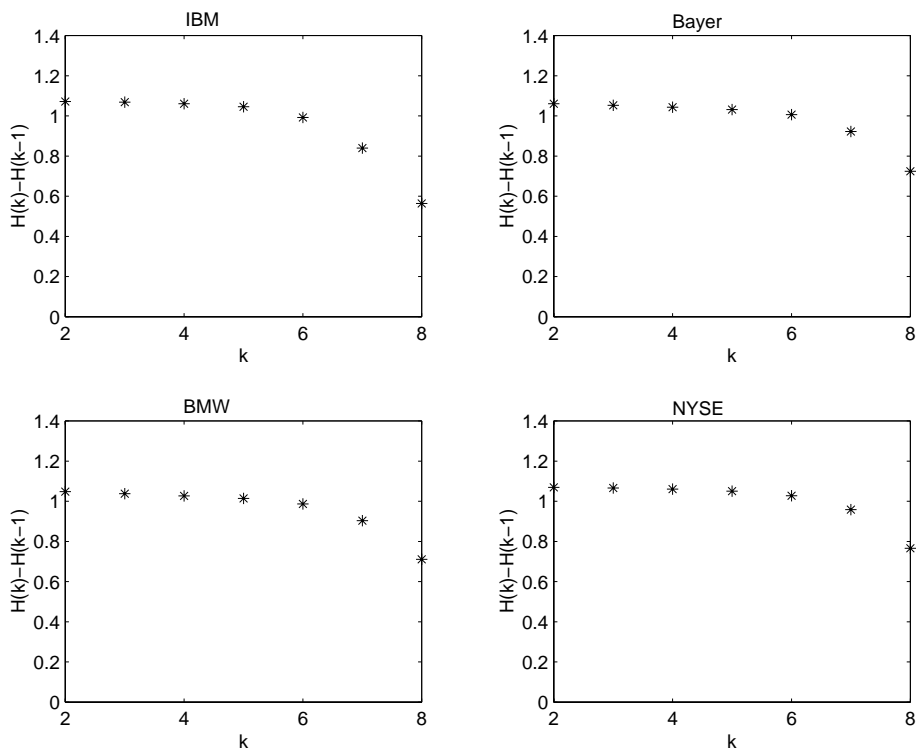

Fig. 15. $H_k - H_{k-1}$ for the $|r(t)|$ process with a 3-symbols alphabet.

The statistical dominance of the short-memory small fluctuations is apparent even for the absolute value process $|r(t)|$, where long-range correlations were found to occur (Fig. 10). This is seen in Fig. 15 where we have plotted $H_k - H_{k-1}$ for the $|r(t)|$ process with a 3-symbols alphabet $\{0, 1, 2\}$. Even in this case, a reliable estimation of the potential range and construction of a Gibbs measure do not seem possible.

In any case, whether a Gibbs measure exists or not, the finite-range potential framework does not seem to be the more convenient way to describe the market fluctuation process. In the next section we will explore another approach specially suited to deal with long-memory processes.

5. Market Fluctuations as a Chain with Complete Connections

Processes with long memory have been studied in the past. Under certain conditions, that is, when the dependence on the past does not decay too slowly, existence and uniqueness of a well defined process may be proved. A particularly well established framework is the one of chains with complete connections and summable decays ([24–26] and references therein).

A stochastic process $\{X_n\}_{n\in\mathbb{Z}}$ with alphabet Σ is said to be a *chain with complete connections* (CCC) if the following conditions are satisfied:

(a) $\forall\, a_i \in \Sigma$,

$$P(X_1 = a_1, \ldots, X_n = a_n) > 0\,. \tag{22}$$

(b) The limit

$$\lim_{m\to\infty} P(X_0 = a_0 | X_j = a_j, -m \leq j \leq -1)$$

$$= P(X_0 = a_0 | X_j = a_j, j \leq -1) \tag{23}$$

exists $\forall\, a_i,\, j \leq -1$.

(c) There is a sequence $(\gamma_m)_{m\geq 1}$ with $\lim_{m\to\infty} \gamma_m = 0$, such that for all $\{a_j, b_j \in \Sigma, j \leq -1\}$ with $a_j = b_j$ for $-m \leq j \leq -1$,

$$\left| \left(\frac{P(X_0 = a_0 | X_j = a_j, j \leq -1)}{P(X_0 = a_0 | X_j = b_j, j \leq -1)} - 1 \right) \right| \leq \gamma_m\,. \tag{24}$$

The process is said to be a *chain with complete connections and summable decay* (CCCSD) if $\sum \gamma_m < \infty$.

Conditions (a) and (b) are implicitly assumed when we considered the processes (and pre-processed the data) to be asymptotically stationary. As for the decays γ_m they may be estimated from a typical sample of the process. From the empirical probabilities for

$$P(a_0 | a_1 \cdots a_m A) = \frac{P(a_0 a_1 \cdots a_m A)}{P(a_1 \cdots a_m A)}\,, \tag{25}$$

where A is a block of arbitrary length, ones computes for each fixed set $a_0 a_1 \cdots a_m$ the maximum and the minimum over A,

$$g(a_0 a_1 \cdots a_m) = \left(\frac{\max_A P(a_0 | a_1 \cdots a_m A)}{\min_A P(a_0 | a_1 \cdots a_m A)} - 1 \right) \tag{26}$$

obtaining for γ_m

$$\gamma_m = \max_{a_0 a_1 \cdots a_m} g(a_0 a_1 \cdots a_m). \tag{27}$$

However if the statistics for very long blocks is poor, which is in general the case for finite samples, the computation of the maximum from empirical data is not reliable. A better estimate of the behavior of decay rates is obtained from the following quantity, which smooths out the large fluctuations due to poor statistics

$$g(m) = \overline{g(a_0 a_1 \cdots a_m)} \tag{28}$$

the average being taken over all sets $a_0 a_1 \cdots a_m$ of size m.

The results obtained for the 5-symbols Σ-coded data of the detrended fluctuations (BMW data) using A blocks of length 5 to 8 $(\times, +, o, *)$ is plotted in Fig. 16. Similar results are obtained for the other data. The result is compatible with exponential decay, which would probably imply the existence of a Gibbs measure (albeit with a long range potential). The data for the maxima of $g(a_0 a_1 \cdots a_m)$ displays large fluctuations and slower decay. However, with the amount of available
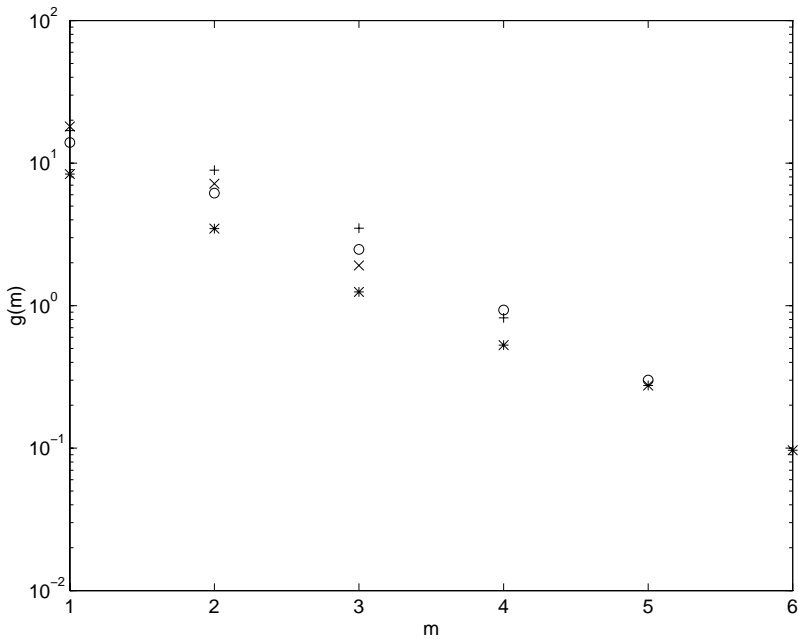

Fig. 16. $g(m)$ computed using A blocks of length 5 to 8 $(\times, +, o, *)$.

data it is not reliable for long blocks. In any case, in the present context of CCC-processes, what the result suggests is the summability of the γ_m's ($\sum_m \gamma_m < \infty$). For practical purposes the most important consequence of this fact is that a CCC-process with summable decays is the $\bar{d}$-limit of its Markov approximations of order k. The nature of this approximation should however be clearly understood.

The $\bar{d}$-*distance* [27] between two processes refers not to the processes themselves but to the process that implements the coupling of the two processes. A *coupling* between two processes $X = \{X_n\}$ and $Y = \{Y_n\}$ over the alphabet Σ is another process $\{\tilde{X}_n, \tilde{Y}_n\}$ defined over $\Sigma \times \Sigma$ such that the marginal probabilities of $\tilde{X}$ and $\tilde{Y}$ coincide with those of X and Y. Then the $\bar{d}$-distance between X and Y is

$$\bar{d}(X,Y) = \inf\{P(\tilde{X}_0 \neq \tilde{Y}_0) : \{\tilde{X}_n, \tilde{Y}_n\} \text{ is a stationary coupling of } X \text{ and } Y\}. \quad (29)$$

For some types of coupling the two processes $\tilde{X}$ and $\tilde{Y}$ are know to coincide after a certain random time. However, for the original processes X and Y, if the $\bar{d}$-distance tends to zero it does mean that the processes will coincide after a certain time. It only means that it will occur for some other processes with the same marginal probabilities.

This fact has an important bearing on the correct interpretation of the "perfect simulation" schemes [26] proposed for CCC's. Perfect simulation is always understood in the $\bar{d}$-distance sense and it does not mean perfect prediction. It simply means that a process is constructed with the same conditional probabilities of the original process, whenever the conditional probabilities of the original process are known. In practice not all conditional probabilities involving infinite pasts are needed, because going back to a regeneration time, only a finite number of back steps are required. Several simulation schemes have been proposed for CCC's with summable decays. The most important one for the applications, when the conditional probabilities are inferred from experiment, is the sequence of canonical Markov approximations of finite order k (k-CMA). A k-CMA of a process X is a Markov chain $Y^{(k)}$ of order k with conditional probabilities $P^{(k)}$ such that

$$P^{(k)}(a_0|a_1 \cdots a_k) = P(a_0|a_1 \cdots a_k) = \sum_A P(a_0|a_1 \cdots a_k A). \quad (30)$$

For a CCC X with summable decays [25]

$$\bar{d}(X, Y^{(k)}) \leq C\gamma_k, \quad (31)$$

$C > 0$ being a constant. Actually the property of the Markov approximation that is essential for the approximation result (31) is

$$\inf_A P(a_0|a_1 \cdots a_k A) \leq P^{(k)}(a_0|a_1 \cdots a_k) \leq \sup_A P(a_0|a_1 \cdots a_k A) \quad (32)$$

meaning that for Markov approximation schemes, other than the canonical one, Eq. (31) holds provided (32) is satisfied. In fact, when the conditional probabilities are inferred from limited experimental data a different Markov approximation is more convenient.

The following approximation scheme is proposed for the market fluctuation data, which we call the $\leq k$-Markov approximation:

(i) Empirical transition probabilities $\tilde{P}(a_0|a_1\cdots a_m)$ are inferred from the occurrence probability of blocks of order $m+1$ up to a certain order m_{Max}. Of course, only probabilities that correspond to blocks $a_1\cdots a_m$ that appear in the data will be available and especially for large m many will be missing.

(ii) For the simulation, with an approximation of order k, one looks at the current block $(a_1\cdots a_k)$ of order k and uses the k-empirical probability to infer the next state a_0. If that block has not appeared in the data, that was used to construct the empirical probabilities, then one looks at the $k-1$ sized block $a_1\cdots a_{k-1}$ and uses the $k-1$ order empirical probabilities. If necessary the process is repeated until an available empirical probability is found. This is the reason why this is called the $\leq k$-Markov approximation.

This approximation scheme has been applied to the market fluctuation data and for each k-order the successor a_0 of each block $(a_1\cdots a_k)$ is compared with a prediction $\tilde{a}_0$ obtained by throwing a random number with the empirical probabilities $\tilde{P}(a_0|a_1\cdots a_k)$. Figure 17 shows some of the results. In all cases the quantity that
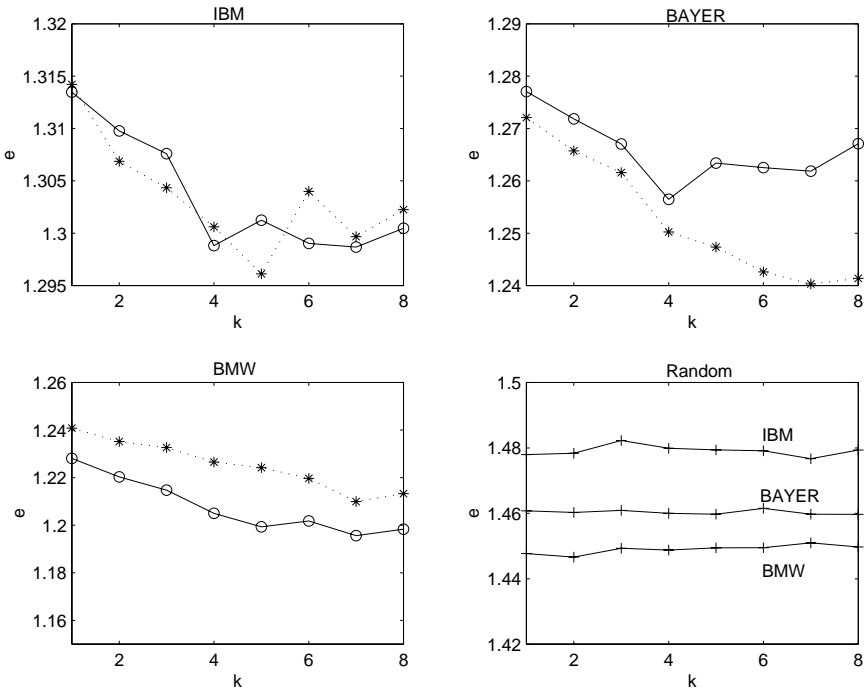

Fig. 17.   The past predicting the future (o) and the future predicting the past (∗), 5-symbols alphabet.

is plotted is the averaged error magnitude

$$e = \langle |\tilde{a}_0 - a_0| \rangle \tag{33}$$

the average being taken over the samples and 100 different runs. The two upper plots and the left lower plot show the results obtained (for each approximation order k) when half of the data for each company is used to predict the other half. The points labelled ($\circ$) correspond to the past used to predict the future and those labelled ($*$) to the future used to predict the past. Finally the right lower plot shows the results obtained when $\tilde{a}_0$ is chosen at random. The main conclusions that may be extracted from these results are:

- The average prediction obtained from using the empirical probabilities is better than a random choice.
- However, the main improvement is a result of a correct accounting of the two-symbol probabilities ($k = 1$).
- After the improvement due to the use of the lowest order blocks a small (but consistent) improvement is found by using the past information up $k = 4$ or 5. No significant improvement is obtained by using higher order approximations. This is consistent with the poorer statistics of large blocks. Actually for each individual simulation the result of using $k > 5$ leads to much larger fluctuations.

The main conclusion is that although the bulk of the data is represented by a short-memory process, there is nevertheless evidence for a small long-memory component that is captured by the higher-order Markov approximations. Depending on the amount of data that is available to infer the empirical conditional probabilities there is a maximum $k = k_m$ that should be used for the simulation process. This k_m value may be estimated from the quantity $p(k)$ plotted in Figs. 12 and 13. Finally, although a mild gain is obtained from using k_m-block probabilities rather than one-symbol probabilities, it should be remembered that perfect simulation in the $\bar{d}$-distance sense is not perfect prediction for the actual process. This is a point to keep in mind when attempting to develop any trading strategies based on the empirical block probabilities.

We have also explored the use of the empirical probabilities of one company to predict the behavior of the others. In all cases the improvement coming from the one-symbol probabilities (as compared to random choice) is obtained. This means that the one-symbol probabilities are similar in all companies. However for the long-memory component the behavior is very much company-dependent. For example there seems to be no correlation of this component between IBM and the other two companies, with the prediction being actually worse when the empirical probabilities for longer blocks are used. The same happens also when the empirical probabilities of BMW and Bayer are used to predict IBM. However there is some statistical correlation between the long-memory components (and some mild prediction improvement) between BMW and Bayer. This suggests that the statistical short-memory component of the market process might be similar for

many different stocks, whether the long-memory component might be different from market to market and to divide the stocks into classes. A similar conclusion follows from the stocks taxonomy obtained by Mantegna [28], although that work does not distinguish between the short- and long-memory components of the process.

In an attempt to distinguish between small and large fluctuations, we have also analysed the data using a 3-symbols alphabet $\{-1, 0, 1\}$ with

$$(r(t) - \overline{r(t)}) > s \Longleftrightarrow 1\,,$$

$$s \geq (r(t) - \overline{r(t)}) > -s \Longleftrightarrow 0\,,$$

$$-s \geq (r(t) - \overline{r(t)}) \Longleftrightarrow -1\,,$$

s being the standard deviation as in (20). With this coding, small fluctuations have better statistics but the number of large events is the same as before.

As before, we have run a $\leq k$-Markov approximation with half the data used to predict the other half. The error magnitude averaged over 500 runs is shown in Fig. 18. One notices a prediction improvement extending to block sizes longer than before. This seems significant because, small fluctuation errors being decreased by better statistics, the persistence of the improvement for longer blocks seems to put into evidence the long-memory component.
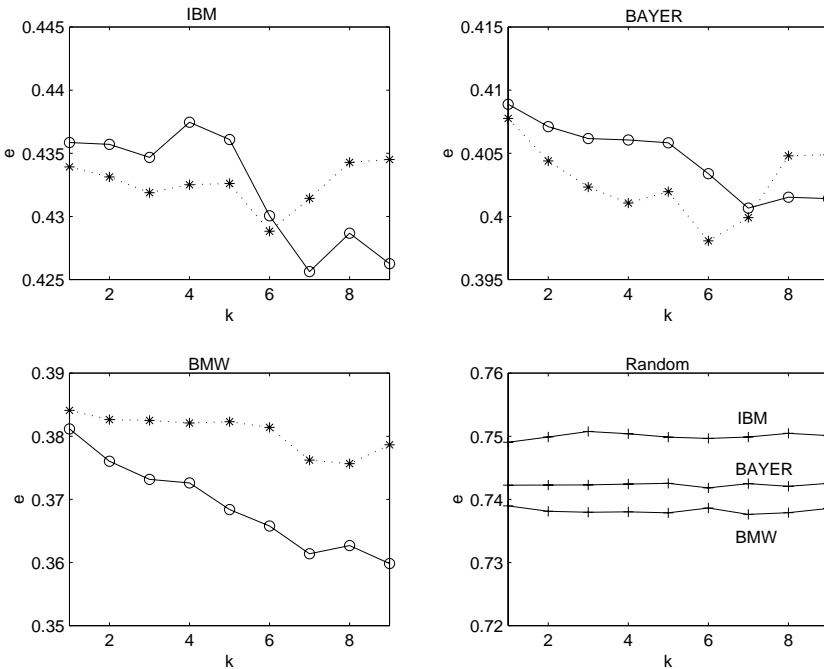

Fig. 18.   The past predicting the future (o) and the future predicting the past (∗), 3-symbols alphabet.

6. Conclusions

6.1 Statistical tools, originally developed to analyse turbulence, may be applied to any stochastic process. However, it must be emphasized that, as already noticed in the past [11–13], the statistical indicator values obtained for the financial process are quite different from those of hydrodynamic turbulence. In particular the scaling exponent $\chi(q)$ and the k-dependence of $H_k - H_{k-1}$ are quite different.

6.2 The bulk of the market fluctuation process seems to be a short-memory process. In addition it has a small long-memory component, which however is very important for practical purposes because it is associated with the large fluctuations of the returns.

6.3 Existence of the long-memory component suggests the *chains with complete connections and summable decays* as the appropriate framework to describe these processes. Although the decays may be exponentially converging, the lack of accurate data concerning long blocks prevents an accurate description by a finite range Gibbs potential.

6.4 The sequence of empirical based $\leq k$-Markov approximations discussed in Sec. 5 seems the most unbiased simulation of the process. Eventual convergence in the $\bar{d}$-distance sense is expected to hold because the market fluctuation process seems to fit in the framework of chains with complete connections and summable decays.

6.5 Except for cases where one is sure of the existence of a finite potential, Markov approximations must always be used if only finite data is available. This is true whether a Gibbs measure exists or not. What the chains with complete connections framework provides though, is a rationale for the convergence of the Markov approximations and a criterium to estimate, through the γ_m decays, how good this approximation is. Notice however the trade-off between higher order approximations and lack of statistics, that leads to an optimal block length for the empirical probabilities to be used in the simulations.

6.6 As work for the future we point out that it would be interesting to analyze in this framework the high frequency market data. Here however attention should be paid to the possibly multi-scale and multi-component nature of the processes.

References

[1] D. W. Stroock, *An Introduction to the Theory of Large Deviations* (Springer, New York, 1984).

[2] R. S. Ellis, *Entropy, Large Deviations and Statistical Mechanics* (Springer, New York, 1985).

[3] E. Ugalde, Self-similarity and finite-time intermittent effects in turbulent sequences, *J. Phys. A: Math. Gen.* **29** (1996) 4425–4443.

[4] E. Ugalde and R. Lima, On a discrete dynamical model for local turbulence, *Physica D* **95** (1996) 144–157.

[5] R. Lima and E. Ugalde, Dynamical lecture of statistical turbulence, *J. Phys. IV France* **8** (1998) 157–162.

[6] P. Collet, A. Galves and A. Lopes, Maximum likelihood and minimum entropy identification of grammars, *Random and Computational Dynamics* **3** (1995) 241–250.

[7] J.-R. Chazottes, E. Floriani and R. Lima, Relative entropy and identification of Gibbs measures in dynamical systems, *J. of Stat. Phys.* **90** (1998) 697–725.

[8] J.-R. Chazottes, Entropie relative, Dynamique symbolique et Turbulence, Ph.D. Thesis, Univ. Provence-Aix-Marseille I (1999), http://www.cpt.univ-mrs.fr.

[9] S. Ghashghaie, W. Breymann, J. Peinke, P. Talkner and Y. Dodge, Turbulent cascades in foreign exchange markets, *Nature* **381** (1996) 767–770.

[10] W. Breymann, S. Ghashghaie and P. Talkner, A stochastic cascade model for FX dynamics, *Int. J. of Theoretical and Applied Finance*, to appear.

[11] R. N. Mantegna and H. E. Stanley, Turbulence and financial markets, *Nature* **383** (1996) 587–588.

[12] A. Arneodo, J.-P. Bouchaud, R. Cont., J.-F. Muzy, M. Potters and D. Sornette, Comment on "Turbulent cascades in foreign exchange markets", cond-math/9607120.

[13] R. N. Mantegna and H. E. Stanley, Scaling behavior in the dynamics of an economic index, *Nature* **376** (1995) 46–49.

[14] E. F. Fama, Market efficiency, long-term returns and behavioral finance, *Journal of Financial Economics* **49** (1998) 283–306.

[15] N. Barberis, A. Shleifer, R. Vishny, A model of investor sentiment, *Journal of Financial Economics* **49** (1998) 307–343.

[16] W. DeBondt and R. Thaler, Does the stock market overreact?, *Journal of Finance* **40** (1985) 793–805.

[17] R. B. Olsen, M. M. Dacorogna, U. A. Müller and O. V. Pictet, Going back to basics — Rethinking market efficiency, Olsen & Associates discussion paper RBO.1992-09-07.

[18] D. Laster and K. Cole, Has the stock market grown more volatile?, *Current Issues in Economics and Finance* **2**(11) (1996) 1–6.

[19] R. Bowen, *Equilibrium states and the ergodic theory of Anosov diffeomorphisms*, Lecture Notes in Mathematics 470 (Springer-Verlag, Berlin).

[20] D. Ruelle, *Thermodynamic formalism*, Encyclopedia of Mathematics and its Applications 5 (Addison-Wesley, 1978).

[21] A. C. D. van Enter and J. Lörinczi, Robustness of the non-Gibbsian property, *J. Phys. A* **29** (1996) 2465–2473.

[22] R. Fernández, Measures for lattice systems, *Physica A* **263** (1999) 117–130.

[23] J. Seixas and R. Vilela Mendes, Large deviation analysis of multiplicity fluctuations, *Nuclear Physics* **B383** (1992) 622–642.

[24] O. Onicescu and G. Mihoc, Le comportement asymptotique des chaines à liaisons complètes, *Disq. Math. Phys.* **1** (1940) 61–62.

[25] X. Bressaud, R. Fernández and A. Galves, Speed of $\bar{d}$-convergence for Markov approximations of chains with complete connections. A coupling approach, *Stoch. Proc. and Appl.* **83** (1999) 127–138.

[26] F. Comets, R. Fernández and P. A. Ferrari, Processes with long memory: Regenerative construction and perfect simulation, arXiv:math.PR/0009204.

[27] D. S. Ornstein and B. Weiss, How sampling reveals a process, *The Annals of Prob.* **18** (1990) 905–930.

[28] R. N. Mantegna, Hierarchical structure in financial markets, cond-math/9802256.