Consistent High-Precision Volatility from High-Frequency Data

Fulvio Corsi [1]
Gilles Zumbach [1]
Ulrich Müller [1]
Michel Dacorogna [2]

January 25, 2001

FCO.2000-09-25

Abstract

Estimates of daily volatility are investigated. Realized volatility can be computed from returns observed over time intervals of different sizes. For simple statistical reasons, volatility estimators based on high-frequency returns have been proposed, but such estimators are found to be strongly biased as compared to volatilities of daily returns. This bias originates from microstructure effects in the price formation. For foreign exchange, the relevant microstructure effect is the incoherent price formation, which leads to a strong negative first-order autocorrelation $\rho(1) \simeq -40\%$ for tick-by-tick returns and to the volatility bias. On the basis of a simple theoretical model for foreign exchange data, the incoherent term can be filtered away from the tick-by-tick price series. With filtered prices, the daily volatility can be estimated using the information contained in high-frequency data, providing a high-precision measure of volatility at any time interval.

[1] Olsen & Associates
Research Institute for Applied Economics
Seefeldstrasse 233, 8008 Zürich, Switzerland.
e-mail: fulvio@olsen.ch, gilles@olsen.ch, ulrichm@olsen.ch
phone: +41-1/386 48 48 Fax: +41-1/422 22 82

[2] Zurich Re
General Guisan - Quai 26, 8022 Zürich, Switzerland.
e-mail: michel.dacorogna@zurichre.com

1 Introduction

Volatility is an essential ingredient for many applied issues in finance and financial engineering, for example in asset pricing, asset allocation or in risk management. In risk assessment, due to the increasing role played by the Value-at-Risk (VaR) approach, it is becoming more important to have a good measure and forecast of short-term volatility, mainly at the one to ten day horizon. Intuitively, the volatility measures how much a random process jitters, and several definitions along this simple idea can be written down. But despite its importance, volatility is still an ambiguous term for which there is no unique, universally accepted definition. Currently, the main approaches to compute volatilities are by historical indicators computed from daily squared (or absolute) returns, by econometric models such as GARCH, by the standard RiskMetrics definition (equivalent to IGARCH with $\nu = 0.94$), or by an indirect computation from option prices and a pricing model such as Black-Scholes. All of those approaches suffer from noticeable drawbacks, especially for short-term volatility. The main drawback is to be fairly insensitive to the recent intraday behavior as they are based on daily data (except for the implied volatility which is an indirect definition). Clearly, a simple solution is to use a direct definition of the volatility, conforming to our intuitive understanding of volatility as outlined above, and to use intraday data.

Several authors went along this line, for example [Taylor and Xu, 1997], [Dacorogna et al., 1998] or [Andersen et al., 1999], where the daily volatility is measured by a sum of short-term intraday squared returns, say for example at a 10 minutes time horizon. This approach is called *realized volatility*, and promises in theory, as claimed by [Andersen et al., 1999], to reach an error-free estimation of the volatility. In practice however, we found that for return intervals less than a few hours, such a definition is affected by a considerable systematic error: the expectation of the volatility computed with high-frequency returns is not equal to the one obtained with daily returns. Yet, the goal is to measure the size of the typical one-day price move, and therefore to compare the volatility definitions based on high-frequency data with the variance of daily price changes.

The purpose of this paper is to present some alternative definitions of realized volatility that make use of very high-frequency data, without suffering from the aforementioned systematic deviation. An empirical study of the lagged correlation of tick-by-tick returns leads to this goal. In Section 2, the motivations and definition of realized volatility are reviewed and its main shortcomings introduced. Section 3 contains the results of the scaling analysis of realized volatility showing its anomalous scaling behavior. In Section 4, the autocorrelation function of high-frequency FX and stock index returns computed with different time scales is analyzed. Section 5 discusses the influence of the multiple contributor structure of the FX spot market on the price formation process, and a simple model which accounts for the empirical findings is presented. Section 6 introduces the EMA price filtering and an unbiased version of the realized volatility estimator.

2 The realized volatility

In the standard framework, volatility is computed by summing squared returns of an artificial regular time series of logarithmic prices $x_{\mathrm{RTS}}(t)$. The usual definition for the annualized (realized) volatility over a time interval T is:

$$\sigma[T,\delta t](t) \quad = \left\{ \frac{1}{n} \sum_{t-T+\delta t \le t' \le t} r^2[\delta t](t') \right\}^{1/2} \tag{1}$$

where the return r has an expectation close to 0 and is defined as

$$r[\delta t](t) = \frac{x_{\mathrm{RTS}}(t) - x_{\mathrm{RTS}}(t-\delta t)}{\sqrt{\delta t / T_{\mathrm{ref}}}} \quad (2)$$

$$n = \sum_{t-T+\delta t \leq t' \leq t} \quad (3)$$

where

- x_{RTS}: a Regular Time Series (RTS) of (logarithmic middle) price. This quantity needs to be computed with some interpolation procedure from the inhomogeneous high-frequency tick-by-tick price time series $x(j)$.
- $r[\delta t]$: the annualized return, observed over time intervals of size δt.
- T_{ref}: the one year normalization period.
- T: the length of the moving window over which the volatility is computed.
- $\sigma[T, \delta t]$: the annualized volatility computed from $r[\delta t](t)$ over the period T.
- n: the number of return observations in the interval T. In the normal case of non-overlapping return intervals: $n = T/\delta t$.

The notations of [Zumbach and Müller, 2000] are used.

It should be emphasized that any definition of realized volatility involves two time parameters: T, and δt, and in order to have a statistically reliable measure of volatility, the parameters must be such that $T \gg \delta t$. For a Gaussian random walk, the order of magnitude of the statistical error of the volatility estimator can be computed. For this model, the sum of square returns has a chi-square distribution with degree of freedoms equal to the number of terms in the sum $\sigma^2[T, \delta t] \sim \chi^2(T/\delta t) = \chi^2(n)$. The root mean square error (RMSE) for a χ^2 distribution is given by $\mathrm{RMSE}(\sigma^2) = \sigma^2\sqrt{2n-1}/n$. For a Gaussian random walk, typical values for the measurement error of the daily volatility measured with return at different time interval are:

$$\begin{aligned} \mathrm{RMSE}(\sigma^2[1\mathrm{d}, \quad 1\mathrm{d}]) &= 100\%\ \sigma^2 \\ \mathrm{RMSE}(\sigma^2[1\mathrm{d}, \quad 1\mathrm{h}]) &= 28.5\%\ \sigma^2 \\ \mathrm{RMSE}\ (\sigma^2[1\mathrm{d}, 1\mathrm{min}]) &= \ \ 3.7\%\ \sigma^2 \end{aligned}$$

These numbers clearly show the advantage of taking returns at small time intervals for measuring the daily volatility. For $T \simeq \delta t$, problems arise since essentially only one observation is used, for example when defining daily volatilities with daily data $\sigma[1\mathrm{d}, 1\mathrm{d}]$. Besides, one advantage of the realized volatility estimator is to be model free, although it is still "definition dependent" as parameters must be chosen in equation 1 (for example, absolute return can be taken instead of squared return).

The idea of using high-frequency data in the computation of volatility traces back to the seminal intuition of [Merton, 1980] according to which higher frequencies are not useful for the mean but essential for the variance. This is deeply rooted in that, for a random walk, a minimal exhaustive statistics for the mean is given by the start and end point of the walk, whereas a minimal exhaustive statistics for the volatility is essentially given by the full set of increments. Yet only recently this idea has been exploited with intraday data: [Taylor and Xu, 1997] and [Andersen et al., 1999] rely on 5-minute returns in the measurement of daily exchange rate volatilities, [Schwert, 1998] utilizes

15-minute returns to estimate daily stock market volatilities, while [Dacorogna et al., 1998] use 1-hour returns to compute a 1-day volatility benchmark for other volatility forecasting models.

A formal justification for taking infinitesimal δt is given in [Andersen et al., 1999]. These authors use a model for the logarithmic price process, which is a standard continuous-time diffusion process with time varying volatility $\sigma_{\text{model}}(t)$ and zero expected drift rate. The name *realized* volatility comes from this context, as the empirical volatility estimate is computed on one *realized* price path of the underlying model. On the other hand, for the empirical data, there is no underlying model and probability space, and we only have the unique realization of the price history. In the theoretical model, the realized variance is a consistent estimator of the 1-day integrated variance (i.e. the one obtained by integrating the instantaneous variance $\sigma^2_{\text{model}}(t)$ over one day). Under those assumptions, the stochastic error of the measure could be arbitrarily reduced by simply increasing the sampling frequency of returns. This convergence property of the volatility is very appealing, in theory. An error-free estimation of volatility would allow us to treat realized volatility as an observable, rather than a latent variable as with a GARCH(1,1) model for example. This opens the possibility to directly analyze, model, forecast and optimize volatility itself. More sophisticated dynamic models can be directly estimated without having to rely on the complicated estimation procedures needed when the volatility is assumed to be unobserved (e.g. log-likelihood for the ARCH-type models). For forecasting purposes, a better estimate of the target function allows to better extract the real underlying signal and to improve forecasting performance.

3 Scaling analysis of the realized volatility

Unfortunately, the empirical data are not that simple. Because of market microstructures effects, the assumption that log asset prices evolve as a diffusion process becomes less realistic as the time scale reduces. At the tick time scale, the empirical data differ from the simple theoretical model, and the volatility computed with very short time intervals is no longer an unbiased and consistent estimator of the daily volatility computed with daily returns. This effect can be analyzed by studying $\mathrm{E}[r^2[\delta t]]$ as a function of δt, because $\mathrm{E}[\sigma^2[T,\delta t]] = \mathrm{E}[r^2[\delta t]]$ (up to finite sample effects that are negligible). The definition 2 of the return $r[\delta t]$ is already annualized, and therefore, for an uncorrelated diffusion process, $\mathrm{E}[r^2[\delta t]]$ is independent of δt. In order to ease the comparison between assets, the empirical average have been normalized such that $\mathrm{E}[r^2[1\mathrm{d}]] = 1$. Figure 1 shows the scaling behavior of $\mathrm{E}[r^2[\delta t]]$ for two currencies and two stock indices. The horizontal line at 1 corresponds to the expected volatility of an i.i.d diffusion process. Clearly, the expectation of the volatility computed with return taken at small time intervals is not equal to the volatility obtained with daily returns:

$$\mathrm{E}[r^2[\delta t]] \neq \mathrm{E}[r^2[1\mathrm{d}]] \qquad \text{and} \qquad \mathrm{E}[\sigma^2[T,\delta t]] \neq \mathrm{E}[\sigma^2[T,1\mathrm{d}]]$$

The volatility estimator computed at a short time interval is strongly *biased* as compared to the mean squared daily return. We use the term bias because for a majority of the agents operating on financial markets, the relevant variable of interest is the daily return or the "one-day risk". Because of widely available daily data, the *de facto* time horizon of reference is one day. These agents are neither interested in the detailed price behavior happening at very short time intervals nor in the volatility observed at a 5-minute time horizon. The use of high-frequency returns to compute the daily volatility is merely a measurement issue: short-term returns are used because we want to improve the estimation of volatility, rather than being interested in risks existing in the extremely short time frame. Therefore, tolerating this bias would lead to distorted daily risk measures when using high-frequency data instead of daily data.

The size of this bias is directly measured by the vertical distance between the i.i.d. horizontal line

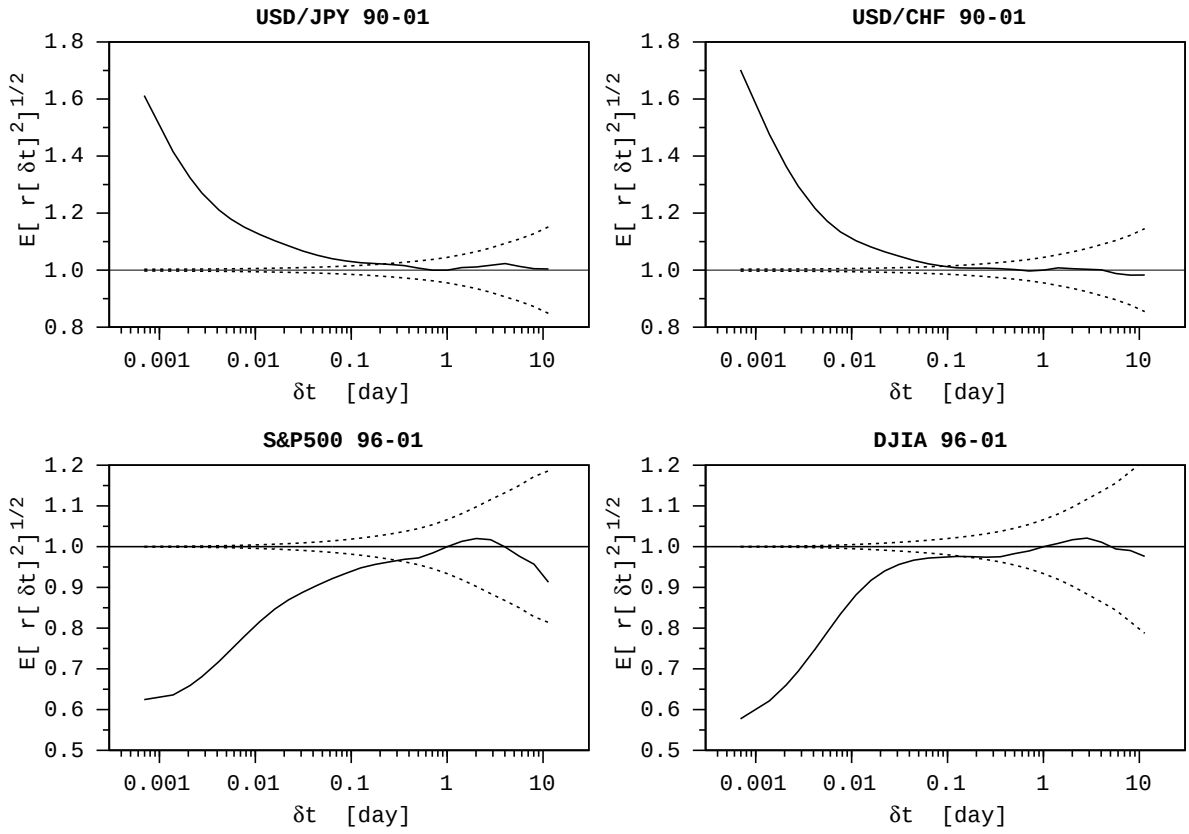

Figure 1: Scaling of the annualized volatilities $\mathrm{E}[r^2[\delta t]]$ for δt ranging from 1 minute to 1 week. The dotted lines are error bounds assuming independant returns. Two FX rates, USD/JPY and USD/CHF, and two stock indices, Standard & Poors 500 and Dow Jones Industrial Average, are investigated.

and the empirical volatility in Figure 1. From a study of many assets, the general behavior of the bias can be summarized as follow:

- **FX:** positive bias. At the 1-minute level, it ranges from 30% to about 80%. We also found that the bias tends to be higher for exchange rates with lower liquidity (such as USD/ITL which has a bias of more than 80% at 3-minute level) than that between major currencies which exhibit higher liquidity.
- **Stock indices:** small or strong negative bias.

Therefore, we are left with the following quandary. On one hand, statistical considerations would impose a very high number of return observations to reduce the stochastic error of the measurement, on the other hand, market microstructure comes into play, introducing a bias that grows as the sampling frequency increases. Because of this bias, there is a trade-off between two opposite types of limitations which precludes the possibility to have a precise and easy measure of the volatility (even in presence of an infinite number of ticks).

Given the trade-off between measurement error and bias previously described, the choice of the underlying return frequency becomes a critical issue if no specific treatment of the bias is made. In a recent manuscript, [Andersen et al., 2000] propose a direct graphical inspection of the scaling law behavior of the realized volatility (which they call "volatility signature plot") as a guidance for the choice of the underlying sampling frequency of returns. The idea is simply to choose, for each financial instruments, the shortest return interval at which the resulting volatility is still not

substantially affected by the bias. They found a time interval δt of only 30-20 minutes for highly liquid exchange rates, and a negative bias for the less liquid ones. We argue that those results are distorted by the linear interpolation used to convert the original inhomogeneous tick-by-tick time series of prices into an homogeneous time series spaced by 5 minutes. When two subsequent ticks are separated by more than 5 minutes, the linear interpolation implicitly assumes a minimum volatility in this interval, and introduces an artificial correlation between the subsequent returns of the generated regular time series. This leads to a systematic underestimation of volatility, that becomes larger as the tick frequency decreases and the number of empty 5-minute intervals without ticks increases. This could explain the negative bias reported by [Andersen et al., 2000] for the less liquid rates, whereas we always found a positive bias for FX rates. In the case of exchange rates, the model developed in Section 5 below can exhibit only positive bias.

For these reasons, when generating a synthetic regular time series for volatility estimation, it is more appropriate to use the "previous-tick interpolation" scheme, in which each tick remains valid until a new tick arrives. This interpolation scheme does not generate an underestimation of volatilities, nor a distortion of the autocorrelation. Using previous-tick interpolation, we found significant evidences of large biases at 30-20-minute level, even for major exchange rates. In our analysis, the return interval at which the bias is no longer significant occurs only at the level of some hours. This means that even for the most liquid assets, the shortest return interval to obtain an unbiased volatility measure is of the order of 2-3 hours, leading to only 8-12 observations per day. Furthermore, this "unbiased return interval" considerably changes from asset to asset. Therefore, without any specific treatment of the bias, the stochastic error of the volatility measure cannot be essentially reduced and cannot be computed with the same return interval for all instruments. Hence, if we want to have a general and homogeneous volatility estimation of optimal precision, an explicit treatment of the bias is required. The first step in this direction is to understand the origin of the bias, which is the subject of the next two sections.

4 Autocorrelation analysis

The anomalous scaling of the second moment can only be explained by a non-zero autocorrelation $\rho(k)$ of the returns i.e. $\mathrm{E}[\sigma^2[T,\delta t]] \neq \mathrm{E}[\sigma^2[T,1\mathrm{d}]]$ if and only if $\rho(k) \neq 0$ for some lags k. In short, the return must be not i.i.d. at a short time scale. This can be derived by the following computations. From the return at high frequency δt, the return at a larger time scale $\delta t' = m\,\delta t$ can simply be computed by aggregation (with an overall multiplication by a factor $1/\sqrt{m}$ to take care of the annualization). Then the variances computed with returns at scale δt and $\delta t'$ can be related. After some algebra, we find

$$\begin{aligned} \sigma^2[T, m\,\delta t] &= \sigma^2[T,\delta t]\frac{1}{m}\sum_{k=1}^{m-1}\sum_{l=1}^{m-1}\rho(k-l) \\ &= \sigma^2[T,\delta t]\left[1+2\sum_{k=1}^{m-1}(1-\frac{k}{m})\rho(k)\right] \end{aligned} \tag{4}$$

where $\rho(k)$ is the autocorrelation function of the return $r[\delta t]$ at lag $k\,\delta t$. We can use this formula to compute the scaling of the mean volatility given the autocorrelation function of the returns. Qualitatively, we can have one of the three following cases:

- no autocorrelation $\rho = 0$ implies $\sigma^2[T,\delta t] = \sigma^2[T,1\mathrm{d}]$, i.e. no anomalous scaling
- negative autocorrelation $\rho < 0$ implies $\sigma^2[T,\delta t] > \sigma^2[T,1\mathrm{d}]$, i.e. positive anomalous scaling

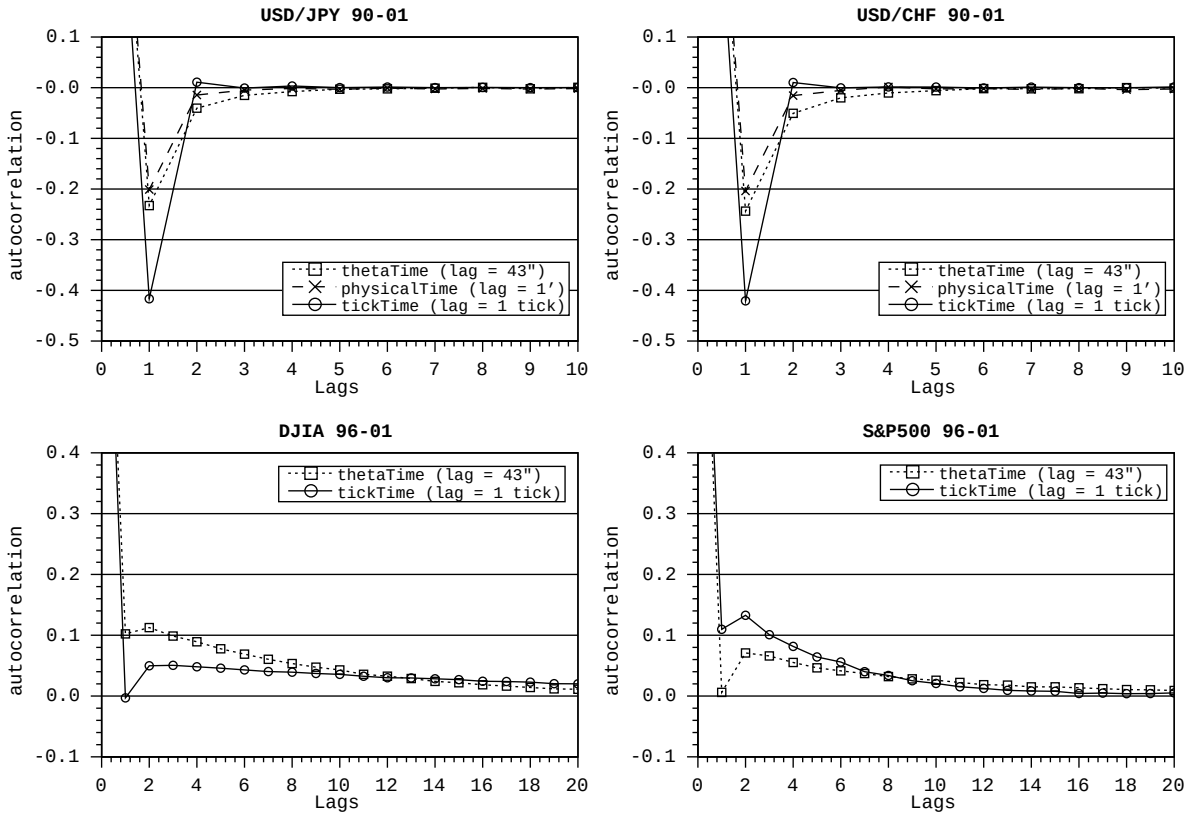

Figure 2: Returns autocorrelations for FX (top panels) and stock ind ices (bottom ones).

- positive autocorrelation; $\rho > 0$ implies $\sigma^2[T, \delta t] < \sigma^2[T, 1\mathrm{d}]$, i.e. negative anomalous scaling

We then expect a positive autocorrelation for stock indices and a negative one for FX. To confirm those expectations, we performed an intraday autocorrelation study of the returns for several stock indices and FX returns. For this computation, synthetic regular time series of price are generated, where "regular" is meant according to different time scales:

- In physical time scale. This is the usual physical time (including weekend).
- In tick time scale. This clock moves by one unit for every incoming tick. When computing the autocorrelation of tick returns on this scale, we ignore the varying time intervals between ticks.
- In dynamic ϑ-time scale. The ϑ-time scale is a sophisticated business time scale designed to remove intraday and intraweek seasonalities by compressing periods of inactivity while expanding periods of higher activity [Dacorogna et al., 1993, Breymann et al., 2000]. It is essentially an intraday generalization of the usual daily business time scale that omits weekends and holidays.

The results are plotted in Figure 2. The most evident result is the very strongly negative first-lag autocorrelation for FX computed in tick time, with a value around -40%. This contrasts with some mildly negative first-order autocorrelations reported in older studies on high-frequency FX data [Goodhart, 1989, Goodhart and Figliuoli, 1991]. Let us emphasize that the bid-ask bounce "a la Roll" cannot be invoked, since logarithmic middle prices are used in these computations. The bid-ask bouncing as described in Roll (1984) is due to the random hitting of the transactions at

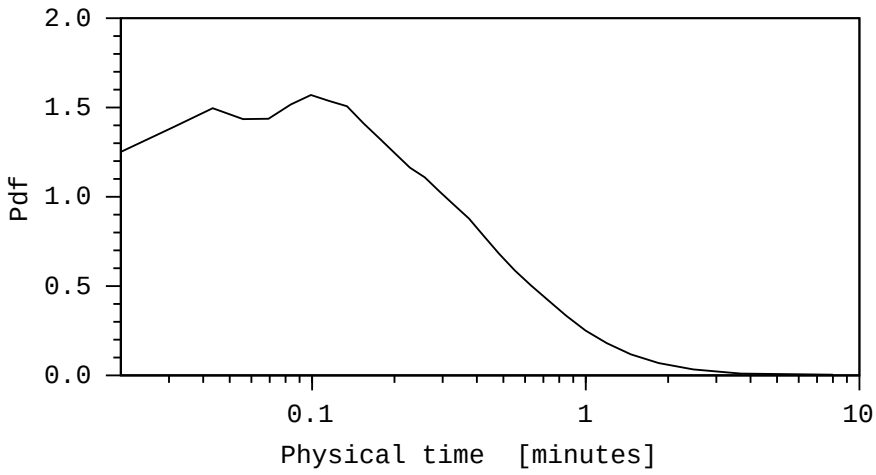

Figure 3: Probability distribution function (pdf) of the time intervals between ticks (semi-logarithmic scale) for USD/CHF. The data sample ranges from 1.1.1990 to 1.1.2000.

the bid or ask quotes, for a fixed bid and ask prices. Roll's explanation of the short-term bouncing thus exclusively applies to time series of transaction prices and not to those of middle prices as used in the present computations.

Compared to the autocorrelation evaluated in physical time and ϑ-time, the one computed in tick time presents a strongly negative value at the first lag, about twice as large. The decay at subsequent lags is faster. These differences between the autocorrelation in tick time and in the other time scales can be understood by considering the time deformation induced by the transformation of the tick time scale into a physical or ϑ-time scale. For example, Figure 3 shows the probability distribution function (pdf) of the time intervals between ticks measured in physical time for 10 years of USD/CHF. Computing the autocorrelation function of, say, one-minute returns in physical time would imply that:

- If many ticks are in the same one-minute interval, most of them are ignored in the computation of one-minute returns, and the returns are determined more by the true process and less by microstructural effects.
- If the time interval between two ticks is (much) larger than one minute, the previous-tick interpolation leads to zero returns in one or more one-minute intervals. Zero returns dampen lagged covariances, but the overall expectation of squared returns is not affected by a change of time scale. Since the autocorrelation is the quotient of lagged covariance and variance of returns, the result is a reduced autocorrelation at lag one. Beside, two returns separated by a string of empty one minute intervals share the incoherent component (see the model in the next section). In this case, the strongly negative first-lag autocorrelation in tick time directly affects two distant one-minute intervals, leading to the slow decay at larger lags of the correlation.

All these effects contribute to the empirically found behavior, notably the attenuated autocorrelation at the first lag.

The existence of a very pronounced mean-reverting behavior of the quotes is also revealed by a closer visual inspection of the price dynamics as displayed in Figure 4. Thus, the considerable anomalous scaling of FX volatility is entirely due to the strong negative correlation occurring between subsequent returns.

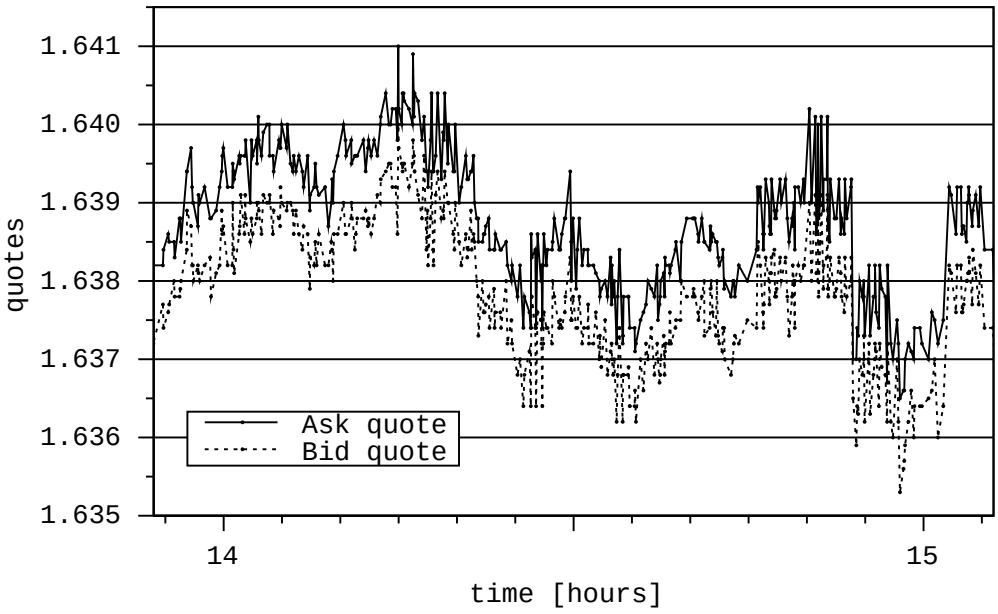

Figure 4: A sample price dynamics for the USD/CHF exchange rate. The data are from Reuters, the 06.04.2000, with GMT time.

On the contrary, for stock indices, we find a large significant positive autocorrelation that lasts up to few hours[3], as already reported by many other authors [Hawawini, 1980, Conrad and Kaul, 1988, Lo and MacKinlay, 1988]. This positive lagged correlation, pervasive across the three different time scales, sample periods and countries, can be explained by the so-called *lagged adjustment model* [Holden and Subrahamanyam, 1992, M. Brennan and Swaminathan, 1993]. According to this model, among the stocks that compose the indices, there are "leading" stocks that react quickly to new information whereas others stocks partially adjust or adjust with a certain delay, due to either information transmission, non-trading or lower volume. Since the lagged covariance of a portfolio is a weighted average of the lagged cross-covariance between the stocks composing it, a positive autocorrelation results.

Therefore, from the statistical point of view, the autocorrelation of returns is not negligible at short time scale (usually less than an hour), causing the volatility scaling to strongly deviate from that of a standard i.i.d. process. A relatively long lasting persistence of stock index returns gives rise to a positive slope of the scaling function (i.e. a negative bias) whereas a strong mean-reverting behavior between ticks causes a positive anomalous scaling in FX.

5 A model for the price process

So far, no economic explanation has been presented to account for those statistical properties. In order to develop a microscopic model for the FX quotes, the microstructure of the market must be taken into account as the return autocorrelation is directly related to microstructure effects arising from the price formation process. We claim that the strong mean-reverting behavior of the FX quotes results from the presence of an "incoherence" in the price formation. This effect comes from the fact that at the tick level, a "true price" cannot be exactly defined, but only a probability distribution of possible price exists. The main source of uncertainty that prevents the FX price from being a well defined concept originates from the multiple contributor structure of the FX spot

[3]While no positive autocorrelation is found in stocks returns themselves or in futures contracts on indices [Ahn et al., 1999].

market. A large part of the FX market is in fact an over-the-counter market where all dealers publish their own price quotes. Some consequences of the multiple contributor structure are:

- Disagreement on what the "true price" should be, due to the fact that opinions on public information, strategies of individual contributors, and private information sets are not uniform.
- Market maker bias towards bid or ask prices. Depending on their inventory position, market makers have preferences for either selling or buying. Hence to attract traders to deal in the desired direction they publish new quotes so that either the bid or the ask is competitive. The other price of the bid-ask pair, being pushed far away, influences the level of the (logarithmic) middle price, inducing a very short-term random bouncing of it.
- Fighting-screen effects for advertising purposes. In order to maintain their name on the data suppliers' screens (such as those of Reuters, Bloomberg or Bridge), some contributors keep publishing fake quotes generated by computer programs that randomly modify the most recent quotes (or a moving average of them).
- Delayed quotes. Trader interviews, comparisons to transaction data from electronic trading systems and lead-lag correlation studies show that many contributors release quotes with a considerable time delay (in some cases larger than a minute).

While individual agents follow their different strategies in a coherent way, the whole market generates an incoherent component due to the price formation process that is responsible for the negative autocorrelation of FX returns. Note again that the random hitting of bid or ask prices by transactions [Roll, 1984] is not considered here because it does not affect the logarithmic middle price.

A simple model that accommodates for the statistical and economic evidences can be build. According to the properties of the return autocorrelation functions computed with different time scales (cf. Section 4), the model is defined in tick time. A Brownian motion in tick time is a subordinate stochastic process which has been shown [Clark, 1973] to easily accommodate many empirical regularities such as heteroskedasticity, volatility clustering, fat tails and others. As in [Hasbrouck, 1993] the process in tick time of the observed price can be simply written as the sum of two terms: the subordinate process of the true price and the incoherent component modeled as an additive noise:

$$x(j) = \tilde{x}(j) + u(j) \tag{5}$$

where

- $j = j(t)$: the tick time scale (which represents the directing process of the subordinate Brownian motion).
- $x(j)$: the observed price at tick time j.
- $\tilde{x}(j)$: the unobserved "true" price.
- $u(j) \sim$ i.i.d$(0,\ \eta^2)$: represents the incoherent term modeled as an i.i.d. noisy component. No assumption on the distributional form of u is made.

The simplest model for $\tilde{x}(j)$ is a standard Brownian motion in tick time $B(j)$. Then the "true" tick return over a period of k ticks ($\tilde{r}_k$) normalized for convenience to 1 unit of tick time (i.e. $T_{\text{ref}} = 1$

tick) is:

$$\begin{aligned} \tilde{r}_k(j) &= \frac{\tilde{x}(j) - \tilde{x}(j-k)}{\sqrt{k}} \\ &= \frac{B(j) - B(j-k)}{\sqrt{k}} = \sigma\epsilon(j) \sim N(0,\ \sigma^2) \end{aligned} \tag{6}$$

where $\epsilon(j)$ is a standard random variable $N(0,1)$. Given this model, the observed k-tick return is

$$r_k(j) = \tilde{r}_k(j) + \frac{u(j) - u(j-k)}{\sqrt{k}} \tag{7}$$

and the variance of the observed returns is given by:

$$\mathrm{Var}(r_k(j)) = \mathrm{Var}(\tilde{r}_k(j)) + \frac{2\eta^2}{k} = \sigma^2 + \frac{2\eta^2}{k} \tag{8}$$

Therefore the observed variance is equal to the "true variance" (the variance of the Brownian motion describing the dynamics of the true price), plus an additional term coming from the incoherent component. This last term is responsible for the observed bias of the volatility. As long as the length of the return interval k is sufficiently long (say a number of ticks equivalent to one day or one week in physical time) the contribution of the incoherent term is negligible and so is the bias of the volatility estimation. But when high-frequency data are used (i.e k becomes small) the contribution of the additional component increases and the size of the bias is no longer negligible.

For the above model, the autocorrelation of the return at a lag h can be easily computed,

$$\rho(h) = \begin{cases} -\eta^2/(\sigma^2 k + 2\eta^2) & \text{for } h = 1 \\ 0 & \text{for } h > 1 \end{cases} \tag{9}$$

This implies $-0.5 \leq \rho(1) \leq 0$. The lower bound -0.5 is reached when σ^2 is completely negligible compared to η^2, and the return is the lag one difference of a noise. An empirical autocorrelation around -0.4, as observed for the USD/JPY and USD/CHF, implies $\eta^2 \simeq 2\sigma^2$. This indicates that at tick-by-tick level *the volatility originating from the incoherent component is largely predominant.* Therefore this effect should be carefully considered before using data at very high frequency.

The model can easily be extended for stock indices by introducing an autoregressive structure in the return $\tilde{r}(j)$:

$$\tilde{r}_1(j) = \phi\, \tilde{r}_1(j-1) + \epsilon(j) \tag{10}$$

with $\epsilon \simeq$ i.i.d.$(0,\ \sigma_\epsilon^2)$. Then the autocovariance structure of the model becomes

$$\mathrm{E}[r_l r_{l-h}] = \begin{cases} \sigma^2 + 2\eta^2 & \text{for } h = 0 \\ \phi\sigma^2 - \eta^2 & \text{for } h = 1 \\ \phi^h \sigma^2 & \text{for } h \geq 2 \end{cases} \tag{11}$$

with $\sigma^2 = \sigma_\epsilon^2/(1-\phi^2)$. This lagged correlation replicates the empirical data for stock indices as shown on Figure 2, where a small incoherent effect at lag one is present.

Thus, the simple model reproduces the empirical evidence for FX and stock indices. In particular it is able to replicate the strong negative first-lag autocorrelation of tick-by-tick returns and the observed anomalous scaling of the volatility.

6 EMA-Filtered Volatility for FX

On the basis of the above model for FX, it is possible to design a new high-frequency estimator for the volatility that discount for the bias induced by the incoherent term. The basic idea behind this new estimator starts from the observation that $r_1(j)$ has a MA(1) representation

$$r_1(j) = w(j) - \theta w(j-1) = (1 - \theta L)w(j) \tag{12}$$

with $w(j) = \text{i.i.d}(0,\ \Omega^2(\sigma, \eta))$ and $\theta = f(\sigma, \eta)$. This representation can be inverted to gives

$$w = (1 - \theta L)^{-1} r_1 \tag{13}$$

The $(1-\theta L)^{-1}$ operator is related to an exponential moving average (EMA) defined by the iterative equation

$$\text{EMA}[\theta; r_1](j) = \theta\, \text{EMA}[\theta; r_1](j-1) + (1-\theta) r_1(j). \tag{14}$$

When iterating the EMA definition, we obtain

$$\begin{aligned} \text{EMA}[\theta; r_1](j) &= (1-\theta)\left\{r_1(j) + \theta r_1(j-1) + \theta^2 r_1(j-2) + \cdots\right\} \\ &= \left\{(1-\theta)(1-\theta L)^{-1} r_1\right\}(j). \end{aligned} \tag{15}$$

Therefore, the white noise term w can be computed by a simple EMA in tick time[4]

$$w = (1-\theta)^{-1} \text{EMA}[\theta; r_1] \tag{16}$$

Furthermore, as $\text{E}[\text{EMA}[\theta; r_1]^2] = (1-\theta)^2 \text{E}[w^2] = \sigma^2$, the EMA filtering does not change the volatility. The appropriate parameters θ can be estimated from the first-lag autocorrelation of the return in tick time. Using the representation 12, the first-lag correlation is

$$\rho(1) = \frac{-\theta}{1+\theta^2} \tag{17}$$

and solving for θ gives

$$\theta = -\frac{1}{2\rho(1)}\left(1 - \sqrt{1 - 4\rho(1)^2}\right). \tag{18}$$

Because the EMA operator is linear, filtering the return is equivalent to filtering the price. Therefore, a tick-by-tick price time series, filtered from the incoherent component, is defined by

$$\mathcal{F}(x) = \text{EMA}[\theta; x]. \tag{19}$$

Then, a regular time series can be computed and the realized volatility estimated using the definitions 1 and 2. We will call this volatility estimator the filtered (realized) volatility $\sigma_{\mathcal{F}}$. The parameter $\hat{\theta}$ must be estimated from the first-lag correlation $\rho(1)$ of the return and eq. 18. According to the model for the price process, this must be computed in principle on the full data set to be filtered (i.e. in-sample). Yet, the incoherent term is related to the structure of the market and is relatively stable over time, as our tests have shown. Therefore, we can estimate the first-lag

[4] In this context, this is equivalent to the application of a Kalman filter since it can be shown that the EMA is the steady-state solution of this particular model. See [Harvey, 1989] pag. 175.

correlation of the return $\rho[T']$ on a moving window of length T' ending at the time t in order to filter the price at t. In this way, a causal estimate is constructed where the information ahead of the current time of the price filtering procedure is not used. The main advantage is that this filter can be used in real time (i.e. out-of-sample). The choice of the parameter T' and the kernel of the moving window only has a moderate influence on results. The procedure behaves well over a wide range of choices; our choice is a T' of the order of few weeks. For the practical implementation, moving autocorrelations can be conveniently computed using a moving average operator (MA) for an inhomogeneous time series as defined in [Zumbach and Müller, 2000]; this permits a highly efficient treatment of inhomogeneous time series (i.e. not equally spaced in time). With the (moving) estimate for the lag one correlation $\rho[T']$, the parameter $\hat{\theta}(t)$ can be computed using eq. 18, and the filtered price computed with $\mathcal{F}(x) = \mathrm{EMA}[\hat{\theta}; x]$ where $\hat{\theta}$ is now a time series. To summarize, the filtered volatility is computed as follows:

- estimate the first-lag autocorrelations of tick returns on a moving sample using moving average operators,
- compute $\hat{\theta}(t)$, using Equation 18,
- filter the log-price time series with an EMA in tick time with parameter $\hat{\theta}(t)$,
- compute a regular time series x_{RTS} from the filtered price $\mathcal{F}(x)$,
- compute the return and volatility at the desired frequency in a given time scale.

Let us emphasize that the filtering is done in a causal way, namely the parameter of the filter is evaluated on past data. Therefore, this filtering technique can be applied in real time, and not only to historical data, and it is "computationally cheap" as it requires only the computation of EMAs.

The results of this EMA filtering to several tick-by-tick FX series are reported in Figure 5 and table 1. The top panels of Figure 5 display the autocorrelation structures in tick time of the original and filtered return series of USD/JPY and USD/CHF rates. The bottom panels show the scaling behavior of original and filtered volatility. Clearly, removing the strong negative first-order autocorrelation, considerably reduces the volatility bias. We report in Table 1 values of the first-order autocorrelation with and without the application of the EMA-filter for different FX rates. Though not always perfect, the reduction of $\rho(1)$ is remarkable. Hence, the filtering of the incoherent component is effective in reducing the bias. Moreover, it is computationally efficient.

For stock indices, the model 10 contains an AR term for the return. This new term changes the character of the problem, as the class of models to consider now is an ARMA(1,1), for which we do not have the convenient EMA representation for w as in eq. 13 and 16. Therefore, an approach based on moving maximum likelihood estimation of ARMA(p, q) models should be done, but this is computationally too demanding, in practice.

The limitations of the EMA filtering mainly originate from the estimate of the coefficient θ which is computed on a past data sample for filtering at t. Therefore, there is a difference $\delta\hat{\theta}(t) = \hat{\theta}(t) - \theta$ between the fluctuating parameter values $\hat{\theta}(t)$ computed on a moving data window and the value θ computed with the full data sample. Note that $\theta < 1$, and therefore $\delta\hat{\theta} \ll 1$. The difference $\delta\hat{\theta}$ induces non-zero error terms in the lagged correlation of the return. As the time series $\delta\hat{\theta}(t)$ varies much more slowly than the tick-by-tick return, we can expand the computation of the correlation in the small parameter $\delta\hat{\theta}$, and then average $\delta\hat{\theta}(t)$. The resulting correlation at second order in

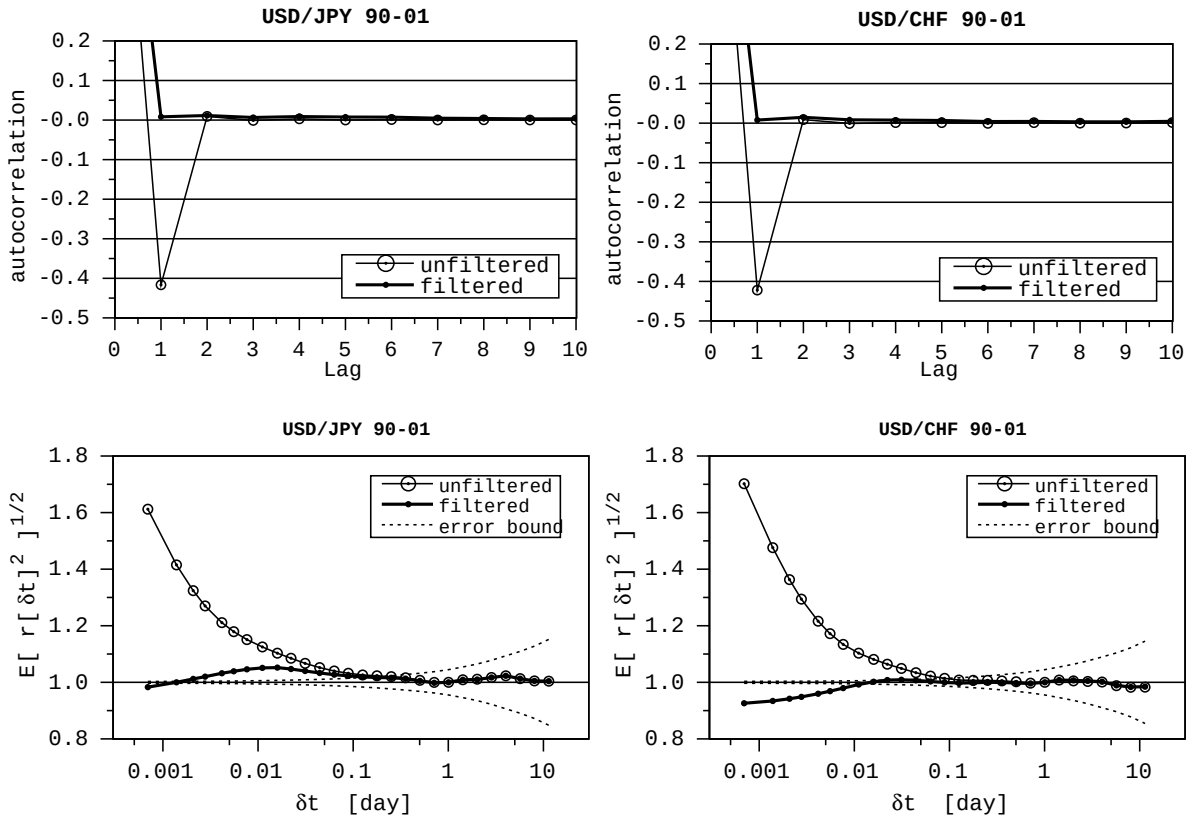

Figure 5: Results on tick time autocorrelations (top panels) and volatility scaling (bottom ones) of the application of the EMA-filter to tick-by-tick price series of USD/JPY and USD/CHF. The sample covers 11 years from January 1, 1990, to January 1, 2001. The first-lag correlation is evaluated on a moving sample of length 20 days.

	$E[\Delta t]$ θ-time [min]	$\rho(1)$ original	$\rho(1)$ filtered
USD/JPY	0.74'	-41.5%	0.78%
USD/CHF	0.82'	-42.0%	0.95%
GBP/USD	0.88'	-40.0%	-0.17%
EUR/USD	0.25'	-41.4%	-1.28%
EUR/GBP	1.9'	-29.8%	-4.66%
USD/ITL	1.6'	-35.0%	-3.15%
USD/DKK	2.3'	-36.0%	-1.38%
GBP/JPY	6.86'	-46.9%	4.8%

Table 1: Reduction of the first order autocorrelation of tick-by-tick returns obtained with the EMA-filter

$\delta\hat{\theta}(t)$ is

$$\begin{aligned} \rho(1) &= \mathrm{E}[\delta\hat{\theta}] + \mathrm{E}[\delta\hat{\theta}^2]\frac{\theta}{1-\theta^2} \\ \rho(k) &= \mathrm{E}[\delta\hat{\theta}]\theta^{k-1} + \mathrm{E}[\delta\hat{\theta}^2]\left(\frac{\theta^k}{1-\theta^2} + (k-1)\theta^{k-2}\right) \qquad \text{for} \quad k \geq 2. \end{aligned} \tag{20}$$

Essentially, the correlation decays exponentially due to the memory kernel of the EMA. The term in $\mathrm{E}[\delta\hat{\theta}]$ should average out for a sufficiently long sample, and the leading error term is of order $\mathrm{E}[\delta\hat{\theta}^2]$ which is much smaller. This analysis is in agreement with a deeper analysis of the results presented in Figure 5. Another limitation of the EMA filtering is the adequacy of the MA(1) model presented in Section 5. As we already discussed, stock indices require an ARMA model. For minor foreign exchange rates such as USD/DKK or GBP/JPY, the lagged correlation of the tick-by-tick return also shows a small AR component, but with a negative coefficient. Although the ticks of minor rates are also genuine Reuters quotes, they are likely to be influenced by or even computed from major rates. This can induce further correlations, not present in the simple MA(1) model. Yet, also for minor rates, the leading effect is clearly the incoherent term, and the EMA filtering achieves a good reduction of the negative autocorrelation.

7 Conclusions

The evident limitations of standard definitions of daily volatility based on daily returns strongly motivates the use of high-frequency data. However, microstructure effects make the empirical returns not i.i.d.. In particular, the multiple contributor structure of the FX spot market generates an *incoherent* component in the observed price. At the tick level, the variance of this component is large, typically twice the variance of the "true" returns. This incoherent effect induces a strong negative first-order autocorrelation of returns in tick time, leading to a considerable anomalous scaling of the realized volatility as estimated by the usual formula. Before high-frequency volatility estimates can be applied in practice, this strong *bias* needs to be corrected.

On the basis of a simple theoretical model that accommodates well the empirical properties of FX data, a simple filter is presented that removes the incoherent component from the raw tick-by-tick time series. The filter is only based on the computation of exponential moving averages and is therefore computationally very efficient, both time-wise and memory-wise. Using filtered prices, the daily volatility can be estimated using the information contained in high-frequency data, allowing for a nearly unbiased measure of high precision. For stock indices, the averaging of many share prices induces further positive correlations, and a more complicated treatment is necessary.

References

[Ahn et al., 1999] **Ahn D.-H., Boudoukh J., Richardson M., and Whitelaw R. F.**, 1999, *Behavioralize this! international evidence on autocorrelation patterns of stock index and futures returns*, Stern Business School Working Paper Series, 1–39.

[Andersen et al., 1999] **Andersen T. G., Bollerslev T., Diebold F. X., and Labys P.**, 1999, *The distribution of exchange rate volatility*, Working Paper, Department of Finance, Stern Business School, 1–45.

[Andersen et al., 2000] **Andersen T. G., Bollerslev T., Diebold F. X., and Labys P.**, 2000, *Microstructure bias and volatility signature*, Manuscript in progress.

[Breymann et al., 2000] **Breymann W., Zumbach G., Dacorogna M. M., and Müller U. A.**, 2000, *Dynamical deseasonalization in otc and localized exchange-traded markets*, Internal document WAB.2000-01-31, Olsen & Associates, Seefeldstrasse 233, 8008 Zürich, Switzerland.

[Clark, 1973] **Clark P. K.**, 1973, *A subordinated stochastic process model with finite variance for speculative prices*, Econometrica, **41**(1), 135–155.

[Conrad and Kaul, 1988] **Conrad J. and Kaul G.**, 1988, *Time varying expected returns*, Journal of Business, **61**, 409–425.

[Dacorogna et al., 1993] **Dacorogna M. M., Müller U. A., Nagler R. J., Olsen R. B., and Pictet O. V.**, 1993, *A geographical model for the daily and weekly seasonal volatility in the FX market*, Journal of International Money and Finance, **12**(4), 413–438.

[Dacorogna et al., 1998] **Dacorogna M. M., Müller U. A., Olsen R. B., and Pictet O. V.**, 1998, *Modelling short-term volatility with GARCH and HARCH models*, published in "Nonlinear Modelling of High Frequency Financial Time Series" edited by Christian Dunis and Bin Zhou, John Wiley, Chichester, 161–176.

[Goodhart, 1989] **Goodhart C. A. E.**, 1989, *'news' and the foreign exchange market*, Proceedings of the Manchester Statistical Society, 1–79.

[Goodhart and Figliuoli, 1991] **Goodhart C. A. E. and Figliuoli L.**, 1991, *Every minute counts in financial markets*, Journal of International Money and Finance, **10**, 23–52.

[Harvey, 1989] **Harvey A. C.**, 1989, *Forecasting, Structural Time Series Models, and the Kalman Filter*, Cambridge University Press, Cambridge.

[Hasbrouck, 1993] **Hasbrouck J.**, 1993, *Assessing the quality of a security market: A new approach to transction-cost measurement*, Review of Financial Studies, **6**(1), 191–212.

[Hawawini, 1980] **Hawawini G.**, 1980, *Intertemporal cross dependence in securities daily returns and the short-term intervalling effect on systematic risk*, Journal of Financial and Quantitative Analysis, **15**, 139–149.

[Holden and Subrahamanyam, 1992] **Holden A. and Subrahamanyam A.**, 1992, *Long-lived private information and imperfect competition*, Journal of Finance, **47**, 247–270.

[Lo and MacKinlay, 1988] **Lo A. W. and MacKinlay A. C.**, 1988, *Stock market prices do not follow random walks: evidence from a simple specification test*, The Review of Financial Studies, **1**, 41–66.

[M. Brennan and Swaminathan, 1993] **M. Brennan N. J. and Swaminathan B.**, 1993, *Investment analysis and the adjustment of stock prices to common information*, Review of Financial Studies, **6**, 799–824.

[Merton, 1980] **Merton R. C.**, 1980, *On estimating the expected return on the market: An exploraty investigation*, Journal of Financial Economics, **8**, 323–361.

[Roll, 1984] **Roll R.**, 1984, *A simple implicit measure of the effective bid-ask spread in an efficient market*, The Journal of Finance, **39**(4), 1127–1139.

[Schwert, 1998] **Schwert G. W.**, 1998, *Stock market volatility: Ten years after the crash*, Brookings-Wharton Papers on Financial Services.

[Taylor and Xu, 1997] **Taylor S. J. and Xu X.**, 1997, *The incremental volatility information in one million foreign exchange quotations*, the Journal of Empirical Finance, **4**(4), 317–340.

[Zumbach and Müller, 2000] **Zumbach G. O. and Müller U. A.**, 2000, *Operators on inhomogeneous time series*, To be published in International Journal of Theoretical and Applied Finance.