Efficient Estimation of Volatility using High Frequency Data

Gilles Zumbach[1], Fulvio Corsi[2], and Adrian Trapletti[3]

Olsen & Associates
Research Institute for Applied Economics
Seefeldstrasse 233, 8008 Zürich, Switzerland.

phone: +41-1/386 48 48 Fax: +41-1/422 22 82

February 21, 2002

Keywords: volatility estimators, high-frequency data, incoherent price formation, daily volatility.

Abstract

The limitations of volatilities computed with daily data as well as simple statistical considerations strongly suggest to use intraday data in order to obtain accurate volatility estimates. Under a continuous time arbitrage-free setup, the quadratic variations of the prices would allow us, in principle, to construct an approximately error free estimate of volatility by using data at the highest frequency available. Yet, empirical data at very short time scales differ in many ways from the arbitrage-free continuous time price processes. For foreign exchange rates, the main difference originates in the incoherent structure of the price formation process. This market micro-structure effect introduces a noisy component in the price process leading to a strong overestimation of volatility when using naive estimators. Therefore, to be able to fully exploit the information contained in high frequency data, this incoherent effect needs to be discounted. In this contribution, we investigate several unbiased estimators that take into account the incoherent noise. One approach is to use a filter for pre-whitening the prices, and then using volatility estimators based on the filtered series. Another solution is to directly define a volatility estimator using tick-by-tick price differences, and including a correction term for the price formation effect. The properties of these estimators are investigated by Monte Carlo simulations. A number of important real-world effects are included in the simulated processes: realistic volatility and price dynamic, the incoherent effect, seasonalities, and random arrival time of ticks. Moreover, we investigate the robustness of the estimators with respect to data frequency changes and gaps. Finally, we illustrate the behavior of the best estimators on empirical data.

[1]Present address: Consulting in financial engineering, Ch. Charles Baudouin 8, 1228 Saconnex d'Arve, Switzerland
e-mail: gilles.zumbach@bluewin.ch

[2]Present address:University of Southern Switzerland Via Buffi, 13 CH-6904 Lugano, Switzerland
e-mail: fulvio.corsi@lu.unisi.ch

[3]Present address:Wildsbergstrasse 31, CH-8610 Uster, Switzerland
e-mail: a.trapletti@bluewin.ch

1 Introduction

Volatility enters as an essential ingredient in many financial computations, like portfolio optimization, option pricing or risk assessment. Although these computations depend critically on the value of the volatility, the estimation of volatility is often done using daily data. Yet, as the volatility measures intuitively "how much the prices jitter", there is a gain in using data at higher frequencies. The relevant statistical concept corresponding to our intuition is the *minimal sufficient statistic*: this is the smallest subset of empirical values needed to evaluate a statistical estimate without loosing information. As a simple example, let us consider a random walk with constant drift μ and volatility σ. Given one realization of the random walk, what are the minimal sufficient statistics to estimate μ and σ? For the drift μ, the answer is that only the start and end points of the random walk are needed. This implies that there is no gain in using high frequency data for estimating the drift. For the volatility σ, the answer is that the absolute value of *every increment* is needed (there is no information in the sign of the increments regarding the volatility). This means that all the increments help in getting a better estimate for the volatility, and any thinning of the original data implies a loss of information. Hence, to estimate volatility of a random walk, we should use data at the highest frequency available, namely tick-by-tick data. Under the more general continuous time arbitrage-free setup, a similar result holds [Andersen et al., 2001b].

Although these theoretical results provide a good hint about the best estimator, the actual data differ in many way compared to the simplest random walk model. A first "naive" estimator for the daily volatility σ using tick-by-tick data results from

$$\sigma^2 = \frac{1}{n}\sum_j r(j)^2 \tag{1}$$

where the sum covers one day of data, n is the number of ticks in this day, and the return $r(j)$ is the difference between two consecutive logarithmic prices. This estimator is the best possible estimator for a random walk without drift. Yet, for empirical foreign exchange (FX) data, this estimator is very strongly biased: the mean σ^2 value is roughly three times larger when compared to the standard deviation of daily returns. This bias originates in the microstructure effects that start to dominate the evolution of prices at the very high frequencies. In short, some time is needed for the market participants to exchange information about their views of the prices, and to reach a consensus. At very short time intervals, the price process is better described by a distribution of quotes around a consensus price, and this quote distribution reflects the incoherence between the traders. According to [Corsi et al., 2001], we call this effect the *incoherent* price formation process, and, in the context of volatility estimation, this effect has first been analyzed by [Zhou, 1996]. A process describing well the tick-by-tick empirical prices is a random walk, plus an additive noise term describing the different opinions, or the incoherence, of the market participants. This additive noise induces a strong negative first autocorrelation of the order of -40% for the tick-by-tick returns, and also the strong bias in the naive volatility estimator 1. Because of the magnitude of the incoherent term, this effect clearly has to be discounted.

Many possible volatility estimators that discount for the incoherent effect can be constructed. Then, we are left with the problem of comparing the different estimators in order to select the best one. The usual comparison criteria are that the volatility estimators should return positive values and have a small bias and variance. Moreover, we would like to have a robust estimator for different tick rates, including possibly week-ends and data gaps. Finally, as we are dealing with tick-by-tick data, the estimator has to be numerically efficient, both in terms of memory requirement and computational time. As the volatility of empirical data is an unknown quantity, we cannot assess the quality of the different estimators directly using financial data. Therefore, we have to perform simulations, using for example ARCH like processes, where the actual volatility is known and can be compared to the estimated volatilities. In this comparison, we have also included two standard volatility measures using daily data, namely risk metric and the squared value of the daily return. In this way, the improvement resulting from using high frequency data can also be assessed.

In this paper, we discuss the volatility σ in the text, but all the definitions are given for the variance σ^2. Similarly, most statistical quantities about the volatilities are indeed computed with the variance σ^2. For example, an unbiased volatility estimator means indeed $\mathrm{E}[\hat{\sigma^2}] = \sigma^2$ where $\hat{\sigma^2}$ is an estimator for the square volatility and σ^2 is the expected variance. The reader should be aware of this potential difference between the text and the formulae. This discrepancy is customary in the literature and originates in the quadratic definitions for the variance whereas intuitively one want a quantity homogeneous (of degree one) with the return.

The notation we are using follows [Zumbach and Müller, 2001]. A time series is denoted by a single letter, like x. The value of a time series at a given time t or tick j is denoted with parenthesis, like $x(t)$ or $x(j)$. The time for the j-th tick is denoted with $t(j)$. The parameters are between square brackets, like $\sigma[\Delta t]$ (for a time series) or $\sigma[\Delta t](t)$ (for a real number). We will use the function "positive part", which is the function $(x)^+ = \max(x, 0)$.

This paper is organized as follows. Section 2 introduces the model for the price process with the incoherent term, and then all the volatility estimators we are investigating. Section 3 is devoted to the Monte Carlo investigations of the various estimators and of their robustness with respect to seasonality. Analysis of empirical data is done in section 4, in particular to validate the model presented in section 2. The conclusions are presented in section 5, with our best estimate for computing the volatility.

2 Definitions of the volatility estimators

2.1 Price process with an incoherent term

In order to motivate the definitions of volatility we use, we describe here briefly the process for the tick-by-tick prices developed in [Corsi et al., 2001]. The price evolution is described in "tick time" with an integer index j. The time increment process $\delta t(j)$ is described by another process. For example, for the Monte Carlo simulations below, the time increments are taken as a simple Poisson process (i.e. drawn i.i.d. randomly from an exponential distribution).

The observed logarithmic prices $x(j)$ are given by

$$x(j) = \tilde{x}(j) + u(j) \tag{2}$$

where $\tilde{x}(j)$ is an unobserved "true" price, which follows a (continuous) diffusion process. The incoherent term $u(j)$ is assumed to be an i.i.d. white noise component with zero mean and variance η^2, $u(j) \sim$ i.i.d.$(0,\ \eta^2)$. No other assumption on the distribution of u is made. The simplest model for $\tilde{x}$ is a random walk, with

$$\tilde{x}(j) = \tilde{x}(j-1) + \tilde{r}(j) \tag{3}$$

with the return $\tilde{r}(j) \sim$ i.i.d.$(0,\ \sigma^2)$. For this process, the tick-by-tick observed return is

$$r(j) = x(j) - x(j-1) = \tilde{r}(j) + u(j) - u(j-1) \tag{4}$$

and the k-ticks return is

$$r_k(j) = x(j) - x(j-k) = \tilde{r}_k(j) + u(j) - u(j-k)\ . \tag{5}$$

The u terms in these equations induce the strong deviation at short time intervals from the usual random walk behavior. The variance for the observed tick return is

$$\mathrm{E}[r_k^2] = \sigma^2 k + 2\eta^2 \tag{6}$$

and the lagged correlation for r_k is

$$\rho(h) = \begin{cases} -\eta^2/(\sigma^2 k + 2\eta^2) & \text{for } h = 1 \\ 0 & \text{for } h > 1 \end{cases} \tag{7}$$

For typical FX data, the empirical value of η^2 is of the order of $2\sigma^2$, meaning that at the tick time horizon, the incoherent term quantitatively dominates the random walk component. Therefore, when computing volatilities at very short time horizons, this term has to be accounted for. With time aggregation, the random walk component $\tilde{r}[\Delta t] = \tilde{x}(t) - \tilde{x}(t - \Delta t)$ scales as $\mathrm{E}[\tilde{r}^2[\Delta t]] \sim \Delta t\ \sigma^2$ whereas the incoherent term has a trivial scaling $\mathrm{E}[(u(t) - u(t - \Delta t))^2] \sim 2\eta^2$. At the daily time horizon, the random walk dominates and the incoherent effect can be neglected.

Model 2 implies that at very small time intervals, we do not have a random walk with "a" well defined price, but instead a distribution of prices around some "consensus" mean price. This is a departure from most of the current models that assume a fixed price at each point in time. Similar models have already been developed by [Moody and Wu, 1997, Moody and Wu, 1998, Zhou, 1996]. In some sense, this is similar to the paradigm shift from classical mechanics to quantum physics, where a description of point particles is replaced by a probabilistic description with the associated uncertainty. The fundamental physical constant measuring the uncertainty is $\hbar$, and the equivalent in the finance world is played by the spread.

Model 2 also differs somewhat from the model process used in [Zhou, 1996]. We are using a description in tick time, with a constant volatility between ticks given by σ. Zhou uses an underlying continuous process in physical time with a constant volatility σ, observed at the tick time $t(j)$. For the Zhou model, the volatility between ticks is $\mathrm{E}[r^2(j)] = (t(j) - t(j-1))\sigma^2$, namely is proportional to the time elapsed between ticks. This difference leads to different analytical results, for example in the variance for the volatility estimator. An empirical analysis of the time series for the hourly volatility, the number of ticks and the (hourly) volatility per tick shows that the volatility per tick is roughly constant, and that the volatility is proportional to the number of ticks (see section 4). This indicates that our model provides a better description of the observed empirical data.

Due to the above random walk scaling, the volatility and the volatility estimators depend on the time horizon at which they are measured. It is more convenient to remove this scaling and to always report the volatility at a reference time scale, which is usually taken to be one year. This is done for example with the definition of the naive estimator as

$$\sigma^2[\Delta t, \delta t] = \frac{1y}{\delta t}\ \frac{1}{n}\sum_t r^2[\delta t](t) \tag{8}$$

where the sum is over a time interval Δt, n is the number of terms in the sum, and $1y$ denotes one year. In this way, the volatility is at leading order independent of the parameters Δt and δt, namely $\mathrm{E}[\sigma^2[\Delta t, \delta t]] \simeq \sigma^2$, and is numericaly comparable to the volatility of the yearly price changes. This discounting of the measurement time horizon δt is called annualization. All the definitions below are directly annualized, including the tick-by-tick definitions, so that all volatilities can be directly compared regardless of the parameter values.

In all the definitions below, we assume that the return r has a zero expectation. As we are working with short time intervals, typically of the order of 1 day, this is a good assumption. Consider for example a time series with a mean annualized drift μ_{ann} of 10% and an equal annualized volatility σ_{ann} of 10%. At the daily level Δt = 1d, we obtain $\mu[\Delta t] = (\Delta t/1y) \cdot\ \mu_{\mathrm{ann}}$ and $\sigma[\Delta t] = \sqrt{\Delta t/1y}\ \sigma_{\mathrm{ann}}$. Because of the different scalings between the mean and the standard deviation, the volatility dominates the mean by a factor $\sqrt{260} \simeq 16$ at the daily level, and the mean can be safely neglected. Should the zero expected return assumption not be valid, or should we be interested in longer time intervals, the modifications in the definitions below are straightforward.

2.2 Volatilities using daily data

Many definitions for measuring volatility using daily data exist. The simplest volatility estimate is the squared value of the daily return

$$\{\ r[1d](t)\ \}^2 . \tag{9}$$

From the statistical point of view, this estimator is very bad. For example, for a Gaussian random walk, the root mean square error (RMSE) of $r^2[1d]$ is 141% σ^2. Yet, this definition has the advantage not to be damped by an average with the past history.

The most accepted volatility estimator using daily data is the RiskMetric definition

$$\sigma^2_{\text{RiskMetric}}(t) = \mu\, \sigma^2_{\text{RiskMetric}}(t - 1d) + (1-\mu)\, r^2[1d](t) \tag{10}$$

with the constant coefficient $\mu = 0.94$. This formula corresponds to an exponential moving average with a characteristic time τ given by the formula $\mu = \exp(-1\text{d}/\tau)$, or $\tau \simeq 16$ business days $\simeq 3$ weeks. This value can be seen as a compromise imposed by daily data between having a good statistical estimator, for which a longer time interval is needed, and the short term dynamic of the volatility, by which most information is in the very recent past. Because of this fundamental trade off, the value $\mu = 0.94$ provides reasonable estimates, regardless of the time series.

More complex definitions can be used, for example a GARCH(1,1) process. Yet, these definitions involve more parameters, which are depending on the time series under consideration. For example, the GARCH(1,1) process involves (implicitly or explicitely) the mean volatility among its parameters, a quantity that strongly depends on the time series. Beside, the quantitative improvement over RiskMetric is typically too small to justify the added complexity and parameters. For these reasons, we have restricted our empirical investigations using daily data to the simplest RiskMetric and $|r[1d]|$ formulas.

2.3 The regular time series volatility

The standard realized volatility estimator is defined by summing squared returns of an artificial regular time series of logarithmic prices $x_{\text{RTS}}(t)$. The usual definition for the annualized (realized) volatility over a time interval Δt is

$$\sigma^2_{\text{RTS}}[\Delta t, \delta t](t) \quad = \quad \frac{1}{n} \sum_{t-\Delta t+\delta t \le t' \le t} r^2[\delta t](t') \tag{11}$$

where the annualized return is defined as

$$r[\delta t](t) \quad = \quad \sqrt{\frac{1\text{y}}{\delta t}}\,(x_{\text{RTS}}(t) - x_{\text{RTS}}(t-\delta t)) \tag{12}$$

$$n \quad = \quad \sum_{t-T+\delta t \le t' \le t} \tag{13}$$

and with

- x_{RTS}: a Regular Time Series (RTS) spaced by δt of (logarithmic middle) prices. This quantity needs to be computed with some interpolation procedure from the irregularly spaced high-frequency tick-by-tick price time series $x(j)$.
- $r[\delta t]$: the annualized returns, observed over a time interval of size δt.

- 1y: the one year normalization period.
- Δt: the length of the moving window over which the volatility is computed.
- n: the number of return observations in the interval Δt. In the usual case of non-overlapping return intervals, the number of observations is $n = \Delta t/\delta t$.

This definition involves two time parameters, Δt and δt.

As several studies have shown (see e.g. [Andersen et al., 2001a, Corsi et al., 2001]), this estimator is strongly biased upward for small return time interval δt due to the incoherent component in the price process. To achieve unbiasedness, the lower bound for the value of the parameter δt should be of the order of 30 minutes to 2 hours for typical FX data [Andersen et al., 2001a]. This limits the usefulness of such an estimator. On the other hand, it is possible to filter the incoherent component of the (logarithmic middle) prices. Then, we can use the estimator 11 on the filtered price series with small time intervals δt, as suggested in [Corsi et al., 2001]. Such a filter can be based on an MA(1) representation for the return which can be inverted to lead to an EMA filter with a parameter depending on the lag one correlation for the tick-by-tick return. [Corsi et al., 2001] shows that even for very small return time intervals δt, the estimator 11 has almost no bias. In the empirical analysis in section 4, we have included σ^2_{RTS} evaluated with unfiltered prices and 30 minute returns, and σ^2_{RTS} evaluated with filtered prices and 5 minute returns.

2.4 The tick-by-tick volatilities

2.4.1 The Zhou volatility

As far as we know, [Zhou, 1996] was the first author to notice the problems induced by the incoherent effect when estimating volatilities from high frequency data and to propose an estimator of volatility that corrects for it. The Zhou estimator is given by

$$\sigma^2_{\mathrm{Zhou}}[\Delta t, k](t) = \frac{1y}{\Delta t}\,\frac{1}{k}\sum_{t-\Delta t\le t(j)\le t} r_k^2(j) + 2r_k(j)r_k(j-k) \tag{14}$$

where the sum $\sum_{t(j)}$ is over all ticks between $t-\Delta t$ and t. The return r_k is given as a k-tick return (see Eq. 5). For $k > 1$, this definition uses overlapping returns (the sum is over all the ticks). Essentially, the "naive" estimator $r_k^2 \sim k\sigma^2 + 2\eta^2$ is corrected by the term $2r_k(j)r_k(j-k) \sim -2\eta^2$. Yet, because of the large cancellation between the two estimators for the variance and lag$-k$ covariance, the resulting estimate may be negative. The non positivity of this definition is a serious drawback, particularly if the number of ticks in the interval Δt is not large enough. This can be corrected by taking the positive part of the volatility, but a zero volatility is not good either.

Another drawback of this formula is to be computationally cumbersome. Because the number of ticks in the interval Δt is not constant, a stack of returns and times has to be kept. This implies that the memory requirement is not fixed, and grows with Δt and the tick frequency. This problem can be solved when the values of the estimator are needed only at regular time points separated with τ, say every hour, and with $\Delta t = p\tau$ with p an integer. In this case, the estimator can be written as a double sum over p and τ, and the memory requirement is fixed to one stack of length p. These fine points are very important when dealing with tick-by-tick data as memory allocation is very costly time wise compared to a few multiplications and additions.

2.4.2 The Zhou volatility with covariance estimated on a longer sample

The previous estimator can be changed by measuring the incoherent correction on a much longer time interval. This should reduce the variance, and therefore the probability to get negative values. In this direction, several variations of the Zhou estimator can be written. One possible estimator is the following:

$$\sigma^2_{\text{Zhou+LC}}[\Delta t, \Delta t', k](t) = \frac{1y}{\Delta t}\frac{1}{k}\left(\sum_{t-\Delta t\le t(j)\le t} r_k^2(j) + \frac{n[\Delta t](t)}{n[\Delta t'](t)}\sum_{t-\Delta t'\le t(j)\le t} r_k(j)r_k(j-k)\right) \quad (15)$$

$$n[\Delta t](t) = \sum_{t-\Delta t\le t(j)\le t} \quad (16)$$

where $n[\Delta t](t)$ is the number of ticks between $t-\Delta t$ and t. Essentially, the ratio Covariance$[\Delta t']/n[\Delta t']$ measures the incoherent term per tick over the interval $\Delta t'$. The value for the parameter $\Delta t'$ needs to be chosen large enough to obtain a good estimator, typically of the order of a few weeks. Again, this formula might give negative values. We call this estimator Zhou with long covariance $\sigma^2_{\text{Zhou+LC}}$.

Other estimators can be defined, for example

$$\sigma^2[\Delta t, \Delta t', k](t) = \frac{1y}{\Delta t}\frac{1}{k}\left(1+2\rho[\Delta t'](t)\right)\sum_{t-\Delta t\le t(j)\le t} r_k^2(j) \quad (17)$$

$$\rho[\Delta t'](t) = \frac{\sum_{t-\Delta t'\le t(j)\le t} r_k(j)r_k(j-k)}{\sum_{t-\Delta t'\le t(j)\le t} r_k^2(j)} \quad (18)$$

where $\rho[\Delta t']$ is the lag k correlation measured on the interval $\Delta t'$. The seasonality analysis of empirical data shows that the covariance is constant whereas the variance has a dependency with respect to the time in the week (see section 4). This induces a seasonality in the correlation, and therefore an undesired dependency with respect to the parameter $\Delta t'$ in the volatility estimator. For this reason, we have not further investigated this estimator.

2.4.3 The quadratic variation

As the incoherent filter developed in [Corsi et al., 2001] allows us to remove the incoherent component, we can use a "naive" estimator based on squared returns of filtered prices. We call this estimator "quadratic variation" as it corresponds to the usual quadratic volatility estimator for a random walk

$$\sigma^2_{\text{QV}}[\Delta t, k](t) = \frac{1y}{\Delta t}\frac{1}{k}\sum_{t-\Delta t\le t(j)\le t} r^2_{f,k}(j) \quad (19)$$

where $r_{f,k}$ is the k-ticks apart return computed from the filtered prices x_f. For the volatility estimate, this corresponds to pre-whitening the prices, and then to compute the volatility of the resulting time series. As the objective is to compute the volatility of the filtered data, no recoloring, or correction of the volatility, needs to be done.

2.4.4 The filtered Zhou definitions

As shown in section 3, the variance for the Zhou estimator depends directly from the incoherent noise level. In order to reduce its variance, we can first apply an incoherent filter, and then use the Zhou estimator to compute the volatility. We denote this estimator as "filter + Zhou".

The filter as presented in [Corsi et al., 2001] is not correctly specified as it assumes a constant volatility per tick. Yet, for empirical data, the volatility per tick shows a seasonality (see section 3). This implies that part of the first lag negative correlation remains after filtering. By using a Zhou estimator on filtered prices, the remaining first lag correlation is correctly discounted (but further lagged correlations induced by the filter remain). This correction is important at short time horizons, below one day. In a broader context, this approach corresponds to pre-whitening the data, then to compute the relevant quantity, and finally to recolor the estimate [Andrews and Monahan, 1992].

2.4.5 The bias corrected Zhou volatility

If the diffusion process $\tilde{x}(j)$ (see Eq. 2) does not have a constant volatility σ^2, then the Zhou estimator 14 is biased for k greater than one. This bias essentially originates from the use of overlapping returns in the sum of Eq. 14. For the case $k = 2$ some simple algebra reveals that the bias is introduced by the first and last term in the sum of Eq. 14 (see also [Zhou, 1996], p. 48 for a more detailed discussion).

Hence, a bias corrected Zhou volatility estimator may be defined by appropriately adjusting the estimator for the terms that introduce the bias. To simplify the formula, we only take into account the two terms which bias the Zhou estimator with $k = 2$. The bias corrected version is then given by

$$\sigma^2_{\text{Zhou+BC}}[\Delta t, k](t) = \sigma^2_{\text{Zhou}}[\Delta t, k](t) + \delta V[\Delta t, k](t) \tag{20}$$

$$\delta V[\Delta t, k](t) = \frac{1y}{\Delta t}\,\frac{k-1}{k}\left(\{r_1^2(j_a) + 2r_1(j_a)r_1(j_a-1)\} - \{r_1^2(j_b) + 2r_1(j_b)r_1(j_b-1)\}\right) \tag{21}$$

where $\delta V[\Delta t, k](t)$ denotes the bias correction and $t(j_a)$ and $t(j_b)$ are the times of the ticks arriving immediately before $t - \Delta t$ and t, respectively. Obviously, the bias corrected Zhou estimator is equal to the (uncorrected) Zhou estimator if $k = 1$. Furthermore, it is easy to show that the estimator 20 is unbiased for $k = 2$. For k greater than two, the estimator 20 is still biased, but the bias correction eliminates the largest terms causing the bias (all other terms are multiplied by a factor of $(k-i)/k$, $2 \le i \le k-1$ instead of the factor $(k-1)/k$ in Eq. 21). Also note, that the annualization factor in Eq. 21 and 14 is only correct, if there are ticks arriving exactly at times $t - \Delta t$ and t. In our implementation, it is ensured that the volatility is always computed on the time interval $[t(j_a), t(j_b)]$ and the volatility estimator is additionally corrected by a factor $\Delta t/(t(j_b) - t(j_a))$. This last correction is important in case of gaps.

2.4.6 Other definitions for the tick-by-tick volatility

The sensitivity of the volatility estimators with respect to the incoherent term is introduced by the term r^2 in formula 14. A way to avoid this term is to take a product of overlapping returns, but at different time points, namely to "alternate" all the time points. This leads to the formula

$$\sigma^2_{\text{alternate}}[\Delta t] = \frac{1y}{\Delta t}\frac{1}{k-m}\sum_{t-\Delta t \le t(j) \le t} r_k(j)\,r_k(j-m) \qquad \text{with } 1 < m < k\,. \tag{22}$$

This definition might also give negative values. When applied to empirical data and Monte Carlo simulations, this estimator behaves similarly to the Zhou estimator. Moreover, it is not possible to improve this estimator along the line described in section 2.4.2. Because it is quite redundant with the Zhou definition, we have not included this estimator in the analysis below.

3 Properties of the volatility estimators

3.1 The Monte Carlo testing set-up

As the actual volatility is unknown for empirical data, we use Monte Carlo simulations to assess the properties of the various volatility estimators. The simulations are done with a constant volatility random walk or with a GARCH(1,1) model. In the simulations, the actual volatility at each time t is known, and this "instantaneous" volatility can be integrated over an interval Δt, say of 1 day, in order to get the "true" integrated volatility $\sigma_{\text{integrated}}$ (some authors called this the realized volatility). Then, the various estimators can be benchmarked against this volatility.

In more details, the set-up for the Monte Carlo simulations is as follows. The unobserved "true" price $\tilde{x}(j)$ follows a process in tick time

$$\begin{aligned} \tilde{x}(j) &= \tilde{x}(j-1) + \tilde{r}(j) \\ \tilde{r}(j) &= \sigma_{\text{eff}}(t)\ \varepsilon(t). \end{aligned} \tag{23}$$

The residuals $\varepsilon(t)$ are i.i.d. with $\mathrm{E}[\varepsilon(t)] = 0$ and $\mathrm{E}[\varepsilon^2(t)] = 1$. For all the simulations, the residuals are drawn from a Student-t distribution with 6 degrees of freedom. This ensures that the distribution of returns has a high degree of kurtosis, and this numerical value is consistent with observed empirical high frequency data. The underlying volatility process is taken to be either

- constant with $\sigma_{\text{eff}} = 1$,
- with a GARCH(1,1) dynamic with parameters corresponding to a mean annualized volatility $\mathrm{E}[\sigma^2_{\text{eff}}] = 1$ and a characteristic time of the decay of the autocorrelation function of the volatility of τ_{corr} = 10 days.

The time interval δt between the simulated prices is taken i.i.d. randomly from an exponential distribution (Poisson process). The "true" integrated volatility is computed with

$$\sigma^2_{\text{integrated}}[\Delta t](t) = \sum_{t-\Delta t \le t(j) \le t} \sigma^2_{\text{eff}}(j) \tag{24}$$

with Δt = 1day.

The observed logarithmic prices $x(j) = \tilde{x}(j) + u(j)$ are obtained by adding the random incoherent term to the "true" price. We have taken a Gaussian distribution for u, with a variance η^2 related to the mean annualized volatility σ^2 of the "true" price process by

$$\eta^2 = z^2 \frac{\mathrm{E}[\delta t]}{1\mathrm{y}} \sigma^2 \tag{25}$$

with $\mathrm{E}[\delta t]$ the mean time interval between the quotes. The factor z^2 fixes the incoherent noise level. All the simulations are done with $z^2 = 2$, in agreement with the empirical value found using Reuters data [Corsi et al., 2001], namely the incoherent term dominates the random walk component at the tick level by a factor of two.

The observed logarithmic prices $x(j)$ are used as the input for the various volatility estimators. Hence, the inaccuracy of the estimators originates from the computation of the return with $\tilde{r} = \sigma_{\text{eff}}\ \varepsilon$ and from the addition of the incoherent component to the prices $x = \tilde{x} + u$. In the studies below, we have included the following estimators: the daily squared return $|r[1d]|^2$, RiskMetric $\sigma^2_{\text{RiskMetric}}$, the (unfiltered) regular time series volatility $\sigma^2_{\text{RTS}}[1d, 30']$ with a return time interval of 30 minutes, the filtered regular time series

volatility $\sigma^2_{\text{RTS}}[1d, 5']$ with a return time interval of 5 minutes, the Zhou volatility $\sigma^2_{\text{Zhou}}[1d]$, the microscopic volatility $\sigma^2_{\text{Zhou+LC}}[1d, \Delta t']$ with the incoherent term measured on $\Delta t'$ = 10 days, the filtered Zhou volatility (i.e. the incoherent filter followed by a Zhou estimator), and the filtered quadratic variations.

We define the estimation error as

$$\Delta\sigma^2 = \sigma^2 - \sigma^2_{\text{integrated}}, \tag{26}$$

where σ^2 is one of the estimators under consideration. Furthermore, we also estimate the pdf (probability density function) of the various estimators. The correlation between σ^2 and $\sigma^2_{\text{Integrated}}$, and between $(\left(\sigma^2\right)^+)^{1/2}$ and $\sigma_{\text{Integrated}}$ is computed. The Monte Carlo simulations are computed for an equivalent length of 42 years (with 260 business days per year).

3.2 Testing for efficiency

The interpretation of the simulations is fairly involved because several competing factors influence the estimation error of the different estimators. First, increasing the aggregation factor k used for measuring the return r_k k-ticks apart is an efficient way to reduce the incoherent term. This is based on the different aggregation properties of a random walk and the incoherent term, and taking larger k makes smaller the incoherent component. Second, when increasing k, the dependency between $r_k(j)$ and $r_k(j+1)$ increases, and therefore the effective number of independent terms in the sum diminishes. These two competing factors lead to an optimal value for k. Third, the incoherent filter applied to the price series is a powerful way to reduce the incoherent noise. As the variance of $\sum r_k^2$ is directly influenced by the level of the incoherent component, reducing the incoherent component reduces the variance of the Zhou estimator. Fourth, the filter is correctly specified only for a random walk with constant volatility, which is not the case for GARCH processes or for empirical data. Therefore, even after filtering, correcting again for the incoherent component in the volatility estimator helps in reducing the estimation error. All these four competing factors influence together the results of the simulations, making the overall picture quite complex.

The standard deviation for the estimation error is given in table 1 for the three simulations below. The first simulation is done with a constant volatility $\sigma^2 = 1$ and with random time intervals with $\text{E}[\delta t]$ = 5 minutes. Because the volatility is constant, the realized volatility depends only on the number of ticks in a day. The average number of ticks is large (288), and therefore the integrated volatility is essentially constant (its standard deviation is 0.03) With this process, the pdf for the estimation error is essentially a translation of the pdf for the various estimators. The estimated probability densities for a selection of the estimators are given in Fig. 1 (the pdf is estimated with an histogram, using a linear interpolation for computing the weights in the two adjacent bins). The very poor properties of the squared return $|r[1d]|^2$ as a daily volatility estimator are obvious from this figure. Therefore, it is very difficult for example to assess the quality of daily volatility forecasts by GARCH(1,1) using only daily data [Andersen and Bollerslev, 1998]. The RiskMetric estimator is good, but as there is no dynamic on the volatility, its properties are not realistic compared to the behavior of the RiskMetric estimator found on empirical data. For the (unfiltered) RTS estimator based on 30' returns, the bias is obvious when looking at the figure. This is because the incoherent noise is added at the 5' time horizon, and there is not much aggregation between 5' and 30' to lower the incoherent contribution. The filtered RTS estimator based on 5' returns is a very good estimator. On the other hand, the Zhou σ^2_{Zhou} and Zhou with long covariance $\sigma^2_{\text{Zhou+LC}}$ estimator with $k = 4$ have a large variance. For $k = 1$, the variance is even larger, and the probability to obtain negative values cannot be neglected. As already noted by [Zhou, 1996], increasing k helps in reducing the variance (see table 1), and therefore the probability to have negative volatility. But clearly, the best estimators are the filtered Zhou volatility and the quadratic variations, both with $k = 1$.

The second simulation is performed with a GARCH(1,1) dynamic for the volatility, and with random time intervals between ticks, with $\text{E}[\delta t] = 5'$. The estimated pdf for the estimation error is given in Fig. 2. The

	Student RW $\mathrm{E}[\delta t] = 5'$	GARCH(1,1) $\mathrm{E}[\delta t] = 5'$	GARCH(1,1) $\mathrm{E}[\delta t] = 30''$
$\lvert r[1d]\rvert^2$	1.49	1.63	1.42
RiskMetric	0.263	0.675	0.308
$\sigma^2_{\mathrm{RTS}}[1d, 30']$	0.370	0.391	0.224
Filter + $\sigma^2_{\mathrm{RTS}}[1d, 5']$	0.145	0.298	0.095
$\sigma^2_{\mathrm{Zhou}}[1d, k=1]$	0.443	0.454	0.140
$\sigma^2_{\mathrm{Zhou}}[1d, k=2]$	0.322	0.353	0.103
$\sigma^2_{\mathrm{Zhou}}[1d, k=4]$	0.324	0.380	0.105
$\sigma^2_{\mathrm{Zhou}}[1d, k=8]$	0.411	0.496	0.133
$\sigma^2_{\mathrm{Zhou}}[1d, k=16]$	0.561	0.669	0.180
$\sigma^2_{\mathrm{Zhou+LC}}[1d, 10d, k=1]$	0.477	0.479	0.152
$\sigma^2_{\mathrm{Zhou+LC}}[1d, 10d, k=2]$	0.300	0.400	0.096
$\sigma^2_{\mathrm{Zhou+LC}}[1d, 10d, k=4]$	0.241	0.276	0.078
$\sigma^2_{\mathrm{Zhou+LC}}[1d, 10d, k=8]$	0.259	0.329	0.084
$\sigma^2_{\mathrm{Zhou+LC}}[1d, 10d, k=16]$	0.332	0.401	0.107
Filter + $\sigma^2_{\mathrm{Zhou}}[1d, k=1]$	0.169	0.230	0.073
Filter + $\sigma^2_{\mathrm{Zhou}}[1d, k=2]$	0.216	0.252	0.070
Filter + $\sigma^2_{\mathrm{Zhou}}[1d, k=4]$	0.290	0.347	0.094
Filter + $\sigma^2_{\mathrm{Zhou}}[1d, k=8]$	0.400	0.484	0.130
Filter + $\sigma^2_{\mathrm{Zhou}}[1d, k=16]$	0.558	0.665	0.178
Filter + $\sigma^2_{\mathrm{QV}}[1d, k=1]$	0.121	0.362	0.121
Filter + $\sigma^2_{\mathrm{QV}}[1d, k=2]$	0.133	0.274	0.093
Filter + $\sigma^2_{\mathrm{QV}}[1d, k=4]$	0.161	0.221	0.070
Filter + $\sigma^2_{\mathrm{QV}}[1d, k=8]$	0.212	0.251	0.071
Filter + $\sigma^2_{\mathrm{QV}}[1d, k=16]$	0.288	0.339	0.093

Table 1: Standard deviation of the estimation error associated with the various estimators. The mean of the estimation error is always negligible, except for $\sigma^2_{\mathrm{RTS}}[1d, 30']$ with values 0.62, 0.62, and 0.044.

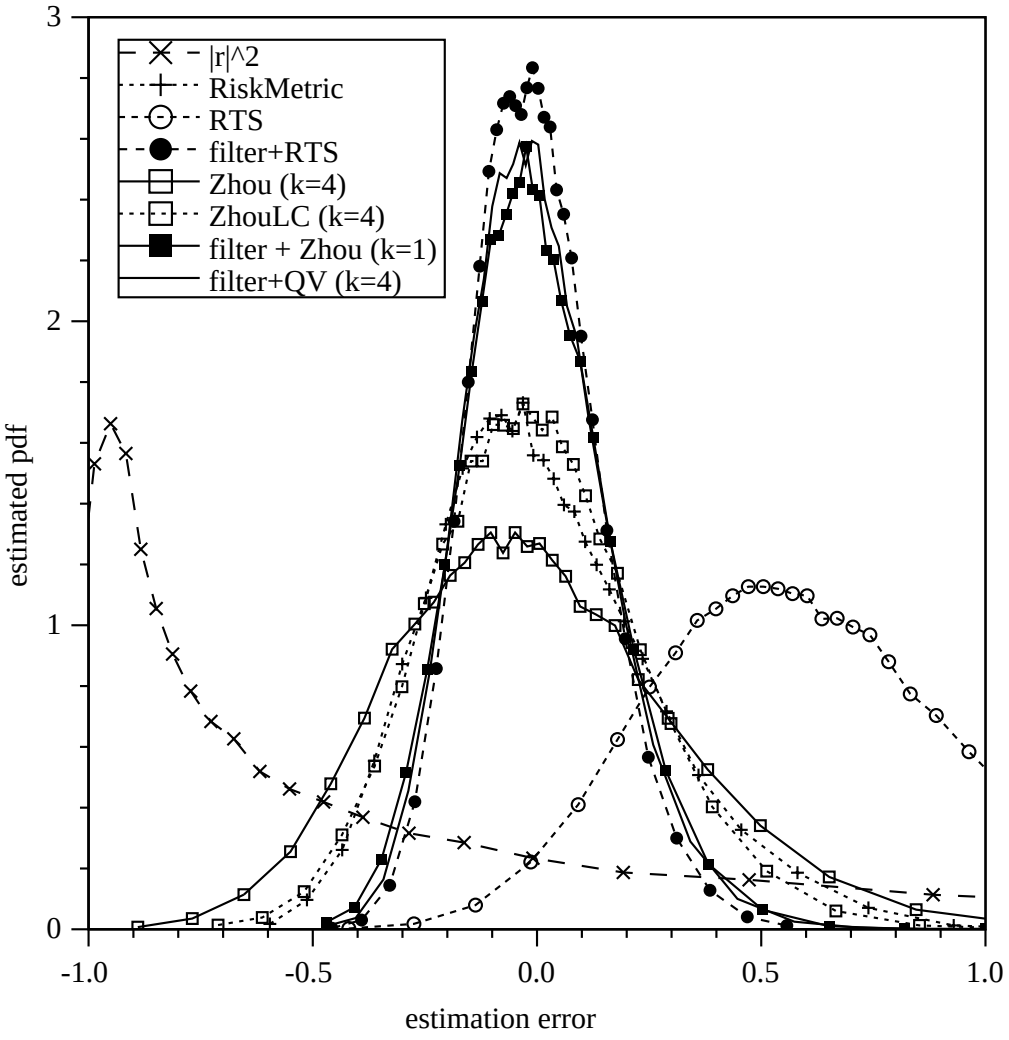

Figure 1: Estimated probability density function (pdf) of the measurement error for a Student-t random walk (constant volatility). The time interval between the prices is random with $\mathrm{E}[\delta t] = 5'$.

major difference with the previous simulation is the larger variance for all the high frequency estimators. This is due to the fact that we are computing the estimation error (and not the *relative* estimation error), and the errors can be larger for large integrated volatilities. Beside, in this particular case, RiskMetric is quite optimal for an estimator using daily data. This good property of RiskMetric originates in the near coincidence of the RiskMetric parameter ($\simeq$ 16 days) and the decay of the autocorrelation function of volatility for the simulated prices (10 day). Moreover, both the memory of the simulated data and the evaluation kernel of RiskMetric decay exponentially. Both coincidences make the RiskMetric estimator nearly optimal in comparison to other simulation setups, like for example with processes including long memory of the volatility. The best estimators are again either filter + Zhou or filter + quadratic variations. Notice that when the volatility is not constant, the filter as given in [Corsi et al., 2001] is partly misspecified as it assumes a volatility per tick changing at time scales longer than the filter time horizon (in the present case, 10 days). Therefore, part of the incoherent noise remains in the tick-by-tick data. This is why Zhou with $k = 1$ shows better properties than the quadratic variations with $k = 1$. Finally, at this tick frequency, the standard deviation of the best high frequency estimator is three times smaller than RiskMetric.

The third simulation is performed with a GARCH(1,1) dynamic for the volatility, and with random time intervals between ticks with $\mathrm{E}[\delta t] = 30''$, namely a 10 fold increase in the tick frequency. The estimated pdf for the estimation error is given in Fig. 3. The unfiltered RTS volatility shows no visual bias on the graph,

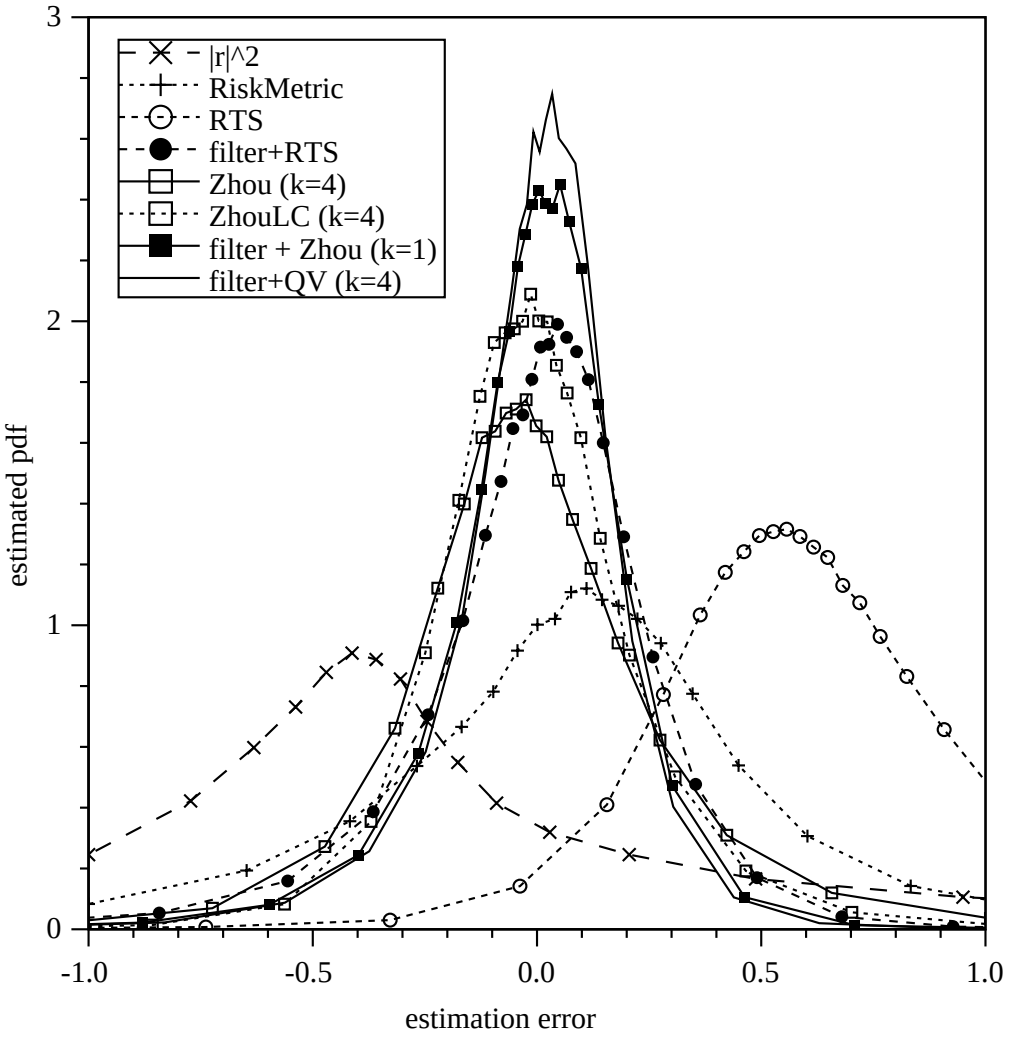

Figure 2: Estimated probability density function (pdf) of the estimation error for a GARCH(1,1) process. The time interval between the prices is random with $\mathrm{E}[\delta t] = 5'$.

and its computed mean estimation error is 0.016. This is due to the aggregation between 30 seconds and 30 minutes that is enough to scale down the incoherent noise. This small bias can be safely neglected. At this tick rate, all the high frequency volatility estimators are clearly better than RiskMetric, with a reduction of the standard deviation up to a factor of 4. The relative performances of the high frequency estimators are similar to the one obtained at 5', with the best estimator being filter + Zhou and filter + quadratic variations.

The correlation between σ^2 and $\sigma^2_{\text{integrated}}$ is given in table 2 for both GARCH(1,1) simulations above (the realized volatility is almost constant for the Student RW, and the correlation is not well defined). For example, the correlation with RiskMetric is of the order of 35% (50% for the 5' simulation), whereas high frequency definitions have correlations in the range 90% to 95%. Table 3 gives the probability to obtain a negative value for the volatility estimators that do not enforce positivity. These probabilities can also be estimated for empirical data, and we have added the values corresponding to USD/CHF measured over 10 years (from 1.1.1991 to 1.1.2001) for daily and hourly volatilities. For the daily estimate, the volatility is measured with 24 hours of data (i.e. in physical time and not in a business time scale removing the week-end). The values are taken every 8 hours, removing the week-ends (from Friday 21:30 to Sunday 21:30 GMT), but keeping holidays. For the hourly estimate, the measure is taken every hour, also removing week-ends. The large difference between USD/CHF and GARCH(1,1) with $\mathrm{E}[\delta t] = 30''$ indicates that the simulated process does not reproduce well the empirical data at this frequency. At least the seasonality is

	GARCH(1,1) $\mathrm{E}[\delta t] = 5'$		GARCH(1,1) $\mathrm{E}[\delta t] = 30''$	
	ρ	R^2	ρ	R^2
$\lvert r[1d]\rvert^2$	43	0.19	17	0.03
RiskMetric	52	0.28	35	0.11
$\sigma^2_{\mathrm{RTS}}[1d, 30']$	88	0.78	73	0.53
$\sigma^2_{\mathrm{RTS}}[1d, 5']$	92	0.85	92	0.85
$\sigma^2_{\mathrm{Zhou}}[1d, k=1]$	86	0.74	87	0.76
$\sigma^2_{\mathrm{Zhou}}[1d, k=2]$	91	0.83	92	0.85
$\sigma^2_{\mathrm{Zhou}}[1d, k=4]$	89	0.80	92	0.85
$\sigma^2_{\mathrm{Zhou}}[1d, k=8]$	82	0.68	88	0.77
$\sigma^2_{\mathrm{Zhou}}[1d, k=16]$	75	0.58	81	0.65
$\sigma^2_{\mathrm{Zhou+LC}}[1d, 10d, k=1]$	85	0.72	85	0.73
$\sigma^2_{\mathrm{Zhou+LC}}[1d, 10d, k=2]$	93	0.86	93	0.87
$\sigma^2_{\mathrm{Zhou+LC}}[1d, 10d, k=4]$	94	0.88	95	0.91
$\sigma^2_{\mathrm{Zhou+LC}}[1d, 10d, k=8]$	91	0.83	95	0.89
$\sigma^2_{\mathrm{Zhou+LC}}[1d, 10d, k=16]$	88	0.77	92	0.84
Filter + $\sigma^2_{\mathrm{Zhou}}[1d, k=1]$	95	0.90	95	0.91
Filter + $\sigma^2_{\mathrm{Zhou}}[1d, k=2]$	95	0.90	96	0.92
Filter + $\sigma^2_{\mathrm{Zhou}}[1d, k=4]$	91	0.82	93	0.87
Filter + $\sigma^2_{\mathrm{Zhou}}[1d, k=8]$	83	0.69	88	0.78
Filter + $\sigma^2_{\mathrm{Zhou}}[1d, k=16]$	75	0.56	81	0.65
Filter + $\sigma^2_{\mathrm{QV}}[1d, k=1]$	88	0.77	86	0.94
Filter + $\sigma^2_{\mathrm{QV}}[1d, k=2]$	93	0.87	92	0.85
Filter + $\sigma^2_{\mathrm{QV}}[1d, k=4]$	95	0.91	96	0.91
Filter + $\sigma^2_{\mathrm{QV}}[1d, k=8]$	94	0.89	95	0.91
Filter + $\sigma^2_{\mathrm{QV}}[1d, k=16]$	90	0.82	93	0.86

Table 2: Linear correlation ρ (in %) and goodness of fit R^2 between σ^2 and $\sigma^2_{\mathrm{integrated}}$.

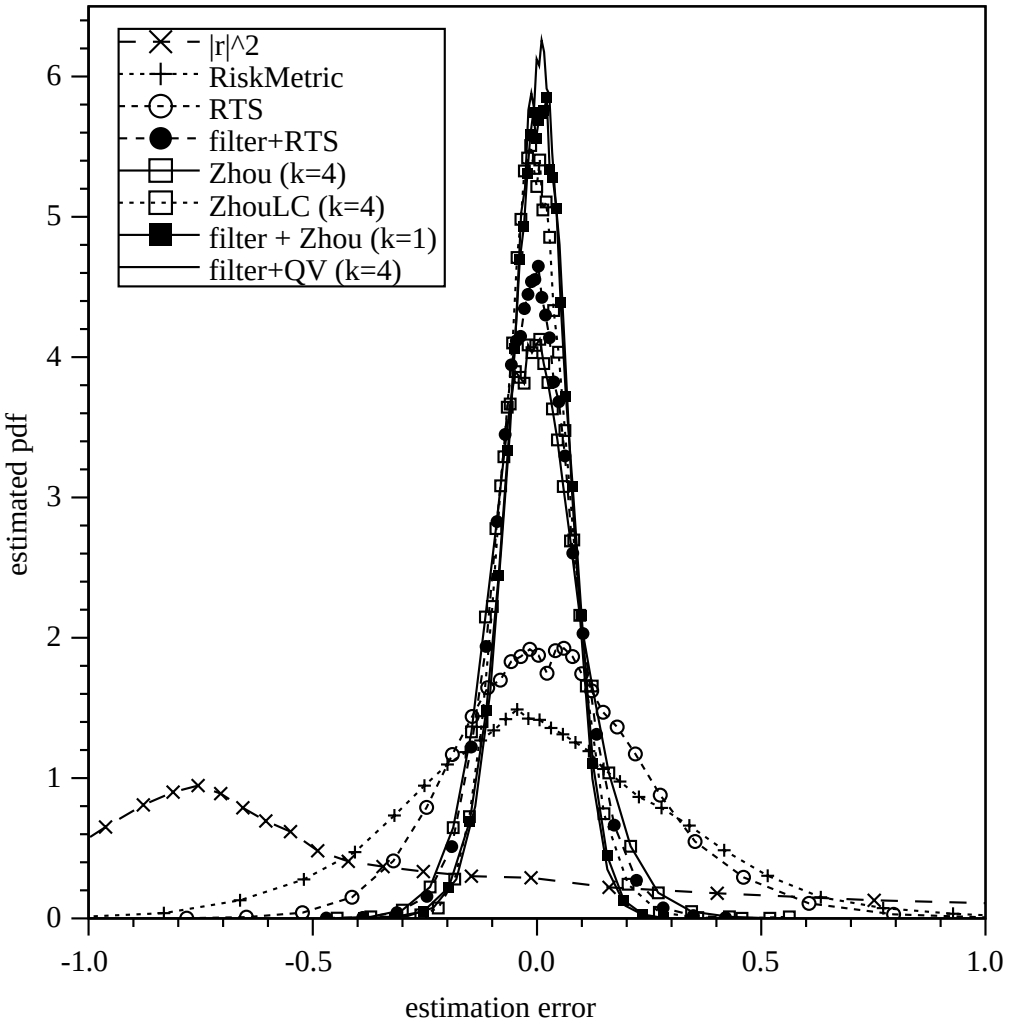

Figure 3: Estimated probability density function (pdf) of the estimation error for a GARCH(1,1) process. The time interval between the prices is random with $\mathrm{E}[\delta t] = 30''$.

missing, but the dynamic of GARCH(1,1) is also not very close to the known empirical properties. The most interesting result regarding the USD/CHF data is the very large probability for negative values for $\sigma^2_{\text{Zhou+LC}}$ at small k. Although this estimator is good on Monte Carlo simulations, this is a very strong argument against using this estimator in practice. On the other hand, the estimator Filter + $\sigma^2_{\text{Zhou}}[1d, k = 1]$ has a sufficiently low probability to be acceptable (1 point for 7818 values for daily volatility, occurring on a Sunday at 00:00). The hourly volatilities estimated from empirical data produce more negative values, as expected from the smaller number of ticks available for the computation. Yet, most of the negative values occur during the opening of the Asiatic market (i.e. between 23:00 and 06:00 GMT), where the volatility and tick rate are low ($\simeq$ 10 tick/hour).

3.3 Simulation for the "week horribilis"

The major salient properties of empirical data are the strong daily and weekly seasonalities induced by the cycle of human activities. In order to test the robustness of the various estimators with respect to rapid changes of the volatility pattern, we simulated an artificial "week horribilis" which includes strong patterns in the volatility and the tick rate. The simulations are done with a simple Student random walk, with a volatility of $\sigma^2 = 1$. The time intervals between ticks are random, with a reference expected tick rate of 12

	Student RW $E[\delta t] = 5'$	GARCH(1,1) $E[\delta t] = 5'$	GARCH(1,1) $E[\delta t] = 30''$	USD/CHF Daily vol	USD/CHF Hourly vol
$\sigma^2_{\text{Zhou}}[k=1]$	1.3	6.0	0.0	3.0	13.8
$\sigma^2_{\text{Zhou}}[k=2]$	0.03	1.3	0.0	0.5	6.9
$\sigma^2_{\text{Zhou}}[k=4]$	0.0	0.17	0.0	0.7	5.7
$\sigma^2_{\text{Zhou}}[k=8]$	0.01	0.04	0.0	1.0	7.8
$\sigma^2_{\text{Zhou}}[k=16]$	0.3	0.4	0.0	1.2	11.2
$\sigma^2_{\text{Zhou+LC}}[10d, k=1]$	1.3	7.3	0.0	29.3	40.6
$\sigma^2_{\text{Zhou+LC}}[10d, k=2]$	0.01	1.5	0.0	11.5	28.9
$\sigma^2_{\text{Zhou+LC}}[10d, k=4]$	0.0	0.15	0.0	3.4	10.0
$\sigma^2_{\text{Zhou+LC}}[10d, k=8]$	0.0	0.03	0.0	1.1	11.1
$\sigma^2_{\text{Zhou+LC}}[10d, k=16]$	0.0	0.1	0.0	0.2	5.0
Filter + $\sigma^2_{\text{Zhou}}[k=1]$	0.0	0.0	0.0	0.01	0.2
Filter + $\sigma^2_{\text{Zhou}}[k=2]$	0.0	0.0	0.0	0.08	0.5
Filter + $\sigma^2_{\text{Zhou}}[k=4]$	0.0	0.0	0.0	0.3	2.5
Filter + $\sigma^2_{\text{Zhou}}[k=8]$	0.0	0.0	0.0	0.7	6.5
Filter + $\sigma^2_{\text{Zhou}}[k=16]$	0.2	0.3	0.0	1.2	10.9

Table 3: Probability of negative values for σ^2, in %. The first three columns are for daily volatilities, the last one for hourly volatility.

tick/hour ($E[\delta t] = 5'$). An incoherent term is added to the prices with an intensity of $z^2 = 2$. The volatilities are measured hourly, and therefore have a very small number of ticks to work with. The set up for this test week is given in the upper graph of Fig. 4, and is as follows: On day one the volatility and the tick rate are simply constant. Day two simulates a gap, namely the underlying process runs regularly with constant volatility while no data is provided to the volatility estimators. As a result, there are potentially large jumps in the price at the opening of day three. Day three tests for a large change in volatility, while the tick rate is constant. To achieve this pattern, the volatility per tick needs to be increased accordingly. Day four simulates a "dead market", namely no tick is provided and the market reopens at the closing price. We introduce this notion in order to differentiate from a night or a week-end. Particularly for stocks and stock indexes, there is new arrivals while the market is closed, leading to an opening price that can be quite different from the previous closing price. Therefore, a real night or a week-end is a combination of a dead market and a gap and is simulated in day six. Day five shows a constant volatility while the tick rate is increased by a factor four. Within this day, the volatility per tick is inversely proportional to the tick rate in order to obtain a constant volatility. Notice that during day three and five, the incoherent filter is misspecified as it assumes a constant noise per volatility. Day six tests for a real week-end, with a gradual reopening of the market, including a slowly increasing tick rate. Finally, day seven shows a simultaneous pattern for the realized volatility and the tick rate. Given the set of patterns as well as the low tick rate and the high incoherent noise, the parameters are quite extreme and provide for a harsh test of the estimators: therefore the name "week horribilis".

The results are displayed in the lower graph of Fig. 4. Overall, the various volatility estimators reproduce well the realized volatility, particularly taking into account the low tick rate and high incoherent noise. On the days three and five, the "filter + quadratic variation" is not able to measure correctly the realized volatility. This is due to the misspecification of the incoherent filter which is not corrected by this simple estimator. On the other hand, the estimator "filter + Zhou" with $k = 4$ measures correctly the realized volatility, as the Zhou estimator will correct for the misspecification of the filter. The correction is not perfect as a misspecified incoherent filter induces an exponential lagged correlation [Corsi et al., 2001], whereas only the first lag (at the given k value) is taken into account in the Zhou estimator. The estimator "filter + Zhou" with $k = 1$ gives similar results as "filter + quadratic variation" at $k = 4$. Nevertheless, given

the quite extreme set-up of the week horribilis, the "filter + Zhou" provides very accurate values.

The data gap on day two deserves a special discussion. The question is whether a "real time" or "historical" behavior is desired. In real time, the value at a given time point needs to be given immediately at this time point (causal estimator), whereas for a historical computation, the values of subsequent ticks can be used (non causal estimator). Therefore, by only using past information, it is not possible to make the difference without extraneous informations between a normal market closure (e.g. a week-end), a non regular market closure (e.g. a Holiday) or a gap (e.g. a technical failure). If a causal estimator, e.g., returns a zero value inside a gap (our choice), then at the end of a gap a large volatility spike appears due to the large price jumps. This behavior is imposed by the causality, and all estimators will suffer similarly. On the other hand, if a historical (non causal) estimator can be used, subsequent values provide a hint that a gap went by. Such an estimator is given with the label "Zhou+BC", and has been introduced in section 2.4.5. Fig. 4 shows that the gap correction is working perfectly on day two. Moreover, the "week-end" in day six, composed of a dead market and a gap, shows exactly the same behavior, although to a smaller extend. This emphasizes the importance of a correct behavior in a gap, as night and week-end are recurring events. At this point, the choice of real time or historical estimators is imposed by other external constraints. The important point is that the gap correction is working perfectly, provided that a historical estimator can be used.

In order to check the dependency with respect to the tick rate, the same set-up has been used, except for the higher tick frequency $\mathrm{E}[\delta t] = 30''$ instead of $\mathrm{E}[\delta t] = 5'$. Essentially the same results are obtained, which confirm the above analysis.

3.4 A first summary

The main results from these simulations are:

- The best high-frequency estimators clearly outperform the estimators using only daily data.
- σ^2_{Zhou} is not a good estimator as its variance is too large (for all k).
- $\sigma^2_{\mathrm{Zhou+LC}}[k=4]$ is an estimator with a small variance, but the probability to obtain negative values with empirical data is far too large.
- Filter + $\sigma^2_{\mathrm{Zhou}}[1d, k=1]$ (or $k=2$) is an efficient estimator, with a small enough probability to obtain negative values when using empirical data. Moreover, the Zhou estimator corrects (at the first lag) for the misspecification of the incoherent filter.
- Filter + $\sigma^2_{\mathrm{QV}}[1d, k=4]$ is efficient and positive. On the down side, this estimator does not correct for the misspecification of the incoherent filter, and the optimal value for k may change with the process used for the simulations (and is unknown for empirical data).

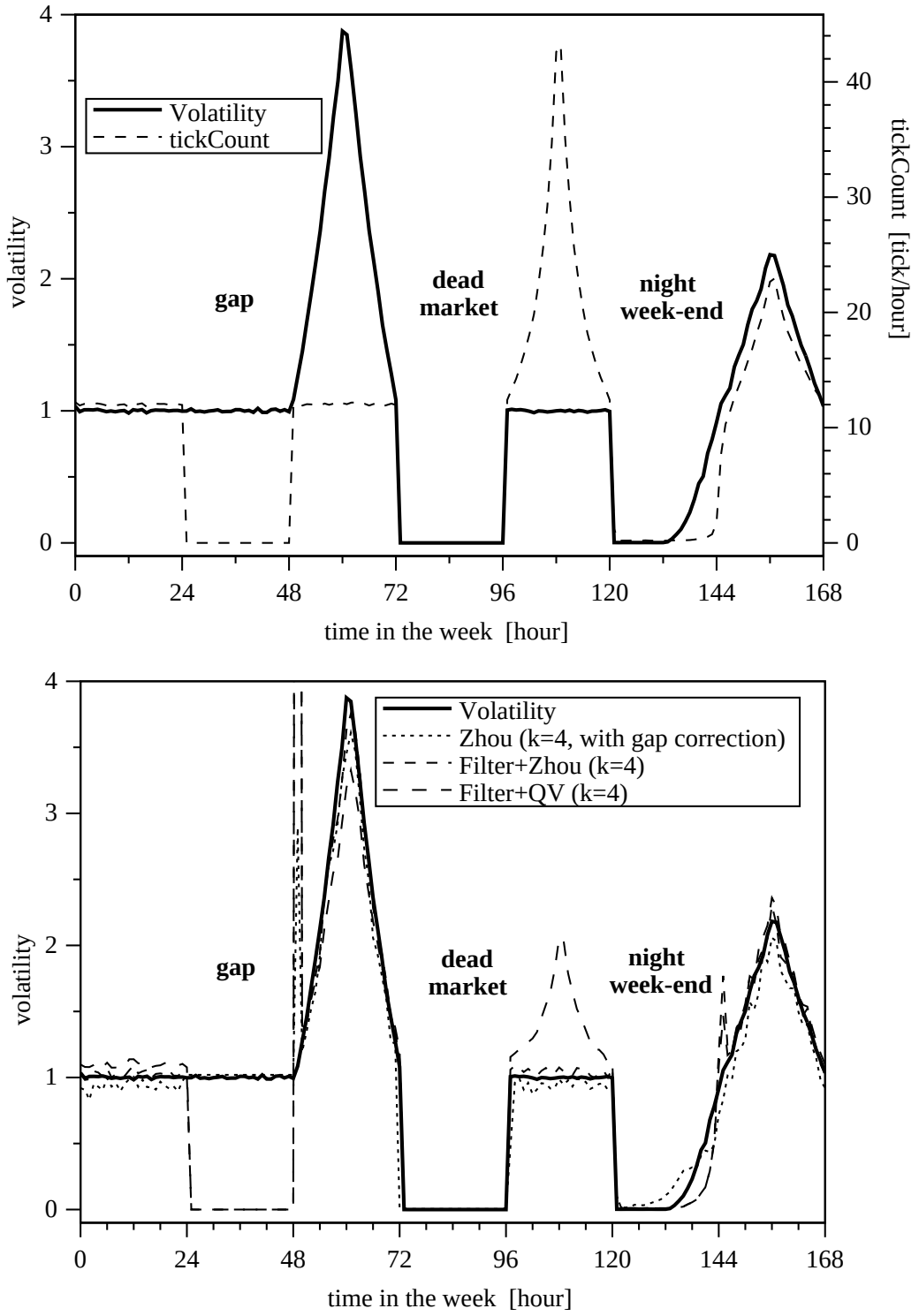

Figure 4: The simulation for the "week horribilis". The upper figure shows the volatility and the tick rate measured from the simulation. The lower figure displays the mean volatility according to a few selected estimators. The parameters for the simulation are given in the text.

4 Volatility estimates on empirical data

The relevance of the Monte Carlo simulations presented in the last section obviously depends on how good the model used to simulate the price process duplicates the properties of empirical data. Yet, a direct comparison is impossible precisely because the volatility of the empirical data is an unknown quantity. As a direct check of the model is not possible, we have to verify that the assumptions used in the price process, as defined in section 2.1, agree with the corresponding properties of empirical data.

An assumption used to model the incoherent term is that its intensity, as measured by η, is independent of the current level of volatility as measured by σ. The rational for this hypothesis is that the incoherence is bounded by the spread, and therefore η must be a fraction of the spread. On the other hand, one can expect that the incoherence, as it measures the level of disagreement between the market participants, must be related to the current level of volatility.

A check for the constant η hypothesis is given by an intra week conditional average. The idea is to use the strong daily pattern to measure the average variance, covariance and volatility as a function of the time in the week. More precisely, we compute the following quantities

$$\text{Var}_{\text{per tick}}(t) = \frac{1}{n[\Delta t](t)} \sum_{t-\Delta t \le t(j) \le t} r_k^2(j) \tag{27}$$

$$\text{Std.Dev.}_{\text{per tick}} = 10000 \sqrt{\text{E}[\text{Var}_{\text{per tick}}(t) \mid t']} \tag{28}$$

where n is the number of ticks in the interval Δt (Eq. 16). The expectation is taken conditional to the time in the week $t' = t \mod 1\text{week}$. The factor 10000 expresses the standard deviation in basis points, a natural unit for "per tick" quantities. We proceed similarly for the covariance per tick and the volatility per tick (according to the Zhou definition). Essentially, the covariance per tick gives an estimate for η^2 and the volatility per tick an estimate for σ^2. The results are displayed in Fig. 5, using 5 years (1996 to 2000) of the FX rate USD/CHF. The number of ticks shows a very strong seasonality, with a factor 30 between the peak ($\sim$ 300 tick/hour during European and American market opening), and the trough ($\sim$ 10 tick/hour during the Asiatic opening). The low number of ticks during the nights induces poor statistics during these periods, and is visible in the larger fluctuations in the other curves. The volatility per tick, a measure of σ, has an inverse behavior. This is due to the lower coverage of the Asiatic market by Reuters, leading to a higher price change "per Reuters tick". The ratio between crest and trough is ~ 3 for the volatility per tick. As volatility is the product of the volatility per tick with the number of ticks, this leads to the usual intra-week volatility pattern, with a factor of ~ 10 for the volatility between European/American and Asiatic markets. The seasonality of σ is not included in the simplest model for the price process, as given in Eq. 2. It is trivial to include a volatility seasonality, and this issue has been already explored in the Monte Carlo simulations with the "week horribilis". This shows the importance of a volatility estimator which is robust against changes in σ. Moreover, the incoherent filter is not correctly specified (as a constant volatility and correlation are assumed). This is an argument for not using "filter + quadratic variation", but instead "filter + Zhou" as this will correct (partly) for the misspecification of the incoherent filter. Beside, as the correlation is the ratio of the covariance with the variance, the seasonality of σ induces a seasonality in the correlation (see the discussion at the end of section 2.4.2).

The covariance per tick, a measure of η, is essentially constant, with a value of 2 basis points per tick. This constant η justifies the model (Eq. 2), in line with the argument that the incoherent term is bounded by the spread ($\sim$ 10 basis point on Reuters). During the opening of the European and American markets, where the statistics are better, one can also see a smaller modulation proportional to the volatility (or to the tick rate). The amplitude of this modulation is of the order of 0.1 basis points, indicating that there is a small influence on the incoherence induced by the current level of volatility. Yet, this modulation is an order of magnitude smaller, enough to be neglected in first approximation. A visual study of the scatter plot for the covariance

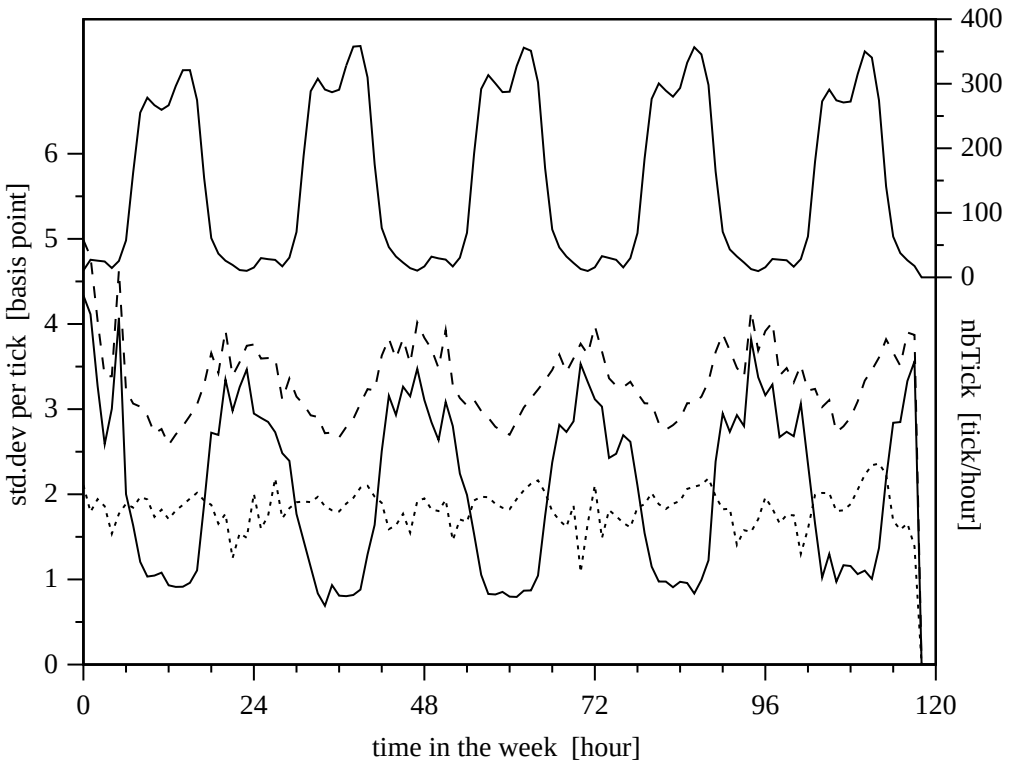

Figure 5: The average hourly volatility per tick conditional to the time in the week, in basis points, for USD/CHF. The averages are computed from 1.1.1996 to 1.1.2001, and the week-ends are omitted from the graph. The upper continuous curve gives the number of ticks (right scale), the dashed curve the standard deviation per tick, the dotted curve the square root of minus the covariance per tick, and the lower continuous curve the volatility per tick (left scale). All the quantities are measured on an hourly basis $\Delta t = 1$ hour, with one tick apart returns r_1 (i.e. $k = 1$).

per tick versus the volatility per tick (or versus the volatility, or versus the number of ticks) for hourly quantities, conditional to a given time in the day (to remove the seasonality), confirms the independence of the incoherent term with respect to the other variables. During the high tick rate period, the volatility per tick is of the order of 1 basis point whereas the incoherent term is of the order of 2 basis points. This shows again that the incoherent contribution dominates the usual random walk component at the tick time scale.

Until this point, we have been concerned with definitions and statistics related to the various estimators. As we have a good idea of their respective merits and weaknesses, we can apply them to actual data. Fig. 6 shows the time evolution of the volatility in May 2001 as measured by some selected volatility estimators. In order not to clutter the graph, only three volatility estimators have been drawn, but essentially all the high frequency volatility definitions give consistent values. Beside, the slow dynamic of RiskMetric is very clear from this graph.

5 Conclusion

Although it is common to talk about *the* volatility, there is no single universally accepted definition of volatility. Instead, as we have shown in section 2, several different estimators can be used to measure the fluctuations of the prices. Simple statistical considerations indicate that we should use tick-by-tick definitions, like the Zhou volatility estimator. Moreover, these definitions do not need an auxiliary regular time series at an arbitrary time scale δt. The tick-by-tick volatility estimators correspond to the best possible approximation of the limit $\delta t \to 0$, given the available information. Indeed, as investigated with Monte Carlo simulations, the advantage of tick-by-tick definitions is clear. As an order of magnitude, the variance of the

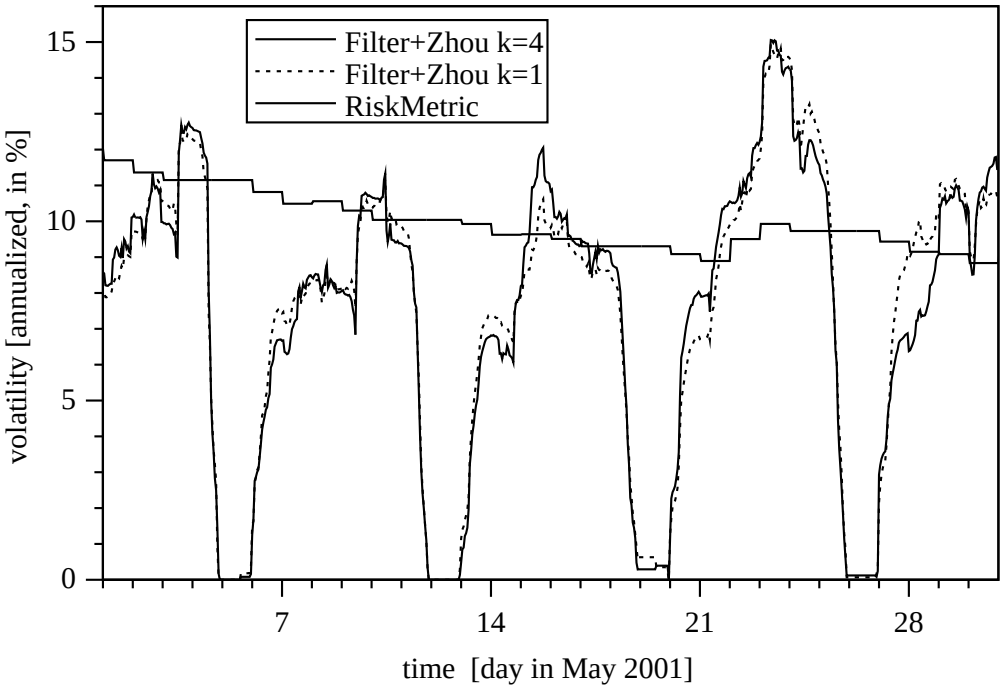

Figure 6: The daily volatility time evolution in May 2001. The two high frequency volatility estimators are filter + $((\sigma^2_{\text{Zhou}}[k=1])^+)^{1/2}$ (dotted curve) and filter + $((\sigma^2_{\text{Zhou}}[k=4])^+)^{1/2}$ (continuous line). The high frequency volatility estimator uses a moving sample of 24 hours, 7 days a week. The plot makes apparent the strong seasonality due to the week-end, but the daily seasonality is completely absorbed by the daily window.

estimation error is reduced by a factor 4 when comparing RiskMetric (daily data) with the estimator "filter + Zhou" (high frequency data).

Yet, it is crucial to discount for the incoherent term induced by the price formation. If not, the volatility is largely overestimated, or biased, compared to measures at the daily time scale. The Monte Carlo simulations indicate that the current best available estimator is "filter + Zhou", namely to filter the prices with an incoherent filter according to [Corsi et al., 2001], followed by a volatility estimator as given by Zhou [Zhou, 1996]. This combination has the advantage to have a small variance, a low probability to give negative values, and to correct for the misspecification of the filter. Another good choice is "filter + QV" (filter + quadratic variation). It is simple, is always positive, and its weakness with respect to the seasonalities are not essential for computing daily volatility. These choices are certainly not the last word: better volatility estimates will allow for a better characterization of the financial tick-by-tick processes. In turn, more refined models will allow for more realistic Monte Carlo simulations and more stringent tests of volatility estimators.

All the volatility estimators investigated in this work are essentially quadratic estimators. Because of the fat tail distribution of returns and of the possible outliers, the present estimators are quite fragile, and the original market data must be carefully filtered before using such tick-by-tick estimators. On the other hand, it is desirable to have more robust estimators. A first step in this direction could be to use formulae based on the sum of absolute values for the returns, like $\sigma_{L1} = \sum |r|$. Yet, the construction of unbiased, efficient and robust estimators is an open problem in the present setting (i.e. in the presence of an incoherent component).

In this work, we have investigated the high frequency volatility estimators of traded assets. To various degrees, the incoherence in the price formation should manifest itself on each "traded" time series. The case of stock indexes is different, as already analyzed in [Corsi et al., 2001], because they are computed quantities. In particular, the lead-lag structure between different stocks induces a more complex lagged correlation function for the indexes, and the simple EMA filter cannot be used. Although the extension of volatility estimators to this case is fairly straight forward (in the Zhou definition, more lagged covariance

terms can be included), the optimal choice of the volatility estimator is still an open problem. For example, an estimator which takes into account the shape of the lagged correlation may give better results than a purely non parametric estimator that includes all the lagged covariances up to some cut-off. Another interesting direction is to revisit the "Epps" effect [Epps, 1979]. This effect denotes the decreasing correlation between the returns $r_1[\delta t]$ and $r_2[\delta t]$ of two different time series when the return time interval δt decreases. Obviously, some time is needed for the market participants to build the appropriate correlation, but the incoherent price formation on both time series should also contribute to the lower correlation at high frequency. It would be interesting to compare the correlation between the raw time series, and between the time series filtered for the incoherent component.

The implications of a better measure for the volatility are far reaching. In broad terms, a good tick-by-tick volatility estimator enlarges our information set about a given time series. This will lead to better forecasts, both because the information set in the past is better, and because the integrated volatility to be forecasted is known accurately. In turn, this will lead to better risk management, portfolio optimization or option pricing.

References

[Andersen and Bollerslev, 1998] Andersen, T. G. and Bollerslev, T. (1998). Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *International Economic Review*, 39:885–905.

[Andersen et al., 2001a] Andersen, T. G., Bollerslev, T., Diebold, F. X., and Labys, P. (2001a). Great realizations. *Risk*, 13:105–108.

[Andersen et al., 2001b] Andersen, T. G., Bollerslev, T., Diebold, F. X., and Labys, P. (2001b). Modeling and forecasting realized volatility. *NBER Working Paper 8160. The National Bureau of Economic Research, Cambridge, MA.*

[Andrews and Monahan, 1992] Andrews, D. W. K. and Monahan, J. C. (1992). An improved heteroskedasticity and autocorrelation consistent covariance matrix estimator. *Econometrica*, 60(4):953–66.

[Corsi et al., 2001] Corsi, F., Zumbach, G., Müller, U. A., and Dacorogna, M. (2001). Consistent high-precision volatility from high-frequency data. *Economic Notes*, 30(2):183–204.

[Epps, 1979] Epps, T. (1979). Comovements in stock prices in the very short run. *Journal of the American Statistical Association*, 74(366):291–298.

[Moody and Wu, 1997] Moody, J. and Wu, L. (1997). What is the true price? – state space models for high frequency fx rates.

[Moody and Wu, 1998] Moody, J. and Wu, L. (1998). High frequency foreign exchange rates: Price behavior analysis and 'true price' models.

[Zhou, 1996] Zhou, B. (1996). High-frequency data and volatility in foreign-exchange rates. *Journal of Business & Economic Statistics*, 14(1):45–52.

[Zumbach and Müller, 2001] Zumbach, G. O. and Müller, U. A. (2001). Operators on inhomogeneous time series. *International Journal of Theoretical and Applied Finance*, 4(1):147–178.