Eur. Phys. J. B **34**, 495–500 (2003)
DOI: 10.1140/epjb/e2003-00249-5

THE EUROPEAN PHYSICAL JOURNAL B

Spot foreign exchange market and time series

F. Petroni[a] and M. Serva

Dipartimento di Matematica and INFM Università dell'Aquila, 67010 L'Aquila, Italy

Received 4 March 2003 / Received in final form 12 June 2003
Published online 9 September 2003 –

Abstract. We investigate high frequency price dynamics in foreign exchange market using data from Reuters information system (the dataset has been provided to us by Olsen and Associates). In our analysis we show that a naïve approach to the definition of price (for example using the spot mid price) may lead to wrong conclusions on price behavior as for example the presence of short term correlations for returns. For this purpose we introduce an algorithm which only uses the non arbitrage principle to estimate real prices from the spot ones. The new definition leads to returns which are not affected by spurious correlations. Furthermore, any apparent information (defined by using Shannon entropy) contained in the data disappears.

PACS. 89.65.Gh Economics; econophysics, financial markets, business and management – 65.40.Gr Entropy and other thermodynamical quantities

1 Introduction

A foreign exchange market is an over the counter (OTC) market not subject to any time restriction, in fact, it is open 24 hours a day 7 days a week. Given also that it is the most liquid market in the world and the availability of tick-by-tick quotes, foreign exchange market is very convenient for the study of high frequency behaviors.

Foreign exchange market is made up of about 2000 financial institutions around the globe which operates by selling or buying certain amount of a given currency. A market maker (any of the financial institutions which make the market) is expected to quote simultaneously for its customers both a bid and a ask price at which it is willing to sell and buy a standard amount of a given currency. Each of the major market makers shows a running list of its main bid and ask quotes, and those quotes are displayed to all market participants. In principle each quote from each market maker is valid until a new quote is displayed by the same market maker. In practice, this is not the case and no information is given about the lifetime of each quote. Many author report a fighting screen effect for advertising purposes [1–3]: to maintain their name on the screen some market maker keep sending fake quotes.

In analyzing recorded financial data [4,5], a difficult and puzzling problem is to define which is the real asset price [1,3,6]. In principle, three different quotes for the asset are available: bid, ask and transaction price (the price at which the transaction is actually made). Using a wrong definition for asset price can lead to wrong evaluation of price dynamics. For example, if the transaction price is used to analyze price dynamics a random zero mean oscillation around the real price will be found at very short time scale and this would generate artificial autocorrelations (bid/ask spread effect) [7,8].

We analyze the DEM/USD exchange quotes taken from Reuters' EFX pages (the dataset has been provided to us by Olsen and Associates) during a period of one year from January to December 1998. In this period 1,620,843 quotes entries in the EFX system were recorded. The dataset provides a continuously updated sequence of bid and ask exchange quotation pairs from individual institutions whose names and locations are also recorded. EFX dataset does not contain any information on traded volume and on the lifetime of quotes. Furthermore EFX quotes are indicative and they do not imply that any amount of currency has been actually traded.

The aim of this work is to find the best definition for the asset price. We start analyzing raw data assuming that the asset price is simply given by spot mid quotes. We find that this leads to an indeterminacy of asset price at very short time scale and to spurious correlations for returns. We investigate one possible explanation assuming that spot quotes contain an estimation error made by the market maker on the real price. In this way we do not find the real price but then we introduce an algorithm which, reducing the spread between bid and ask quotes, is able to determine the real price and solve the indeterminacy. We use information theory and a moving average based analysis to strengthen our results. The key of our work is that we are able to determine the real price with a parameter free algorithm which uses only the non-arbitrage principle.

[a] e-mail: filippo_petroni@virgilio.it

2 A naïve approach to the study of FX microstructure

The aim of this section is to show that a naïve approach to the analysis of foreign exchange market may lead to wrong conclusions on price dynamics.

We analyze data taken from EFX Reuters' information system of DEM/USD exchange quotes of the entire year 1998. In the dataset each pair of bid and ask quotes as given by the market operators is recorded. The dataset does not contain information on transaction prices or on volumes of currencies traded but only tick-by-tick exchange quotes. Notice that, nowadays, transaction prices for FX market are becoming more and more available.

Prices are irregularly time-spaced and we decided, instead of sampling the data in some arbitrarily fixed sampling time, to use business time (basically a tick time) as our time flow index. In the calendar time framework, prices are modelled as random processes evolving in ordinary time. Clearly, prices in the markets are not fixed at every t, but only at discrete intervals. Nevertheless, according to the calendar time picture, prices are usually considered as discrete samples of an underlying process. In the business-time approach, price dynamics is directly modelled as a discrete-time random process. Indeed, the time basis is the ordered sequence of times at which prices are quoted in the markets. Although in this work we are not forced to make a precise choice we prefer to use business-time approach because it is simpler to handle and because there are strong indications of being more fundamental (see [9] and Refs. therein). According to our choice t takes all integer values up to N which is the number of quotes in the dataset.

We indicate with $S_t^{(b)}$ and $S_t^{(a)}$ respectively bid and ask quotes at time t. For our analysis we consider mid price as given by the geometric average of bid and ask quotes $S_t = \sqrt{S_t^{(a)} \times S_t^{(b)}}$ [2]. We stress that this choice for the mid price is not stringent, the same results can be obtained if bid or ask quotes are used [10].

We define return at two consecutive business time as:

$$r_t \equiv \ln \frac{S_{t+1}}{S_t} \tag{1}$$

and, in general, returns at time t and lag τ as

$$r_t(\tau) \equiv \ln \frac{S_{t+\tau}}{S_t}. \tag{2}$$

We estimated using the above cited dataset the τ dependent variance of returns:

$$\left\langle r_t^2(\tau) \right\rangle, \tag{3}$$

the non-overlapping first order covariance of two consecutive returns after s lags

$$\langle r_{t+s}(\tau) r_t(\tau) \rangle \tag{4}$$

and the non-overlapping higher order covariances of returns

$$\langle r_{t+s+\alpha}(\tau) r_t(\tau) \rangle \tag{5}$$

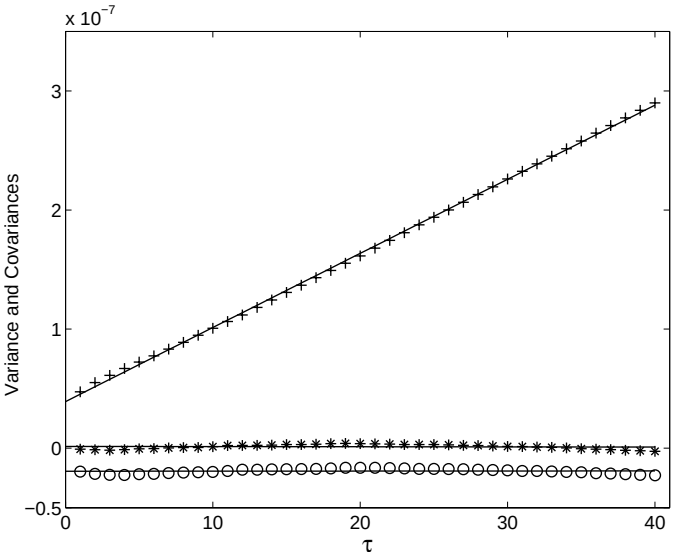

Fig. 1. DEM/USD spot exchange rates: variance (3) (crosses) compared with a linear fit $2A + B\tau$, non-overlapping first order covariance (4) (circles) compared with -A, non-overlapping higher order covariance (5) (stars) compared with zero. A and B are identified with $\langle \epsilon_t^2 \rangle$ and $\langle \tilde{r}_t^2 \rangle$.

where $\alpha \geq 1$. In the three definitions $\langle \cdot \rangle$ indicates an average over the probability distribution. Results are shown in Figure 1. The variance of returns is a linear function of time lags s, as expected, but it is different from zero in the limit $s \to 0$. This implies the existence of an implicit indeterminacy in the price estimation for vanishing time lags. The same indeterminacy is responsible for the negative covariance of two consecutive returns (see below) [1].

In order to explain the previous facts, it has been suggested [6,11] that the mid price is the composition of two different stochastic processes: a real price change and a noise contribution which is the result of microstructure frictions and imperfections (we do not enter in details of the different terms generating the error contribution: for a better explanation see [1,2]).

Given that S_t is the mid price at business time t we can express the two contributions as:

$$S_t = \tilde{S}_t e^{\epsilon_t} \tag{6}$$

where $\tilde{S}_t$ is the real price and ϵ_t is the error contribution to the real price ($\epsilon_t \equiv \ln(S_t/\tilde{S}_t)$, $\tilde{r}_t = \ln(\tilde{S}_{t+\tau}/\tilde{S}_t)$). The relation between returns is then given by:

$$r_t = \tilde{r}_t - \epsilon_t + \epsilon_{t+1}. \tag{7}$$

In this framework we can explain the behavior of the variance and of the other quantities reported in Figure 1. In fact, with the above definitions, the τ dependent variance can be calculated analytically:

$$\left\langle r_t^2(\tau) \right\rangle = 2 \left\langle \epsilon_t^2 \right\rangle + \left\langle \tilde{r}_t^2 \right\rangle \tau. \tag{8}$$

Where it has been assumed that ϵ_t and $\tilde{r}_t$ are uncorrelated random variables. The non-overlapping first order covariance of two consecutive returns after s business time

$$\langle r_{t+s}(\tau) r_t(\tau) \rangle = - \left\langle \epsilon_t^2 \right\rangle \tag{9}$$

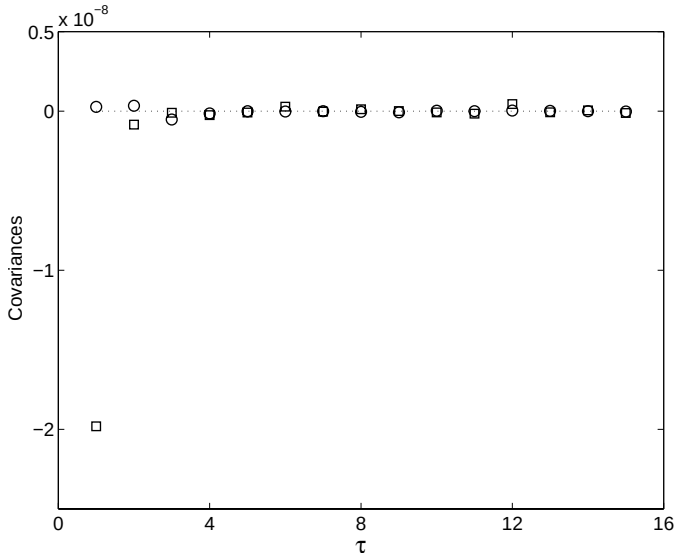

Fig. 2. Covariance $\langle r_t r_{t+s}\rangle$ and $\langle \tilde{r}_t \tilde{r}_{t+s}\rangle$ for spot (squares) and real (circles) returns.

and the non-overlapping higher order covariances of returns

$$\langle r_{t+s+\alpha}(\tau) r_t(\tau)\rangle = 0. \tag{10}$$

The above picture corresponds exactly to what one can see in Figure 1. Therefore, it can be estimated the experimental value for $\langle \epsilon_t^2\rangle$ which is $(2.0 \pm 0.2) \times 10^{-8}$ and $\langle \tilde{r}_t^2\rangle = (0.64 \pm 0.05) \times 10^{-8}$ for the particular dataset analyzed. We stress that equations (8) and (9) give two independent estimation of the variance allocated in the error contribution. We find that the two values, computed from data of Figure 1, coincide within errors.

In order to complete our picture we also estimated the covariance function on time intervals s, defined as

$$\langle r_{t+s} r_t\rangle \tag{11}$$

where we considered $\langle r_t\rangle = 0$. Results are plotted in Figure 2. The figure shows that the spot returns are one step negatively correlated ($\langle r_{t+1} r_t\rangle = -\langle \epsilon_t^2\rangle$) while for $s > 1$ we have $\langle r_{t+s} r_t\rangle \simeq 0$ (with respect to the first term) according to previous findings [1–3,10,12].

We stress that the correlation we find here is not due to the bid-ask spread [7,8] given that in our analysis we are not using transaction prices but mid quotes. Even the discreteness [13,14] of prices cannot be invoked to explain the increase of volatility and this because in our dataset price changes are restricted to less than 1×10^{-4} of the actual price and this effect, if exist, is very small [1,3,10].

3 A more realistic approach

Different methods already exit in the literature which try to estimate the real underlying price from the mid quote [1,2,10]. Those methods are based on a trade-matching algorithm [10], on the assumption that the mid quote return is an MA(1) stochastic process [1] or on the assumption that each market maker quote has a mean life time (of about 2 minutes) before elapsing. The aim of this work is to find a possible algorithm which is able to separate the two contributions in the mid price without any assumption on the nature of the mid quote return process and without fixing any arbitrary parameter.

This algorithm should be able to solve the indeterminacy found when the mid price is used to analyze high frequency price dynamics. From the previous paragraph we have constraints on the variance allocated in the real price and in the error distribution, the algorithm should then take this constraints into account.

In DEM/USD 1998 dataset, each quote at each business time is associated with the financial institution which fixed that quote. In principle this quote should be valid until the same bank gives a different exchange quote (both for bid and ask prices). In practice between two different quotes from the same bank there are several quotes fixed by other institutions around the world. This suggests that a bank quote elapses after a certain time even if a new quote has not been fixed by the same bank. If the dataset contained information on the time duration of each quote, or the life time of each quote would be a know constant, there would be no problem in establishing real price at each time: it would be the best bid and ask quotes valid at that time. But this information is not available and a different strategy has to be found to establish real price at each time.

The algorithm we propose is the following: let us suppose that we are observing the bid and ask price of a given currency and that we are able to detect each quotes from all the financial institution in the business time t.

We define the spread between bid and ask as: $D_t = S_t^{(a)} - S_t^{(b)}$. Notice that for the non-arbitrage principle this quantity is greater than or equal to zero. Considering k time lags previous to business time t we consider the following effective spread:

$$D_{t,k} = \tilde{S}_i^{(a)} - \tilde{S}_i^{(b)} \tag{12}$$

where $\tilde{S}_i^{(a)} = \min_{i\in\{t-k,t\}} S_i^{(a)}$ and $\tilde{S}_i^{(b)} = \max_{i\in\{t-k,t\}} S_i^{(b)}$. For each t our algorithm find $\tilde{k}$ which gives $D_{t,\tilde{k}} \geq 0$ and $D_{t,\tilde{k}+1} < 0$. The real price is then given by

$$\tilde{S}_t = \sqrt{\tilde{S}_i^{(a)} \times \tilde{S}_i^{(b)}}. \tag{13}$$

In this way we can then define one currency quote at each time. Notice that the number of steps the algorithm has to go backwards in time is only given by the non-arbitrage principle and it is different for every t. We stress that we do not know the lifetime of each quote and that the non-arbitrage principle is needed to fix an upper bound on this life time. It is, in fact, very unlikely that a given quote is still valid even if it produces arbitrage (except for a very small number of data) and this is fully supported by the results presented in the following.

To compare our algorithm with the one proposed in [2] we compute the average number of steps needed to find the real price: we find that, on average, about 6 ticks are

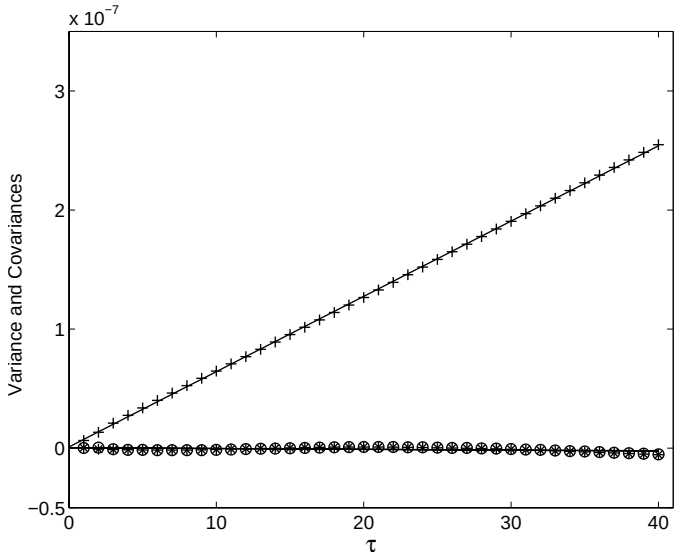

Fig. 3. DEM/USD real exchange rates: variance (3) (crosses) compared with a linear fit $B\tau$, non-overlapping first order covariance (4) (circles) compared with zero, non-overlapping higher order covariance (5) (stars) compared with zero, B is identified with $\langle \tilde{r}_t^2 \rangle$.

needed and given that, in the calendar time approach, one tick in our dataset corresponds on average to 20 seconds also our algorithm finds an average life time of quotes of about 2 minutes.

Once we have obtained $\tilde{S}_t$ we can define $\tilde{r}_t(\tau) = \ln(\tilde{S}_{t+\tau}/\tilde{S}_t)$ and compute all quantities (variances and correlations) already computed for the naïve price definition.

As stated above if our algorithm is correct we should have that the indeterminacy contained in the mid price is removed for the real price. We then replicate the analysis described in the first paragraph for the mid price using the above defined real price $\tilde{S}_t$. Results for this analysis are presented in Figure 3. It can be seen that the variance of returns goes to zero when business time goes to zero, in fact the experimental value of $\langle \epsilon^2 \rangle$ in equation (8) for the real price is $(0.03 \pm 0.01) \times 10^{-8}$, two order of magnitude smaller than for the mid price. Also the first order non-overlapping covariance of two consecutive returns goes to zero. Another interesting results is that we obtain for the real returns variance a value ($\langle \tilde{r}_t^2 \rangle = 0.64 \times 10^{-8}$) which is identical, within error, to the one predicted in equation (8).

If we estimate the covariance of returns as defined in equation (11), we obtain that the real price returns are uncorrelated at every step (see Fig. 2 where covariance is compared with that of 'naïve returns' given in Eq. (11)).

We notice that while our algorithm produces returns which are uncorrelated the trade-matching algorithm proposed in [10] finds a positive first order correlation of the same magnitude of the negative first order correlation present in the observed returns. We want also to stress that while the results from our algorithm and the procedure proposed in [1] are more or less comparable (they both produce uncorrelated returns), in our approach we do not need to assume any particular stochastic model for returns and we do not need to estimate any parameter from the past quotes.

Indeed, the idea we have used here is very simple, we assume that old quotes are still valid until they produce arbitrage. In spite of the simplicity we are able to remove all artifacts in the data without introducing any free parameter.

4 Information analysis

To be able to perform information analysis on our dataset first of all we need to code the original data in a sequence of symbols [15]. There are several way to build up such a sequence: one should make sure that this treatment does not change too much the structure of the process underlying the evolution of financial data. A partition process of the range of variability of the data is needed in order to assign a conventional symbol to each element of the partition. A symbol corresponds then unambiguously to each element of the partition. The procedure described below permits to code financial data in a sequence of binary symbols from which is then possible to quantify available information.

We fix a resolution value Δ and define

$$r_{t_i}(\tau) \equiv \ln \frac{S_{t_i+\tau}}{S_{t_i}} \tag{14}$$

where t_i is a given business time. We wait until an exit time τ_i such as

$$|r_{t_i}(\tau_i)| \geq \Delta. \tag{15}$$

In this way we only consider market fluctuations of amplitude Δ. We can build up a sequence of $r_{t_i}(\tau_i)$, where $t_1 = t_0 + \tau_0$ and $t_{i+1} = t_i + \tau_i$, then we code this sequence in a binary code according to the following rules:

$$c_k = \begin{cases} -1 \; if \; r_{t_i}(\tau_i) < 0 \\ \\ +1 \; if \; r_{t_i}(\tau_i) > 0 \end{cases}. \tag{16}$$

The procedure described above corresponds to a patient investor who waits to update his investing strategy until a certain behavior of the market is achieved, for example, a fluctuation of size Δ.

Once we have build a symbolic sequence we can estimate the entropy which is defined, for a generic sequence of n symbols, as:

$$H_n = -\sum_{C_n} p(C_n) \ln p(C_n) \tag{17}$$

where $C_n = \{c_1 \dots c_n\}$ is a sequence of n objects and $p(C_n)$ its probability. The difference

$$h_n \equiv H_{n+1} - H_n \tag{18}$$

represents the average information needed to specify the symbol c_{n+1} given the previous knowledge of the sequence $\{c_1 \dots c_n\}$.

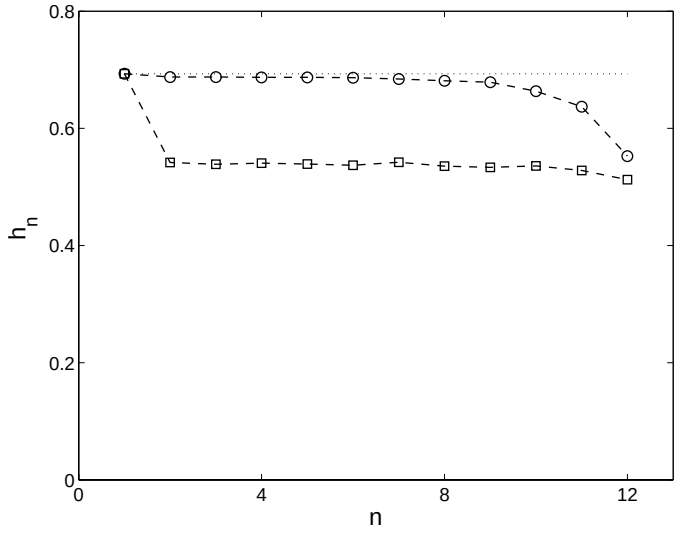

Fig. 4. Information for spot (squares) and real (circles) prices.

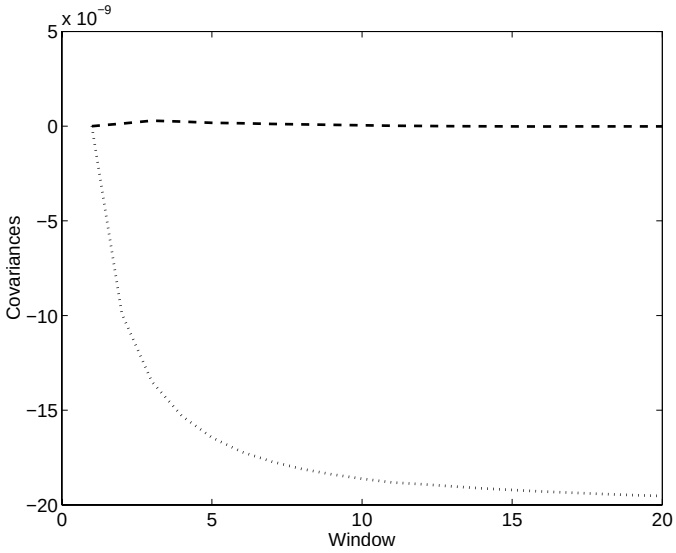

Fig. 5. Covariances between returns r_t and $\phi_t(\tau)$, defined in equation (21), as a function of lags window τ for spot (dots-line) and real (dashed line) prices.

The series h_n is monotonically not increasing and for an ergodic process one has

$$h = \lim_{n\to\infty} h_n \tag{19}$$

where h is the Shannon entropy [16]. It can be shown that if the stochastic process $\{c_1 \dots c_n\}$ is Markovian of order k (*i.e.* the probability of having c_n at time n depends only on previous k steps $n-1, n-2, \dots, n-k$), then $h_n = h$ for $n \geq k$. In other cases either h_n goes to zero for increasing n, which means that for n sufficiently large the $(n+1)$th-symbol is predictable knowing the sequence C_n, or it tends to a positive finite value. The maximum value of h is $\ln(2)$ for a dichotomic sequence. It occurs if the process has no memory at all and the 2 symbols have the same probability. The difference between $\ln(2)$ and h is intuitively the quantity of information we may use to predict the next result of the phenomenon we observe, *i.e.* the market behavior.

In Figure 4 h_n is estimated both for real ($\tilde{S}_t$) and mid prices (S_t). From the results it is obvious the different behaviors of the two definition for currency price. In fact, while for the mid price we find a non zero available information ($\ln 2 - h_n \neq 0$), the stochastic process is a Markov process of order 1, the real price does not show this behavior. The available information for the real price is zero and it remains zero at every step (due to the finite number of data we can only estimate h_n until $n \simeq 9$ but we can extrapolate its behavior for $n \to \infty$). This show that the real price (unfortunately) is a stochastic process with no memory and predictability at least at the frequency at which the analysis has been performed.

5 Moving average and price forecasting

As seen in the previous section, the Shannon information, which seemed to be available in the naïve approach, vanishes when the dataset is pre-processed using our algorithm. The information carried by the naïve data could be used to predict future prices and gain money from this. We will use a quantity, the moving average, which is one of the most popular analysis tools widely used both in financial literature and from market traders [17–19], to show that this prediction power is only apparent and vanishes when our algorithm is used to define real prices.

The lag dependent moving average is defined as:

$$S_t(\tau) = \frac{1}{\tau}\sum_{i=0}^{\tau-1} S_{t-i} \tag{20}$$

where S_t can be the real or the mid price. The lags window τ can be varied in order to consider the influence of prices far back in time. Once the moving average is estimated we can define the following quantity

$$\phi_t(\tau) = \ln \frac{S_t}{S_t(\tau)} \tag{21}$$

which measures the relative position of the quote S_t with respect to its moving average. One would like to know if the knowledge of $\phi_t(\tau)$ can give some information about the successive evolution of prices, *i.e.* if this knowledge is useful to make a better probabilistic prediction about next returns. For this reason we compute the covariance between this quantity and returns

$$\chi(\tau) = \langle r_t \phi_t(\tau) \rangle \tag{22}$$

where again $\langle\cdot\rangle$ is an average over the entire dataset. Results for real and mid price are presented in Figure 5. Once again it is evident that while the real price defined here satisfies the efficient market hypothesis this does not hold true for the mid price which is negatively correlated with the moving average. This negative correlation between returns and moving average allows forecast of the

next coming price. In fact the evolution of mid price is such to reduce the distance between the mid price and the average [20].

6 Conclusions

The aim of this work is to find the exact way to extract real prices from quotes taken form Reuters' Information system. Our dataset contains 1,620,843 bid and ask DEM/USD quotes recorded during the entire year 1998, from the 1st of January until the 31st of December 1998.

In Section 2 we review the results that show that a wrong behavior of price dynamics can be obtained when the raw dataset is naïvely processed. In fact, one finds an implicit indeterminacy in price specification which increases the volatility and produces spurious covariances. Following [6] we then explain this indeterminacy by means of an error contribution which is responsible for the increased volatility and for the covariances.

At this point we introduce a parameter free algorithm, only based on the non arbitrage principle, which is able to extract the real prices from the spot ones. The correctness of the procedure is corroborated by the many results presented in this work. First of all we show that with the new price definition the indeterminacy and the one step anti-correlation drop to zero. We also show, through information analysis, that the stochastic process for the new defined price has no short range memory.

Given our results we think that when studying price dynamics a strong attention has to be posed on the definition of prices to be used in the analysis in order to avoid wrong conclusions as, for example, the existence of short term return correlations.

We stress that we are able to define real prices directly from spot quotes without the need of further information (like for example time of validity of quotes [2] or the estimation of parameters from past quotes [1]).

In conclusion we would like to propose our method as a general tool to process raw high frequency dataset in order to obtain a new dataset of the same length whose data are a better representation of price evolution in the very short time scale.

We thank Michele Pasquini for illuminating discussions in the early stage of the present work and for continuous interest and suggestions. F.P. acknowledges the financial support of Cofin MIUR 2002 prot. 2002027798_005.

References

1. F. Corsi, G. Zumbach, U.A. Müller, M. Dacorogna, Economic Notes **30**, 183 (2001)
2. M.M. Dacorogna, R. Genay, U.A. Müller, R.B. Olsen, O.V. Pictet, *An introduction to high-frequency finance* (Academic Press, 2001)
3. B. Zhou, Journal of Business & Economic Statistics **14**, 45 (1996)
4. S. Taylor, *Modeling financial time series* (John Wiley & Sons, New York, 1986)
5. P.J. Brockwell, R.A. Davis, *Time series: theory and methods* (Springer-Verlag, New York, 1991)
6. J. Hasbrouck, Rev. Financial Studies **6**, 191 (1993)
7. M. Blume, S. Stambaugh, J. Financial Economics **12**, 387 (1983)
8. R. Roll, J. Finance **39**, 1127 (1984)
9. L. Berardi, M. Serva, Int. J. Appl. Finance (submitted)
10. T. Bollerslev, I. Domowitz, J. Finance **48**, 1421 (1993)
11. M. Pasquini, M. Serva, Physica A **277**, 228 (2000)
12. F. De Jong, R. Mahieu, P. Schotman, J. Int. Money, Finance **17**, 5 (1998)
13. G. Glottlieb, A. Kalay, J. Finance **40**, 135 (1985)
14. L. Harris, J. Financial, Quantitative Analysis **25**, 291 (1990)
15. R. Baviera, M. Pasquini, M. Serva, D. Vergni, A. Vulpiani, Eur. Phys. J. B **20**, 473 (2001)
16. C.E. Shannon, Bell Syst. Techn. J. **27**, 623 (1948)
17. M.J. Pring, *Technical Analysis Explained* (McGraw-Hill, 1985)
18. T.A. Meyers, *The Technical Analysis Course* (Probus Publishing, 1989)
19. D. Meyers, Futures **29**, 60 (2000)
20. R. Baviera, M. Pasquini, J. Raboanary, M. Serva, Int. J. Theor. Appl. Finance **6**, 575 (2002)