Ann. Inst. Statist. Math.
Vol. 55, No. 2, 437–446 (2003)

STABILITY OF MARKOVIAN STRUCTURE OBSERVED IN HIGH FREQUENCY FOREIGN EXCHANGE DATA

MIEKO TANAKA-YAMAWAKI

Department of Computer Science and Systems Engineering, Miyazaki University, Miyazaki 889-2192, Japan, e-mail: mieko@cs.miyazaki-u.ac.jp

(Received June 3, 2002; revised January 20, 2003)

Abstract. Contrary to the common sense in economics and financial engineering, price fluctuations at very fine level of motion exhibit various evidences against the efficient market hypothesis. We attempt to investigate this issue by studying extensive amount of foreign currency exchange data for over five years at the finest level of resolution. We specifically focus on the proposed stability in binomial conditional probabilities originally found in much smaller examples of financial time series. In order to handle very large data, we have written an efficient program in C that automatically generates those conditional probabilities. It is found that the stability is maintained for extremely large time duration that covers almost the entire period. Based on the length of conditions for which the conditional probabilities are distinguishable each other, we identify the length of memory being less than 3 movements.

Key words and phrases: Markovian structure, memory length, conditional probability, high frequency data in finance, tick data, foreign exchange rates, prediction.

1. Introduction

Although efficient market hypothesis is a basic assumption in financial engineering that regards any competitive market to be a perfectly fair gamble, investors believe in the opportunities of taking advantage by using various means such as arbitrage chances, skewness in probability distributions, and various patterns, because those opportunities indeed exist in the real financial time series. Details are, however, not clear enough to explain why some empirical 'rules' work, and what kind of stochastic process they are, to what degree they are stationary and how to deal with them. Much discussion has been made as to how we can incorporate a deviation from the efficient market in order to improve the theory of finance (Fama (1991)).

Studying real data in the financial systems, especially high frequency data of price movements is of special importance in order to answer those questions. Mainly due to their enormous sizes, high-frequency data in finance have been difficult to access until recently.

An epoch-making publication appeared in 1995 in which Mantegna and Stanley (1995) demonstrated that the shape of probability density distribution of price increments of S&P 500 stock index per minute follows the Lévy's stable distribution of index 1.4, instead of Gaussian distribution. This implies that the price fluctuation is not a pure random walk but a walk with burst effects, often referred as the 'fat-tail'. This was the first in the literature to show that high resolution financial data indeed obey to the

scale invariant statistics, proposed by Mandelbrot (1963) much earlier. Many different analyses following this work have derived various values of indices and the results are not conclusive so far. Moreover the auto-correlations of the price as a function of time resolution vanish after a few seconds, while the auto-correlations of volatility stay non-zero for months, which remains as an open problem.

Ghashghaie *et al.* (1996) compared approximately 1.47 million data points U.S. Dollar–German Mark exchange rates recorded in HFDF (1993) with three dimensional fully-developed turbulence and showed that both the financial time series in HFDF (1993) and velocity differences of three dimensional turbulence follow Kolmogorov's scaling law of moments. They argued the resemblance of those two processes caused by the common cascade structure in the energy flow for turbulence and the information flow for the financial market. This issue is also an open problem.

Those high-resolution financial data are called as 'tick data', since every price movement (called tick) is recorded in them together with related information. From them we expect to obtain detailed information on the short-term behavior of financial data. Especially it is expected to approach the question of how the no arbitrage chance hypothesis is to be altered at such high level of time resolution. Although there is no doubt that arbitrage chance is scarcely observed at the time scale of human reaction, it may be possible to find such chances at the resolution much smaller than the relaxation time. To this end we view the financial fluctuation from the same viewpoint as the microscopic motions of molecules compared to the macroscopic thermodynamics of bulk materials.

2. Short-term patterns

It is normally assumed that the fine fluctuations of financial data (such as currency exchange rates, or stock prices) can be regarded as the Brownian motion. This is another way of representing the dogma asserting that a competitive market is efficient and is very close to a perfect gamble, since a chance of making clear profit is immediately wiped out by prompt actions of shrewd investors. Luis Bachelier (1900) is said to be the first who identified the mathematical structure of competitive prices as random walks, in his Ph. D. thesis in 1900, five years earlier than the publication of Albert Einstein's famous paper on the Brownian motion. This hypothesis has long been a backbone of financial technology, based on which famous Black-Sholes formula was derived, for example. In short, the financial fluctuation is Brownian in the macroscopic level but no one has studied what really happens in the microscopic level before examining the tick data.

Recently, Ohira *et al.* (2002) has shown that the naive myth of efficient market hypothesis needs to be reconsidered at tick level. They showed by comparing two sets of data, A and B, which are well separated in time, that many of conditional probabilities for up/down motion take almost the same numerical values in the two data.

Motivated by this observation, we investigated much wider sets of data and found that the stability persists for a very long term compared to the average decision time of most investors. Also we have roughly identified the upper limit of the memory length of foreign currency market.

3. Tick data

Prior to the First International Conference on High Frequency Data in Finance held in Europe in March 1995, the idea of distributing high frequency data to the academic

community was suggested at the planning session of the Conference held in August 1993. The test data HFDF (1993) distributed by Olsen & Associates was prepared under such circumstances.

Tick data are presented as a set of information on the time, the price, position of the price (bid, or ask), and other information. The bid price is the dealer's buying price on which we sell, and the ask price is the dealer's selling price on which we buy. Always the bid price is lower than the ask price at each time. We are bounded to sell low and buy high. Tables 1 and 2 are the typical examples of HFDF (1993) and CQG (2001), respectively.

Tick data are available for various kinds of price movements. The CQG (2001) is the largest data set available to us at this moment and we choose it to take advantage of statistics. Another reason that we adopt this data set is that the foreign exchange rates have an advantage of continuity compared to other financial data. Since the exchange market is open somewhere on the Earth, prices are recorded without interruption except weekends. This means we have continuous time series of length 30,000 to 40,000 on the average, although we use here the whole data from 1995 to 2001 as a continuous time series.

There is a disadvantage in foreign exchange rates. First, all the available data are quotes only and we must judge by ourselves how many of them were actually executed. Stocks such as NYSE-TAQ (1993) and futures are, on the other hand, usually equipped with information on the actually executed prices, though they have discontinuities at every closing time of the market and a large amount of accumulated orders rush into the market almost simultaneously at the opening time of the market.

Another possible problem is the diversity of information sources. Recorded data are gathered from all over the world and the price movements do not necessarily indicates the logical order of the price moves. Test data such as HFDF (1993) have detailed information on the country and the bank at which each price quote comes from. However, many tick data including CQG (2001) miss such information and the resulted data are

Table 1. Example of HFDF (1993) data of U.S. Dollar in terms of German Mark. The time, bid-price, ask-price, country/city code, bank code, filter (1 means a good data) are shown.

HFDF (1993) Data Set Description

CCYY-MM-DD (GMT)	bid	ask	country/city	bank	filter
1990-03-25 23:59:44	1.7190	1.7195	344 01	0056	1
1990-03-25 23:59:56	1.7185	1.7192	036 02	0065	1

http://www.olsendata.com/

Table 2. Example of CQG (2001) data of U.S. Dollar in terms of Japanese Yen. Contract number, date, session, time, bid (B) or ask (A) are listed.

Contract	Date	Session	Time	Price	(T)
JY1995U	19950731	0	1730	11413	B
JY1995U	19950731	1	0002	11376	B
JY1995U	19950731	1	0002	11394	A

http://www.cqg.com/

simply a time series of quotes from many different countries/markets. However, this does not immediately mean that such tick data miss all the information on the logical order of price movements, since many investors nowadays depend on the on-line quotes coming from computer terminals, which are essentially the same information as seen in the tick data.

Although many financial databases are gradually prepared for distribution, still most of those data sets are not easy to use for various reasons, such as recording errors, inconvenient formats, and high costs, etc.

Recording errors are the most serious problem among all. Especially until early 1990's when recordings were made by human hands, excess rates of jumping/missing digits are observed (Moriya (2002)). We use, as a standard data for cross-checking, HFDF (1993).

Excess of zero in price movements causes another problem. Probability distribution of price increments can be approximated by Gauss distribution or other stable distribution only when a part of zeros are taken off. This process is necessary not only for quotes data but also for actually traded data in order to approximate the probability distribution by a known distribution function. Therefore we need to fit a raw distribution by a stable distribution and subtract appropriate amount of zeros then rescale the distribution and fit it again. This process can be avoided in many cases by appropriately choosing the histogram width. A careful treatment is called for in the course of statistical analysis. We do not get into this problem in this paper any further because we focus on the tick-wise motions and dot not deal with the delicate question of the statistical property.

Finally we mention the problem of ask-bid spreads. Bid is a price at which the dealer wants to buy, and ask is the price to sell. Those are mostly paired at every tick. However, the total numbers of ask prices are usually different from the number of bid prices. It is said that bid prices are more reliable because ask prices are vulnerable to recording errors and other external conditions. We still need more study before drawing any conclusion about this.

4. Remarkable stability

We first examine the largest currency exchange tick data that we have, and then treat HFDF (1993) for the purpose of cross checking. We adopt the 'ask' quotes of U.S. Dollar vs. Japanese Yen from January 2, 1995 to April 12, 2001 having 10,127,289 data points (Hereafter we abbreviate this data set as CQG_UJA).

The process of our analysis is as follows. We first split the first 10 million data points into twenty sets, each of which contains 500,000 data points. We use {0,1} to represent {down, up} motion of the tick level price change and neglect data if there is no change from the previous one. We then compute all the 2^m conditional probabilities for memory depth m. We have written a program in C that automatically generates 2^m files containing each conditional probability for 20 data sets for a fixed memory depth m. In the case of $m = 1$, for example, two files are generated after scanning 20 data sets, one of which containing 20 values of $P(1 \mid 0)$ for 20 data sets and the other containing 20 values of $P(1 \mid 1)$. Figure 1(a) shows those two conditional probabilities for $m = 1$. Here $P(1 \mid 0)$ indicates the rate for the price to rise ($1 = \uparrow$) after fall ($0 = \downarrow$) and $P(1 \mid 1)$ indicates the rate for the price to rise ($1 = \uparrow$) after rise ($1 = \uparrow$). Note that other elements such as $P(0 \mid 0) = 1 - P(1 \mid 0)$, and $P(0 \mid 1) = 1 - P(1 \mid 1)$ are not independent and thus

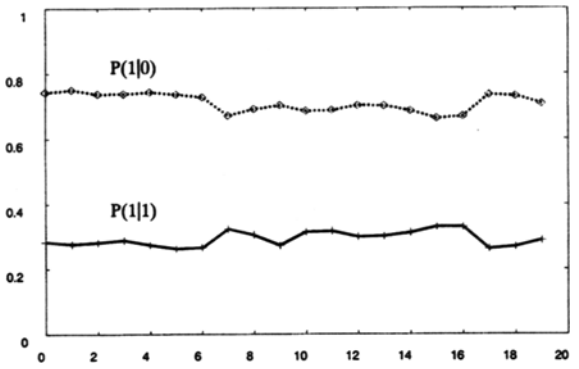

Fig. 1(a). Two conditional probabilities, $P(1 \mid 0)$ and $P(1 \mid 1)$ for memory depth $m = 1$ are drawn for the data explained in the text (CQG_UJA) that are cut into 20 samples of size 500,000.

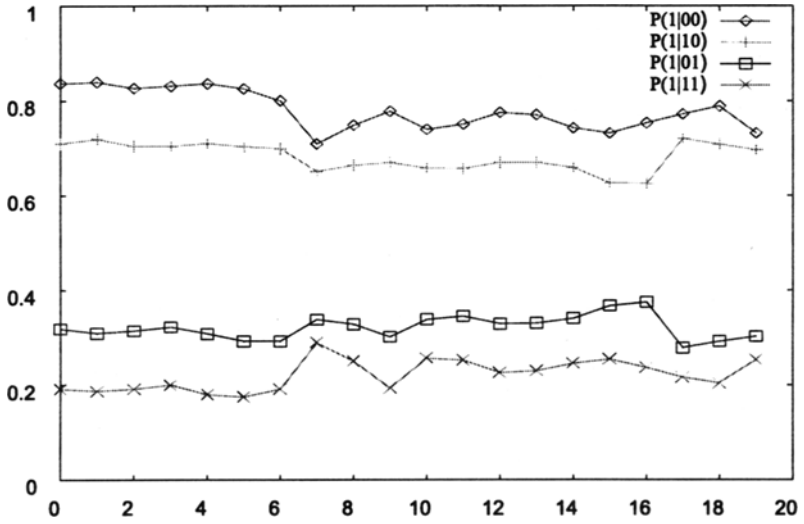

Fig. 1(b). Four conditional probabilities, $P(1 \mid 00)$, $P(1 \mid 10)$, $P(1 \mid 01)$, and $P(1 \mid 11)$ for the memory depth $m = 2$ are drawn according to the same method for the same data as in (a).

omitted in our discussion. Note that $P(1 \mid 0)$ and $P(1 \mid 1)$ are both remarkably stable for the entire period from 1995 to 2001.

Increasing the depth of memory is straightforward. For example there are eight conditional probabilities, $P(1 \mid 000), P(1 \mid 001), \ldots, P(1 \mid 111)$ for $m = 3$, which are plotted in Fig. 1(c). Note that those time series of conditional probabilities keep high level of stability throughout the entire time period except the seventh sample and the seventeenth, which presumably correspond to rapid rises of U.S. Dollars.

Another notable outcome of this study is on the information of memory length that the price fluctuations are supposed to have. We can read it out from Fig. 1(a) and Fig. 1(b) in the following manner. The lines are stable and well separated each other in (a) and (b) corresponding to $m = 1$ and 2, then at $m = 3$ shown in Fig. 1(c) lines begin

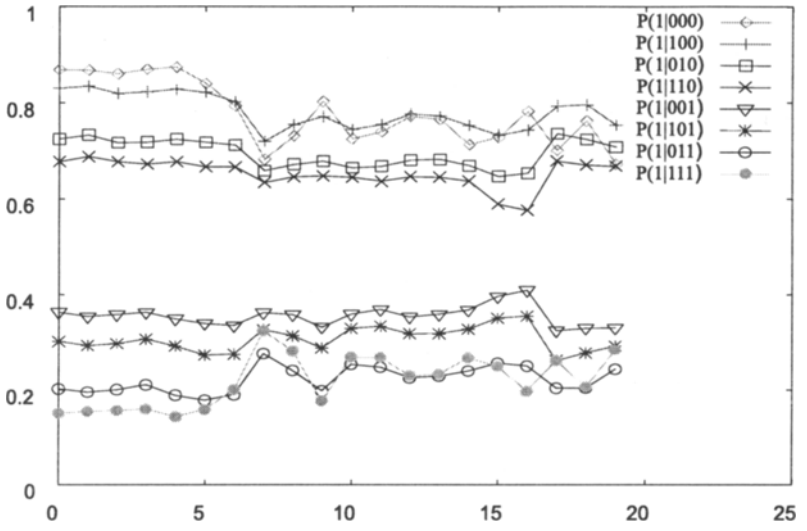

Fig. 1(c). Eight conditional probabilities for $m = 3$, $P(1 \mid 000)$, $P(1 \mid 100)$, $P(1 \mid 010)$, $P(1 \mid 110)$, $P(1 \mid 001)$, $P(1 \mid 101)$, $P(1 \mid 011)$, $P(1 \mid 111)$, from the top to the bottom, are drawn according to the same method for the same data as in (a) and (b).

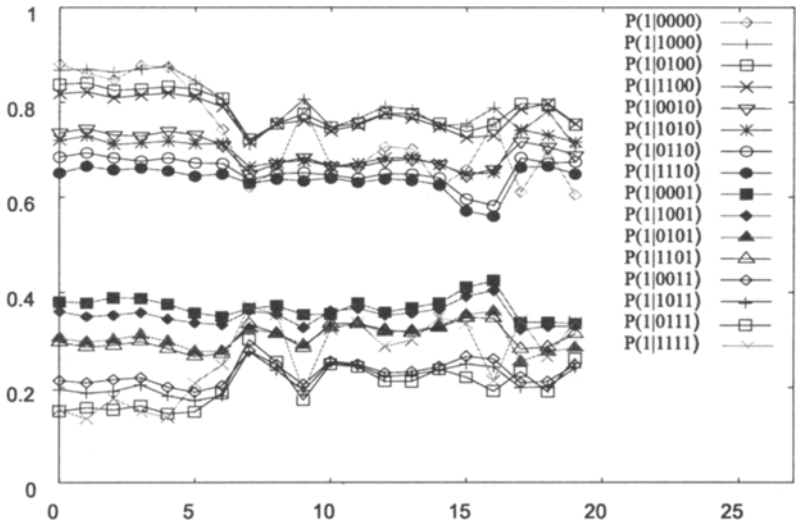

Fig. 1(d). Sixteen conditional probabilities for $m = 4$ are drawn according to the same method for the same data as in (a)–(c). The first set of four lines including $P(1 \mid 0000)$, $P(1 \mid 1000)$, $P(1 \mid 0100)$, $P(1 \mid 1100)$ are bunched into the upper most part of the figure. The next bunch below the first one consists of the four lines including $P(1 \mid 0010)$, $P(1 \mid 1010)$, $P(1 \mid 0110)$, and $P(1 \mid 1110)$. The third bunch is made by $P(1 \mid 0001)$, $P(1 \mid 1001)$, $P(1 \mid 0101)$, $P(1 \mid 1101)$, from the upper to the lower. $P(1 \mid 0011)$, $P(1 \mid 1011)$, $P(1 \mid 0111)$, $P(1 \mid 1111)$, which indicates that the memory length larger than three, does not provide an extra knowledge on the price movements.

pairing, which means the following identity.

$$P(1 \mid 000) = P(1 \mid 100), \qquad P(1 \mid 011) = P(1 \mid 111), \qquad \text{etc.}$$

Those equations imply that the condition of three consecutive fall of the price is equivalent to the two consecutive falls, and the three consecutive rises is equivalent to the two consecutive rises.

Figure 1(d) showing the case $m = 4$, in which many lines overlap each other. For example, the four lines corresponding to $P(1 \mid 0000)$, $P(1 \mid 0100)$, $P(1 \mid 1000)$, and $P(1 \mid 1100)$ almost overlap into a single bunch at the upper part of the figure, and another four lines of $P(1 \mid **10)$ making the second highest bunch, and so on. As a result, 16 lines in Fig. 1(d) appear to be the four thick lines due to the loss of information. Here we must point out that $P(1 \mid 0000)$ and $P(1 \mid 1111)$ largely deviate from this rule in appearance, due to the fact that the four consecutive moves to the same direction such as 0000 or 1111 are extremely rare events thus suffered from the low statistics.

In other words, the memory length of the price movement is found to be about two (or up to three) ticks roughly speaking. We have checked this fact in studying mutual information between the past movements and the next, which also show that the meaningful length of memory being about two ticks.

This fact indicates that only the two movements (i.e., actually moved ticks) in the past control the next movement of the price and the memory beyond two movements is irrelevant to determine the next movement. This situation reminds us various other examples of off-random time series, such as human-generated random series (Tanaka-Yamawaki (1999)) that we studies some time ago thus could be fitted by means of a hidden Markov model (Rabinier and Huang (1993)).

To this end, a question arises whether there is a difference in cutting the data into a different size of samples. The result for the case of 500 samples of size 20,000 is shown in Fig. 2, which is very similar to Fig. 1(a) in the outline.

In order to make the stability more quantitatively we compute mean and standard deviation for each time series of conditional probability and list the result in Table 3. Note that mean and standard deviation for both columns coincide very well.

We further examine the effect of downsizing the samples. In Fig. 3 we show the errors as a function of the size of samples. From this we can read out that the results are essentially the same as the previous cases if the sample size is not very small compared to 10,000.

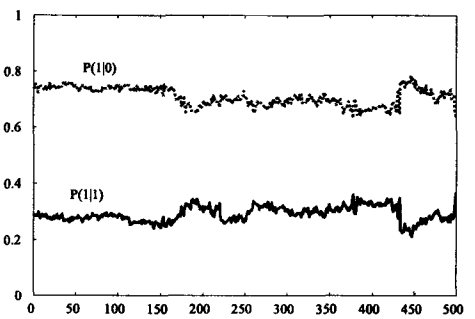

Fig. 2. Conditional probabilities calculated for the case of 500 samples of size 20,000, much finer than the case in Fig. 1, which has almost the same as Fig. 1(a) in the outline.

Table 3. Mean and standard deviation of conditional probabilities of binomial motions in the foreign exchange data for various lengths of memory (up to four ticks) are listed at two different ways of sampling. The data used in this table are the first ten million data points of the ask position of Japanese Yen vs. U.S. Dollar in CQG (2001). The left is the case where the sampling size is 20,000, while the right column is 500,000. Both coincide very well.

Memory	20,000 points * 500 (Fig. 2)			500,000 points * 20 (Fig. 1)		
depth	Conditional prob.	Mean	SD	Conditional prob.	Mean	SD
1	$P(1 \mid 0)$	0.709454	0.031540	$P(1 \mid 0)$	0.709468	0.027599
	$P(1 \mid 1)$	0.292142	0.026769	$P(1 \mid 1)$	0.292127	0.022062
2	$P(1 \mid 00)$	0.780571	0.047476	$P(1 \mid 00)$	0.779413	0.040372
	$P(1 \mid 01)$	0.321634	0.028868	$P(1 \mid 01)$	0.321320	0.024599
	$P(1 \mid 10)$	0.680840	0.032513	$P(1 \mid 10)$	0.681142	0.029041
	$P(1 \mid 11)$	0.220092	0.039745	$P(1 \mid 11)$	0.221170	0.031636
3	$P(1 \mid 000)$	0.783128	0.074298	$P(1 \mid 000)$	0.777754	0.065382
	$P(1 \mid 001)$	0.354890	0.026408	$P(1 \mid 001)$	0.355078	0.020703
	$P(1 \mid 010)$	0.694159	0.032194	$P(1 \mid 010)$	0.694682	0.028650
	$P(1 \mid 011)$	0.220839	0.034717	$P(1 \mid 011)$	0.221135	0.027110
	$P(1 \mid 100)$	0.781533	0.041745	$P(1 \mid 100)$	0.781253	0.035437
	$P(1 \mid 101)$	0.306406	0.029698	$P(1 \mid 101)$	0.305832	0.025239
	$P(1 \mid 110)$	0.652805	0.033508	$P(1 \mid 110)$	0.652568	0.028362
	$P(1 \mid 111)$	0.213118	0.064068	$P(1 \mid 111)$	0.218003	0.053657
4	$P(1 \mid 0000)$	0.748125	0.110421	$P(1 \mid 0000)$	0.732846	0.094989
	$P(1 \mid 0001)$	0.368033	0.035580	$P(1 \mid 0001)$	0.368679	0.022827
	$P(1 \mid 0010)$	0.695044	0.037562	$P(1 \mid 0010)$	0.694816	0.032082
	$P(1 \mid 0011)$	0.228630	0.033577	$P(1 \mid 0011)$	0.229240	0.024018
	$P(1 \mid 0100)$	0.784996	0.043715	$P(1 \mid 0100)$	0.784954	0.036416
	$P(1 \mid 0101)$	0.307030	0.030511	$P(1 \mid 0101)$	0.306176	0.026225
	$P(1 \mid 0110)$	0.656575	0.033938	$P(1 \mid 0110)$	0.656526	0.028420
	$P(1 \mid 0111)$	0.201415	0.054516	$P(1 \mid 0111)$	0.203979	0.043949
	$P(1 \mid 1000)$	0.796713	0.062471	$P(1 \mid 1000)$	0.794030	0.053197
	$P(1 \mid 1001)$	0.351561	0.026444	$P(1 \mid 1001)$	0.351566	0.020340
	$P(1 \mid 1010)$	0.693953	0.031783	$P(1 \mid 1010)$	0.694742	0.028230
	$P(1 \mid 1011)$	0.216675	0.037123	$P(1 \mid 1011)$	0.216797	0.028588
	$P(1 \mid 1100)$	0.775189	0.040976	$P(1 \mid 1100)$	0.774522	0.033464
	$P(1 \mid 1101)$	0.305087	0.032059	$P(1 \mid 1101)$	0.305270	0.024924
	$P(0 \mid 1111)$	0.638904	0.038500	$P(1 \mid 1110)$	0.638109	0.027176
	$P(1 \mid 1111)$	0.246158	0.097646	$P(1 \mid 1111)$	0.259902	0.080375

Finally a comment is in order. At first look, the two lines in Fig. 1(a) appear to be dependent since the sum of the two at each moment seems to be one. This is caused by the fact that the total number of up motions and that of down motions are approximately the same in the time series we have used here. For the same reason Fig. 1(b), Fig. 1(c), and Fig. 1(d) are approximately symmetric against the horizontal line of height 1/2. However we consider them independent quantities because the slight difference between the upper half and the lower half of the figures carries important

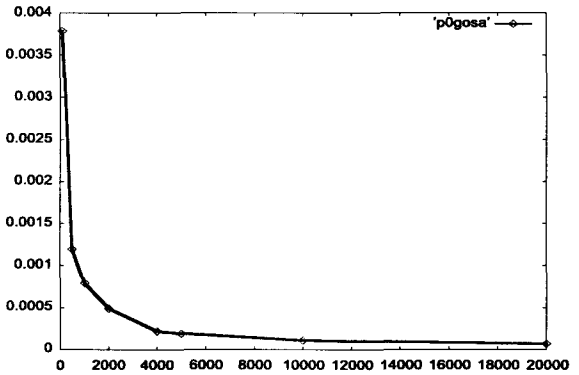

Fig. 3. Standard deviations of $P(1 \mid 0)$ as a function of sample sizes. The minimum size required is about 5,000–10,000.

information characterizing the data set.

5. Concluding remarks

We have studied tick data of currency exchange rate that contains over 10 million 'ask' quotes of U.S. Dollar vs. Japanese Yen from January 2, 1995 to April 2001.

First we wrote a program that extracts the conditional probabilities of binary motions by assigning the 'up' motion to 1 and 'down' motion to 0, and neglecting unmoved motions and automatically generates 2^m files corresponding to the conditional probabilities of memory length m.

We found that those conditional probabilities are very stable over years, and the effective length of memory is two, roughly speaking from the separation of independent conditional probabilities. We have also checked that this process is independent of the size of data sampling, by showing that the mean and standard deviation of the time series for two sizes, 20,000 and 500,000. Also the minimum sample size required for this analysis is 5,000–10,000, from the magnitude of errors at each size.

Acknowledgements

We thank Professor Tamura (The Institute of Statistical Mathematics) for sharing the data set HFDF93 used in this paper.

References

Bachelier, L. (1900). Théorie de la speculation, Doctor Thesis, *Annales Scientifiques de l'Ecole Normale Sperieure*, **III-17**, 21–86 (Translation: Cootner, P. H. (ed) (1964). The Random character of stock market prices, 17–18, MIT Press, Cambridge, Massachusetts).

CQG (2001). A set of tick data foreign currency exchange rates of Japanese Yen vs. U.S. Dollar, taken from CQG-terminal from 1995 to 2001 by H. Moriya. CQG is a chart-graphics distribution company (http://www.cqg.com/).

Fama, E. F. (1991). Efficient capital market II, *Journal of Finance*, **46**, 1575–1617, and the references therein.

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

HFDF (1993). A set of tick data from October 1, 1992 to September 30, 1993, obtained from Olsen & Associates at http://www.olsendata.com/

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

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

Moriya, H. (2002). Tick data filtering and cleaning, Talk given at the First ISM/GUAR Economics Meeting, Institute of Mathematical Statistics, July 18–19, 2002: http://www.cs.miyazaki-u.ac.jp/joho2/mieko/econo_e/ (to appear in SOKENDAI publication, 2003).

NYSE-TAQ (1993). A collection of tick data of all the trades and quotes of stock prices at the NYSE market since 1993, which are available from New York Stock Exchange at http://www.nyse.com/

Ohira, T. *et al.* (2002). Predictability of currency market exchange, *Physica A*, **308**, 368–374.

Rabinier, L. and Huang, B.-H. (1993). *Fundamentals of Speech Recognition*, Prentice Hall; More general discussions on Markov Models are found, e.g., in Hideki Imai (1986). *Information Theory*, Shokodo, Tokyo.

Tanaka-Yamawaki, M. (1999). Human generated random numbers and a model of the human brain functions, *Proceedings of 1999 IEEE International Conference on Systems, Man, & Cybernetics*, **III**, 223–228.