

Medium and small-scale analysis of financial data

Andreas P. Nawroth, Joachim Peinke*

Institut für Physik, Carl-von-Ossietzky Universität Oldenburg, D-26111 Oldenburg, Germany

Available online 30 March 2007

Abstract

A stochastic analysis of financial data is presented. In particular we investigate how the statistics of log returns change with different time delays τ . The scale-dependent behaviour of financial data can be divided into two regions. The first time range, the small-timescale region (in the range of seconds) seems to be characterised by universal features. The second time range, the medium-timescale range from several minutes upwards can be characterised by a cascade process, which is given by a stochastic Markov process in the scale τ . A corresponding Fokker–Planck equation can be extracted from given data and provides a non-equilibrium thermodynamical description of the complexity of financial data.

© 2007 Elsevier B.V. All rights reserved.

Keywords: Econophysics; Financial markets; Stochastic processes; Fokker–Planck equation

1. Introduction

One of the outstanding features of the complexity of financial markets is that very often financial quantities display non-Gaussian statistics often denoted as heavy tailed or intermittent statistics [1–9]. To characterise the fluctuations of a financial time series $x(t)$, most commonly quantities like returns, log returns or price increments are used. Here, we consider the statistics of the log return $y(\tau)$ over a certain timescale τ , which is defined as

$$y(\tau) = \log x(t + \tau) - \log x(t), \quad (1)$$

where $x(t)$ denotes the price of the asset at time t . A common problem in the analysis of financial data is the question of stationarity for the discussed stochastic quantities. In particular we find in our analysis that the methods seem to be robust against nonstationarity effects. This may be due to the data selection. Note that the use of (conditional) returns of scale τ corresponds to a specific filtering of the data. Nevertheless the particular results change slightly for different data windows, indicating a possible influence of nonstationarity effects. In this paper we focus on the analysis and reconstruction of the processes for a given data window (time period). For further information concerning stationarity and the methods used here, refer to Refs. [10,11]. The analysis presented is mainly based on Bayer data for the time span of 1993–2003. The financial

*Corresponding author.

E-mail address: peinke@uni-oldenburg.de (J. Peinke).

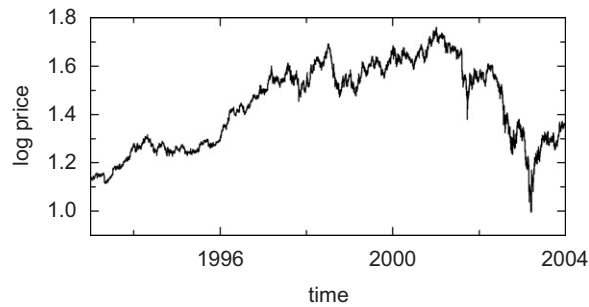


Fig. 1. Log price for Bayer for the years 1993–2003.

data sets were provided by the Karlsruher Kapitalmarkt Datenbank (KKMDB) [12]. In order to illustrate the underlying data, the graph of the logarithm of the price time series is shown in Fig. 1.

2. Small-scale analysis

The statistics of $y(\tau)$ are shown in Fig. 2. One remarkable feature of financial data is the fact that the probability density functions (pdfs) are not Gaussian, but exhibit heavy tailed shapes. Another remarkable feature is the change of the shape with the size of the scale variable τ . To analyse the changing statistics of the pdfs with the scale τ a non-parametric approach is chosen. The distance between the pdf $p(y(\tau))$ on a timescale τ and a pdf $p_T(y(T))$ on a reference timescale T is computed. As a reference timescale, $T = 1$ s is chosen, which is close to the smallest available timescale in our data sets and on which there are still sufficient events. The independence of the results with respect to the chosen timescale within a certain range is shown in Ref. [13]. In order to be able to compare the shape of the pdfs and to exclude effects due to variations of the mean and variance, all pdfs $p(y(\tau))$ have been normalised to a zero mean and a standard deviation of 1.

As a measure to quantify the distance between the two distributions $p(y(\tau))$ and $p_T(y(T))$, the Kullback–Leibler entropy [14]

$$d_K(\tau) := \int_{-\infty}^{+\infty} dy p(y(\tau)) \ln \left(\frac{p(y(\tau))}{p_T(y(T))} \right) \quad (2)$$

is used. In Fig. 3 the evolution of d_K with increasing τ is illustrated. This quantifies the change of the shape of the pdfs. For different stocks we found that for timescales smaller than about 1 min a linear growth of the distance measure seems to be universally present, see Fig. 3a. If a normalised Gaussian distribution is taken as a reference distribution, the fast deviation from the Gaussian shape in the small-timescale regime becomes evident, as displayed in Fig. 3b. For larger timescales d_K remains approximately constant, indicating a very slow change of the shape of the pdfs, in accordance with Ref. [15]. The independence of this small-scale behaviour of the particular choice of the measure and on the choice of the stock is shown in Ref. [13].

3. Medium scale analysis

Next the behaviour for larger timescales ($\tau > 1$ min) is discussed. We proceed with the idea of a cascade. As it has been shown in Refs. [9,16,17] it is possible to grasp the complexity of financial data by cascade processes running in the variable τ . In particular it has been shown that it is possible to estimate directly from given data a stochastic cascade process in the form of a Fokker–Planck equation [16,17]. The underlying idea of this approach is to access statistics of all orders of the financial data by the general joint n -scale probability densities $p(y_1, \tau_1; y_2, \tau_2; \dots; y_N, \tau_N)$. Here we use the shorthand notation $y_1 = y(\tau_1)$ and take without loss of generality $\tau_i < \tau_{i+1}$. The smaller log returns $y(\tau_i)$ are nested inside the larger log returns $y(\tau_{i+1})$ with common end point t .

The joint pdfs can be expressed as well by the multiple conditional probability densities $p(y_i, \tau_i | y_{i+1}, \tau_{i+1}; \dots; y_N, \tau_N)$. This very general n -scale characterisation of a data set, which contains the

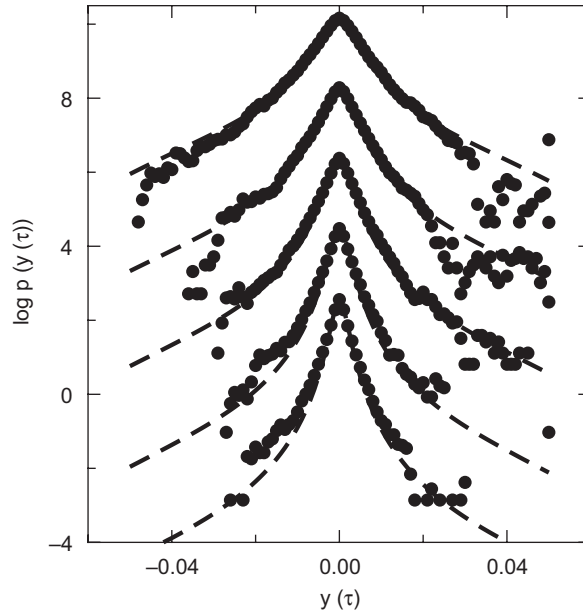


Fig. 2. Unconditional probability densities $p(y(\tau))$ for the timescales of $\tau = 240, 454, 955, 1800$ and 3766 s (bottom up) obtained from the original data (dots) and reconstructed from the extracted Fokker–Planck equation (dashed lines). The pdfs for different scales are shifted in y direction for clarity of presentation.

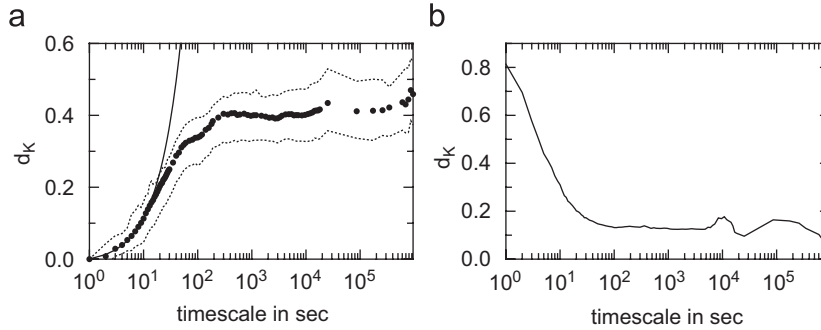


Fig. 3. Distance measure d_K for a reference distribution $p_T(y)$ for Bayer data. (a) As a reference timescale $T = 1$ s is chosen. The bold dots represent the estimated value, the dotted lines the one-sigma error bound and the solid line the linear fit for the first region, after Ref. [13]. (b) A normalised Gaussian distribution is chosen as a reference distribution $p_T(y)$.

general n -point statistics, can be simplified essentially if there is a stochastic process in τ , which is a Markov process. This is the case if the conditional probability densities fulfil the following relations:

$$p(y_1, \tau_1 | y_2, \tau_2; y_3, \tau_3; \dots; y_N, \tau_N) = p(y_1, \tau_1 | y_2, \tau_2). \tag{3}$$

Consequently,

$$p(y_1, \tau_1; \dots; y_N, \tau_N) = p(y_1, \tau_1 | y_2, \tau_2) \cdot \dots \cdot p(y_{N-1}, \tau_{N-1} | y_N, \tau_N) \cdot p(y_N, \tau_N) \tag{4}$$

holds. Eq. (4) indicates the importance of the conditional pdf for Markov processes. Knowledge of $p(y, \tau | y_0, \tau_0)$ (for arbitrary scales τ and τ_0 with $\tau < \tau_0$) is sufficient to generate the entire statistics of the increment, encoded in the N -point probability density $p(y_1, \tau_1; y_2, \tau_2; \dots; y_N, \tau_N)$.

For Markov processes the conditional probability density satisfies a master equation, which can be put into the form of a Kramers–Moyal expansion for which the Kramers–Moyal coefficients $D^{(k)}(y, \tau)$ are defined as

the limit $\Delta\tau \rightarrow 0$ of the conditional moments $M^{(k)}(y, \tau, \Delta\tau)$:

$$D^{(k)}(y, \tau) = \lim_{\Delta\tau \rightarrow 0} M^{(k)}(y, \tau, \Delta\tau), \quad (5)$$

$$M^{(k)}(y, \tau, \Delta\tau) = \frac{\tau}{k! \Delta\tau} \int_{-\infty}^{+\infty} (\tilde{y} - y)^k p(\tilde{y}, \tau - \Delta\tau | y, \tau) d\tilde{y}. \quad (6)$$

For a general stochastic process, all Kramers–Moyal coefficients are different from zero. According to Pawula’s theorem, however, the Kramers–Moyal expansion stops after the second term, provided that the fourth order coefficient $D^{(4)}(y, \tau)$ vanishes. In that case, the Kramers–Moyal expansion reduces to a Fokker–Planck equation (also known as the backwards or second Kolmogorov equation):

$$-\tau \frac{\partial}{\partial \tau} p(y, \tau | y_0, \tau_0) = \left\{ -\frac{\partial}{\partial y} D^{(1)}(y, \tau) + \frac{\partial^2}{\partial y^2} D^{(2)}(y, \tau) \right\} p(y, \tau | y_0, \tau_0). \quad (7)$$

$D^{(1)}$ is denoted as drift term, $D^{(2)}$ as diffusion term. The probability density $p(y, \tau)$ has to satisfy the same equation, as can be shown by a simple integration of Eq. (7).

4. Results for Bayer data

From the data shown in Fig. 1 the Kramers–Moyal coefficients were calculated according to Eqs. (5) and (6). The timescale was divided into half-open intervals

$$\left[\frac{1}{2}(\tau_{i-1} + \tau_i), \frac{1}{2}(\tau_i + \tau_{i+1}) \right]$$

assuming that the Kramers–Moyal coefficients are constant with respect to the timescale τ in each of these subintervals of the timescale. The smallest timescale considered was 240 s and all larger scales were chosen such that $\tau_i = 0.9 \cdot \tau_{i+1}$. The Kramers–Moyal coefficients themselves were parameterised in the following form:

$$D^{(1)} = \alpha_0 + \alpha_1 y, \quad (8)$$

$$D^{(2)} = \beta_0 + \beta_1 y + \beta_2 y^2. \quad (9)$$

The coefficients obtained by this procedure are shown in Fig. 4. This result shows that the rich and complex structure of financial data, expressed by multi-scale statistics, can be pinned down to coefficients with a relatively simple functional form.

As mentioned in the Introduction the parameterisation of the Kramers–Moyal coefficients in Eqs. (8) and (9) assumes that there is no explicit dependence on the time, i.e., the Markov process in scale is stationary with respect to the time. We have analysed the influence of nonstationarity by splitting the data set into two sub-sets of equal size and checked whether the coefficients are the same. For both data sets the coefficients α_0 , β_0 and β_1 are again essentially zero. For α_1 and β_2 there are only minor differences between both sub-sets, which are about 0.26 (MAE) and 0.26 (RMSE) for α_1 and 0.09 (MAE) and 0.12 (RMSE) for β_2 . The differences become even smaller if each sub-set is normalised to the same mean and standard deviation. These errors are close to the intrinsic error due to the finite data on which the estimation process for the Kramers–Moyal coefficient is based.

To show the quality of our results we reconstructed the measured statistics by the estimated Fokker–Planck equations. At first, the conditional probability densities $p(y(\tau_i) | y(\tau_{i+1}))$ were reconstructed. As an example the conditional probability density $p(y(\tau = 3389 \text{ s}) | y(\tau = 3766 \text{ s}))$ is shown in Fig. 5. The reconstructed conditional probability density and the one calculated directly from the data are in good agreement. As a next step we used the pdf on the scale of $\tau = 27,900 \text{ s}$ and the reconstructed conditional probability densities to calculate the increment pdfs on timescales between 4 min and 1 h. The results for the timescales of $\tau = 3766, 1800, 955, 454$ and 240 s are shown in Fig. 2. Again the agreement between unconditional probability densities $p(y(\tau))$ of the original data (dots) and the reconstructed ones (broken lines) is very good.

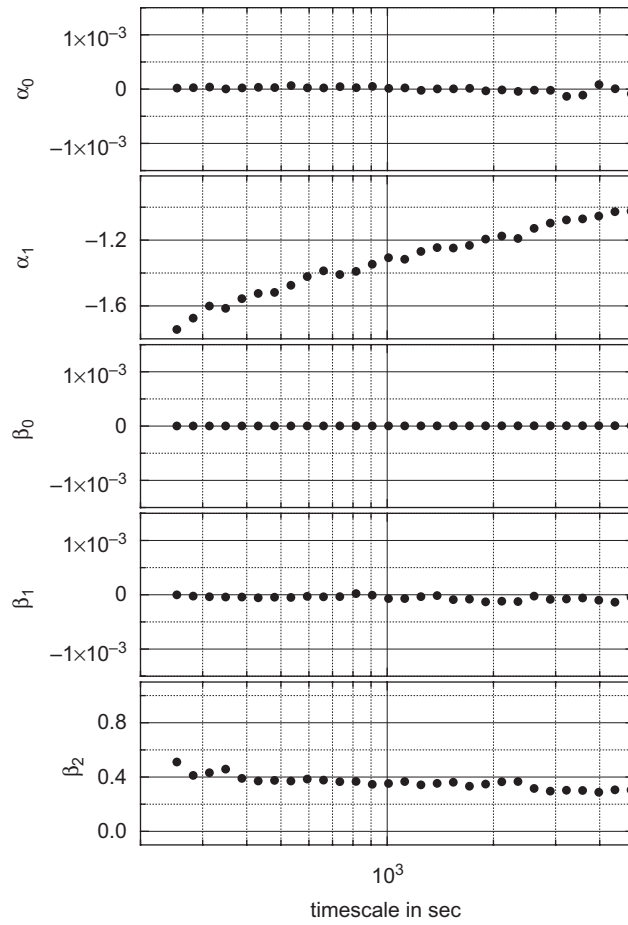


Fig. 4. The parameters α_0 , α_1 , β_0 , β_1 and β_2 of the parameterisation of the Kramers–Moyal coefficients used for the reconstruction.

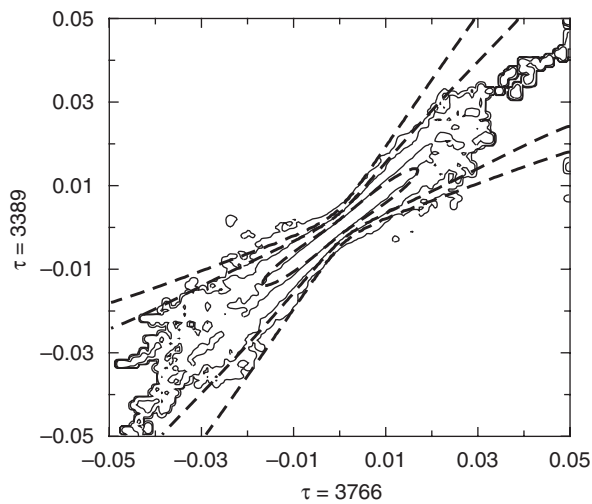


Fig. 5. Conditional probability density $p(y(\tau = 3389 \text{ s}) | y(\tau = 3766 \text{ s}))$ of given data (unbroken lines) and reconstructed by the numerical solution of the Fokker–Planck equation (broken lines).

5. Discussion

The results indicate that for financial data there are two scale regimes. In the small-scale regime the shape of the pdfs changes very fast and a measure like the Kullback–Leibler entropy increases linearly. At timescales of a few seconds not all available information may be included in the price and processes necessary for price formation take place. Nevertheless this regime seems to exhibit a well-defined structure, expressed by the very simple functional form of the Kullback–Leibler entropy with respect to the timescale τ . The upper boundary in timescale for this regime seems to be very similar for different stocks [13].

Based on a stochastic analysis we have shown that a second time range, the medium scale range exists, where multi-scale joint probability densities can be expressed by a stochastic cascade process. Here, the information on the comprehensive multi-scale statistics can be expressed by simple conditioned probability densities. This simplification may be seen in analogy to the thermodynamical description of a gas by means of statistical mechanics. The comprehensive statistical quantity for the gas is the joint n -particle probability density, which describes the location and the momentum of all the individual particles. One essential simplification for the kinetic gas theory is the single particle approximation. The Boltzmann equation is an equation for the time evolution of the probability density $p(\mathbf{x}, \mathbf{p}, t)$ in one-particle phase space, where \mathbf{x} and \mathbf{p} are position and momentum, respectively. In analogy to this we have obtained for the financial data a Fokker–Planck equation for the scale τ evolution of conditional probabilities, $p(y_i, \tau_i | y_{i+1}, \tau_{i+1})$. In our cascade picture the conditional probabilities cannot be reduced further to single probability densities, $p(y_i, \tau_i)$, without loss of information, as it is done for the kinetic gas theory.

As a last point, we would like to draw attention to the fact that based on the information obtained by the Fokker–Planck equation it is possible to generate artificial data sets. As pointed out in Ref. [18], the knowledge of conditional probabilities can be used to generate time series. One important point is that increments $y(\tau)$ with common right end points should be used. By the knowledge of the n -scale conditional probability density of all $y(\tau_i)$ the stochastically correct next point can be selected. We could show that time series for turbulent data generated by this procedure reproduce the conditional probability densities, as the central quantity for a comprehensive multi-scale characterisation.

Acknowledgements

For helpful discussion we want to thank R. Friedrich, Ch. Renner and D. Sornette.

References

- [1] E. Fama, *J. Bus.* 38 (1965) 34.
- [2] B. Mandelbrot, *J. Bus.* 36 (1963) 394.
- [3] P.K. Clark, *Econometrica* 41 (1973) 135.
- [4] R.N. Mantegna, H.E. Stanley, *Nature* 376 (1995) 46.
- [5] B. Castaing, Y. Gagne, E.J. Hopfinger, *Physica D* 46 (1990) 177.
- [6] T. Lux, M. Marchesi, *Nature* 397 (1999) 498.
- [7] J.P. Bouchaud, M. Potters, *Theory of Financial Risks*, Cambridge University Press, Cambridge, 2001.
- [8] J. Muzy, J. Delour, E. Bacry, *Eur. Phys. J. B* 537 (2000) 17.
- [9] S. Ghashghaie, W. Breymann, J. Peinke, P. Talkner, Y. Dodge, *Nature* 381 (1996) 767.
- [10] S. Siegert, *Entwicklung eines Verfahrens zum Schätzen deterministischer und stochastischer dynamischer Strukturen*, Shaker Verlag, 2001.
- [11] A.M. van Mourik, A. Daffertshofer, P.J. Beek, *Phys. Lett. A* 351 (2005) 13.
- [12] T. Lüdecke, Discussion Paper No. 190, University of Karlsruhe, 1998.
- [13] A.P. Nawroth, J. Peinke, *Eur. Phys. J. B* 50 (2006) 147.
- [14] S. Kullback, *Information Theory and Statistics*, Dover, New York, 1968.
- [15] V. Plerou, P. Gopikrishnan, L.A.N. Amaral, M. Meyer, H.E. Stanley, *Phys. Rev. E* 60 (1999) 6519.
- [16] R. Friedrich, J. Peinke, C. Renner, *Phys. Rev. Lett.* 84 (2000) 5224.
- [17] C. Renner, J. Peinke, R. Friedrich, *Physica A* 298 (2001) 499.
- [18] A.P. Nawroth, J. Peinke, *Phys. Lett. A* 360 (2006) 234.