FORECASTING VOLATILITY USING HIGH-FREQUENCY DATA

Giovanni De Luca

Dipartimento di Statistica e Matematica per la Ricerca Economica, Università di Napoli Parthenope.
e-mail: giovanni.deluca@uniparthenope.it

Abstract

This paper deals with the analysis and the forecast of volatility using data characterized by a frequency higher than daily. The aim is to check the adequacy of the traditional ARCH modelling for high frequency returns as well as the use of them for better forecasts of the daily volatility. Finally, tick-by-tick data whose main feature is the irregular spacing in time are investigated.

Keywords: GARCH, high-frequency data, volatility.

1. INTRODUCTION

Volatility plays a central role in asset pricing, asset allocation and risk management. In the last decades modelling and forecasting volatility has become a rapidly growing interest research area. The class of Autoregressive Conditional Heteroskedasticity (ARCH) models has been introduced by Engle (1982). After more than twenty years and an impressive scientific production, these models still represent the most popular statistical instrument to capture the dynamics of asset return volatility and are largely used in all areas of finance by academics and practitioners alike.

Most of ARCH-type models are thought for application using daily or weekly (or even lower frequencies) data. However, in recent years, higher frequency data, e.g. with a hourly or 15-min frequency, have been available. This has allowed many researchers to shed a new light on the analysis of volatility.
The first issue is the adequacy of traditional GARCH modelling to high-frequency data for the prediction of the short-term volatility. Secondly, the effort has been addressed in understanding how the use of high-frequency data could be a support for the analysis carried out on daily (or lower frequency) data. The concept of realized volatility, introduced to evaluate the prediction of daily volatility carried out using daily returns, is a remarkable example. Finally, the availability of ultra-high-frequency (or tick-by-tick) data, that is data without a fix sampling frequency, has further moved the frontier in financial time series analysis.

On the whole, this review tries to answer to three simple questions:

1. Can ARCH-type models be successfully used for high-frequency data volatility?
2. Can high-frequency data help to better understand (that is model or predict) daily volatility?
3. Can ultra-high-frequency data help to discover features otherwise not capturable?

In the following pages, after briefly introducing the class of ARCH-type models (Section 2), we review some recent papers that can contribute to find answers to the three questions (Sections 3, 4 and 5). Some conclusions are reported in the last Section.

2. ARCH-TYPE MODELS

Let \( P_t \) be the price of an asset at day \( t \) and let \( r_t \) be the daily (log)-return, that is \( r_t = \log P_t - \log P_{t-1} \). ARCH-type models can be generally defined as

\[
\begin{align*}
  r_t &= \sqrt{h_t} \epsilon_t \\
  h_t &= f(r_{t-1}^2, r_{t-2}^2, \ldots, h_{t-1}, h_{t-2} \ldots)
\end{align*}
\]

(1)

with \( \epsilon_t \sim iid(0,1) \). The distribution of the returns conditionally on past information turns out to have a constant (null) mean and a time-varying variance equal to \( h_t \), that is \( \text{Var}(r_t|I_{t-1}) = h_t \), where \( I_{t-1} \) denotes the information up to time \( t - 1 \). This feature is known as conditional heteroskedasticity. The specific equation of the conditional variance characterizes the variety of models proposed in literature. The original ARCH(q) model (see Engle, 1982) considers

\[
h_t = \omega + \alpha_1 r_{t-1}^2 + \ldots + \alpha_q r_{t-q}^2.
\]
However, a high number of lags, that is a high number of parameters to estimate, is usually needed to have a satisfactory performance. A more parsimonious formulation is the Generalized ARCH (GARCH), proposed by Bollerslev (1986). In the general GARCH(p,q) model

\[ h_t = \omega + \alpha_1 r_{t-1}^2 + \ldots + \alpha_q r_{t-q}^2 + \beta_1 h_{t-1} + \ldots + \beta_p h_{t-p}, \]

but the key of the success is the simple and popular GARCH(1,1) where

\[ h_t = \omega + \alpha r_{t-1}^2 + \beta h_{t-1}, \]

with \( \omega > 0, \alpha, \beta \geq 0 \) and \( \alpha + \beta < 1 \). Empirical evidence has shown that a high number of financial time series can be satisfactorily described by a GARCH(1,1) model. As a result, this model is often seen as benchmark for novel formulations. Note that when \( \alpha + \beta = 1 \), we have the Integrated GARCH (1,1) model.

A possible refinement is given by the introduction in the mean equation of a function of the contemporaneous volatility. In fact, according to the most of financial theories, a direct relationship between risk and return holds. So the first equation of (1) becomes

\[ r_t = f(h_t) + \sqrt{h_t} \epsilon_t. \]

When the associated model for volatility is the GARCH, we have the GARCH-in-mean model.

In theory, we distinguish symmetrical and asymmetrical volatility models. ARCH and GARCH models belong to the former. A past return has the same effect on future volatility regardless of its sign. In the latter, negative past returns have a greater impact on future volatility than positive ones (this feature is also known as leverage effect, see Black, 1976). The Exponential GARCH (EGARCH) proposed by Nelson (1991) and the Threshold GARCH (TGARCH) proposed by Glosten et al. (1993) are popular examples. In former case, the conditional log-variance of the EGARCH(1,1) is given by

\[ \log h_t = \omega + \alpha r_{t-1} + \gamma (|r_{t-1}| - E(|r_{t-1}|)) + \beta \log h_{t-1}. \]

In this framework, the negative shocks have an impact equal to \( \alpha - \gamma \) on the log-variance, while for positive ones the impact is \( \alpha + \gamma \), which is smaller with \( \alpha < 0 \) and \( 0 \leq \gamma < 1 \).

In the latter, also known as GJR-GARCH, with the orders \( p = q = 1 \), we have

\[ h_t = \omega + \alpha r_{t-1}^2 + \beta h_{t-1} + \gamma r_{t-1}^2 d_{t-1} \]
where \( d_{t-1} \) is a dummy variable equal to 1 if the associated return is negative and 0 otherwise.

Further asymmetric representations include the Quadratic GARCH (Sentana, 1995), the Volatility-Switching GARCH (Fornari and Mele, 1997), the Asymmetric Nonlinear Smooth Transition GARCH (Anderson et al., 1999), the Logistic Smooth Transition GARCH (Gonzàles-Rivera, 1998), and the Asymmetric Power ARCH (Ding et al., 1993).

Forecast evaluations allow the comparison among competitive models. Unfortunately, the task is made difficult because of the latent nature of the volatility which is usually estimated ex-post through daily squared returns calculating from market closing prices. Popular measures (e.g. Mean Square Error, Mean Absolute Percent Error) are then evaluable. Furthermore, the evaluation of the performance of the daily volatility forecasts is made through a regression technique. The observed daily squared return is the ex-post volatility measure and represents the dependent variable while the explicative variable is given by the in-sample forecast. The regression model is then

\[
r^2_{t+1} = a + b\hat{\sigma}^2_{t+1} + \epsilon_{t+1}.
\]

If the forecasts of the model are correct, then the intercept and the slope of the regression should be, respectively, zero and one. An F-test is then carried out. Note that logarithmic transformations are usually encountered in order to reduce the impact of abnormal values.

Finally, volatility model performance can be evaluated exploiting measures of economic significance (e.g. portfolio or Value-at-Risk improvement according to the volatility forecasts). A comprehensive survey is in Poon and Granger (2003).

3. CAN ARCH-TYPE MODELS BE SUCCESSFULLY USED FOR HIGH-FREQUENCY DATA VOLATILITY?

ARCH-type models are usually found to be adequate for different frequencies returns. However, Diebold (1988) shows that conditional heteroskedasticity tends to disappear with the increasing of the sampling interval. What happens with higher frequency data?

A first contribution on the effect of the temporal aggregation is provided by Drost and Nijman (1993), henceforth DN, who distinguish strong, semi-strong, and weak GARCH processes. Recalling (1), a strong GARCH process is characterized by \( \epsilon_t = r_t/\sqrt{h_t} \sim iid(0,1) \), whereas for a semi-strong GARCH process we have \( E(r_t|r_{t-1}, r_{t-2}, \ldots) = 0 \) and \( E(r^2_t|r_{t-1}, r_{t-2}, \ldots) = \).
Finally, a GARCH process is weak if \( P(r_t | r_{t-1}, r_{t-2}, \ldots) = 0 \) and \( P(r_t^2 | r_{t-1}, r_{t-2}, \ldots) = h_t \), where \( P(\cdot) \) is the best linear predictor. DN showed that the weak GARCH(1,1) process is close under temporal aggregation. High-frequency parameters determine the corresponding low-frequency parameters and vice versa. In other words, the temporal aggregation formula provides a mapping between equally spaced representations with different sampling frequencies. Focusing on the temporal aggregation of weak GARCH(1,1) processes for flow variables, let \( \{r_t\} (t = 1, \ldots, T) \) be a weak GARCH(1,1) process with a symmetrical marginal distribution (e.g. Normal or \( t \)-Student), \( h_t = \omega + \alpha r_{t-1}^2 + \beta h_{t-1} \) and unconditional kurtosis \( k_r \). Increasing sampling frequency such that in the interval \((t-1, t)\) we have \( M \) observations, it is straightforward to see that observations are at times \( m, 2m, \ldots, T \) with \( m = 1/M \), and we can write

\[
h_{mt} = \omega_m + \alpha_m r_{mt-m}^2 + \beta_m h_{mt-m}
\]

where

\[
\omega_m = m\omega \frac{1 - (\alpha + \beta)^m}{1 - (\alpha + \beta)}, \quad \alpha_m = (\alpha + \beta)^m - \beta_m,
\]

and \( \beta_m \) is the solution of a quadratic equation also depending on \( k_r \) not reported here.

From (3), it turns out that \( \alpha_m + \beta_m = (\alpha + \beta)^m \) which implies:

1. conditional heteroskedasticity disappears when \( m \) is large (see Diebold, 1988). Given \( (\alpha + \beta) < 1 \), \( \lim_{m \to \infty} (\alpha + \beta)^m = 0 \);
2. very high-frequency processes are close to Integrated GARCH(1,1).

Given \( (\alpha + \beta) < 1 \), \( \lim_{m \to 0} (\alpha + \beta)^m = 1 \).

Starting from daily parameters, Figures 1 and 2 report GARCH(1,1) parameters at lower and higher frequencies. Note that in the two figures the starting parameters \( \alpha \) and \( \beta \) are the same. What changes is the kurtosis index \( k_r \). The points 1) and 2) above outlined are clearly distinguishable. An application to real data (six exchange rates considering daily, weekly and monthly frequencies) reveals that high-frequency GARCH(1,1) model can be identified from low-frequency data.

Andersen and Bollerslev (1997), henceforth AB97, suggest that an important limitation of the work of DN is to neglect a possible daily periodic component usually documented in high-frequency time-series. In presence of strong intraday periodicity, standard ARCH-type methodology can fail in modelling volatility of high frequency returns. Moreover, the theoretical relationships among parameter estimates at different sampling frequencies do not generally apply.
The major contribution of AB97 is the encountering of a usually pronounced U-shaped intraday seasonality in volatility pattern. In particular, volatility is high at the open and the close of the trading day and low in the middle. Intraday periodicity can have a strong impact on the autocorrelation patterns of absolute intraday returns. These features are detected by AB97 in their analysis of the 5-minute returns for the (DM-$) exchange rate (74880 observations) and the stock index SP500.

Andersen and Bollerslev propose a model specifically for high-frequency returns which explicitly incorporates the intradaily periodicity. After dividing the day $t$ into $N$ intervals indexed by $n$ such that $n = 1, \ldots, N$, the return of day $t$ at interval $n$ is defined as $r_{t,n} = \log \frac{P_{t,n}}{P_{t,n-1}}$ (but if $n = 1$, $r_{t,n} = \log \frac{P_{t,n}}{P_{t-1,n}}$). AB97 propose to model $r_{t,n}$ as

$$ r_{t,n} = E(r_{t,n}) + h_t \frac{1}{N^{1/2}} s_{t,n} \epsilon_{t,n} \tag{4} $$

where $E(r_{t,n})$ denotes the unconditional mean, $h_t$ is a daily volatility factor, $N$ is the number of returns in one day, $s_{t,n}$ is the deterministic, day-specific

![Graph showing the relationship among the parameters of GARCH(1,1) processes at different frequencies with k_r = 6. Starting daily parameters are α = 0.05 and β = 0.87.](image)
intraday periodic component, such that \( s_{t,n} = s_{t,n+jN} \) for any integer \( j \) and \( \epsilon_{t,n} \sim iid(0, 1) \). The model (4) can be simplified replacing \( s_{t,n} \) with a term, \( s_n \), not depending on the day \( t \).

The Authors suggest to estimate \( s_{t,n} \) through the Fourier flexible functional form. After removing the seasonal component, the filtered returns, defined as

\[
\tilde{r}_{t,n} = \frac{r_{t,n}}{\hat{s}_{t,n}},
\]

can be used for standard statistical analysis to discover their main characteristic (such as the pattern of the autocorrelation of returns and absolute returns). Moreover, they should conform more closely to the theoretical aggregation results for the GARCH(1,1) model given by DN.

Finally, the filtered standardized returns can be defined as

\[
\hat{r}_{t,n} = \frac{\tilde{r}_{t,n}}{\hat{h}_t}
\]

On the whole, AB97 remark that traditional time series methods applied to high-frequency data returns may give rise to erroneous inference about the return

Fig. 2: **Relationship among the parameters of GARCH(1,1) processes at different frequencies with**

\( k_r = 3 \). Starting daily parameters are \( \alpha = 0.05 \) and \( \beta = 0.87 \).
volatility dynamics. They suggest that it is necessary an explicit allowance for the influence of the usually strong periodicity. After removing it, the salient intraday volatility features can be discovered.

The High-Frequency Multiplicative Component GARCH, proposed by Chanda, Engle and Sokalska (2005), is a new model for intraday returns. The main feature of the model is the decomposition of the volatility of high-frequency returns into multiplicative components.

The intraday return \( r_{t,n} \) is modelled as

\[
r_{t,n} = \sqrt{h_t s_n q_{t,n}} \epsilon_{t,n},
\]

where \( h_t \) is the daily variance component, \( s_n \) is the diurnal (calendar) variance pattern, \( q_{t,n} \) is the intraday variance component with unity mean and \( \epsilon_{t,n} \) is the usual error term.

The innovation with respect to the model (4) of AB97 consists in assuming the existence of two distinct intradaily components, the stochastic component \( q_{t,n} \) and the deterministic intradaily component \( s_n \).

The daily component \( h_t \) is not estimated, but a commercially available volatility forecast is used while the diurnal component \( s_n \) is computed as the standard deviation of returns in each interval (after deflating by the daily volatility). Then we get

\[
z_{t,n} = \frac{r_{t,n}}{\sqrt{h_t s_n}}
\]

and

\[
z_{t,n} | I_{t,n-1} \sim N(0, q_{t,n})
\]

\[
q_{t,n} = \omega + \alpha z_{t,n-1}^2 + \beta q_{t,n-1}
\]

An application to 2721 individual companies (3 months period) and to 54 grouped companies has been carried out.

A different approach is followed by Müller et al. (1997) who proposed the Heterogeneous interval ARCH (HARCH) model emphasizing that the conditional variance depends on past squared returns taken at different frequencies.

4. CAN INTRADAILY RETURNS HELP TO BETTER UNDERSTAND DAILY VOLATILITY?

The first example of using intradaily returns for refining the analysis of daily volatility is provided by Andersen and Bollerslev (1998), henceforth AB98, who
deal with the goodness of the forecasts provided by standard GARCH models. Actually, in spite of highly significant in-sample parameter estimates, numerous studies find that standard volatility models explain little of the variability in ex-post squared returns. AB98 demonstrate how high-frequency returns permit the construction of improved ex-post volatility measures.

The problem is the following: when using regression model (2), are the observed daily squared returns good ex-post measures of volatility? AB98 show that the answer is negative. Daily squared returns provide unbiased estimator of the underlying latent volatility, but they are influenced by the idiosyncratic component of daily returns which is very large and so the $R^2$'s from regression (2) are usually very poor (even if it is shown that for a GARCH process those $R^2$'s have an upper bound given by $k^{-1}$ where $k$ is the kurtosis index of the conditional distribution of the returns).

AB98 suggest to replace the squared returns with the sum of the corresponding squared intraday returns. So, instead of $r^2_{t+1}$, they use

$$\sum_{n=1}^{N} r^2_{t,n}$$

which is an error-free volatility measure.

The regression is now

$$\sum_{n=1}^{N} r^2_{t,n} = a_N + b_N \hat{\sigma}^2_{t+1} + \epsilon_{N,t+1}$$

Using this regression, we find a much stronger evidence that volatility forecasts from GARCH model are good forecasts. Hence the famous title of the paper: Answering the skeptics. Yes, standard volatility models do provide accurate forecast.

Expression (5) is the so-called Realized Volatility and has fastly become a core concept in financial time-series literature. The choice of the specific high-frequency to be used has to take into account the trade-off between negligible sample variation and market microstructure effects, such as the bid-ask bounce. Andersen et al. (2000) suggest a graphical tool, the volatility signature plot, that is a plot of realized volatility against sampling interval. The relationship is usually negative in corresponding of very high frequencies (small sampling intervals). With the increase of sampling interval the plot becomes approximately constant. The stabilization of the realized volatility permits to specify the optimal sampling interval. The search for the optimal sampling frequency has been investigated also by Bandi and Russell (2005).
Hansen and Lunde (2006), henceforth HL, deal with the issue of the market microstructure noise affecting the realized variance estimator. Taking into account the market microstructure noise prevents the estimator from being unreliable. In order to remedy, however, one should know the properties of the market microstructure noise. According to the empirical analysis carried out by HL, the market microstructure noise turns out to be negatively correlated with the latent efficient price and autocorrelated. Moreover its properties tend to change over time.

To take into account a possible first-order autocorrelation, (5) could be replaced by

$$\sum_n r_{t,n}^2 + \sum_n r_{t,n} r_{t,n-1} + \sum_n r_{t,n} r_{t,n+1}$$

as proposed by Zhou (1996). Higher autoregressive structures may be encountered with caution.

Engle and Gallo (2006), henceforth EG, observe that various measures of volatility exist in literature. The squared daily return, the daily high-low range and the daily realized volatility are those considered by the authors. Accordingly, many forecasts of volatility can be constructed. As a result, an improvement of the forecasts could be obtained by jointly considering these indicators. Defined the log-return at time $t$ as $r_t = \log(C_t/C_{t-1})$ where $C_t$ is the close price at time $t$, squared returns can be described by a multiplicative error model,

$$r_t^2 = h_t \epsilon_t$$

where $\{\epsilon_t\}$ is a sequence of iid positive definite random variable whose distribution conditionally on the information at time $t-1$ is characterized by a unit mean, that is $\epsilon_t | I_{t-1} \sim iid(1, \sigma^2)$. Then, the conditional mean of $r_t^2$ turns out to be $h_t$. One of the many GARCH-type models can be considered for $h_t$, such as the symmetrical GARCH(1,1),

$$h_t = \omega + \alpha r_{t-1}^2 + \beta h_{t-1}, \quad (6)$$

or an asymmetrical version to take into account the possible leverage effect, e.g.

$$h_t = \omega + \alpha r_{t-1}^2 + \beta h_{t-1} + \gamma r_{t-1}^2 d_{t-1}, \quad (7)$$

where the dummy variable $d_t$ is defined in Section 2.

EG propose to enlarge the lagged part of equation (6) or (7) through the inclusion of other measures of volatility. In particular they propose to insert the
(lagged) squared high-low range and the (lagged) square of the realized volatility. The high-low range is defined as $HL_t = \log(H_t/L_t)$, with $H_t$ and $L_t$ indicating the highest and the lowest price during the trading day. The realized volatility, $v_t$, is the square root of the sum of squared returns over $N$ subperiods within the day. The resulting expression for $h_t$ is now:

$$h_t = \omega + \alpha r_{t-1}^2 + \beta h_{t-1} + \gamma r_{t-1}^2 d_{t-1} + \varphi HL_{t-1}^2 + \theta HL_{t-1}^2 d_{t-1} + \psi v_{t-1}^2 + \lambda v_{t-1}^2 d_{t-1}$$

(8)

The ratio of the approach is that different measures of volatility emphasize different aspects of the volatility. A high squared daily return can be associated to a very low high-low range if the opening price is very high. By the same token, in correspondence of a high squared daily return we can observe a high or low realized volatility. Note that in (8) asymmetric effects are encountered also in $HL_{t-1}^2$ and $v_{t-1}^2$.

An analysis has been carried out using daily SP500 returns (January 4, 1988 - December 14, 1998). They detected a high correlation coefficient both between squared return and daily range and between daily range and realized volatility. On the contrary the correlation between squared return and realized volatility is not strong (about 0.5). This confirms that using multiple indicators for predicting volatility can be a good strategy.

The results show that the enlargement of the information set adds explanatory power to the expression (8). The Authors have also experimented the possible enrichment of the equation for the prediction of the volatility along the same lines.

Giot and Laurent (2004) compare the daily volatility forecasts using two models, with daily and intradaily data respectively. In order to evaluate the competitive models, the Authors refer to the computation of a one-day ahead Value-at-Risk (VaR). The former is an ARCH-type model estimated using daily returns, the latter makes use of intradaily data and, in particular, it exploits the definition of realized volatility. Specifically, the daily model is an Asymmetric Power ARCH (Ding et al., 1993) with skewed $t$ innovations. The realized volatility is modeled as a long-memory model. The Authors concludes that there is not a model which forecasts the volatility in a better way. However, the key is the use of a sophisticated model for daily data. A simple Gaussian GARCH(1,1) would provide different results.

A very recent and alternative approach is proposed by Ghysels et al. (2006), henceforth GSV. They consider various mixed data sampling (MIDAS)

---

1 The Value-at-Risk is the loss that can occur over a given period, at a given confidence level, due to exposure to market risk.
The ultimate step of the increase of the sampling frequency is reached when the data are recorded each time a market event occurs. In this context, instead of sampling prices at shorter and shorter intervals, we record the price whenever a change occurs. In this case the time between two events (price change) is not constant. We have irregularly spaced time series for which the traditional tools of financial econometrics do not apply. In this case we have to model the time between two events (duration) (Engle, 2000).

The class of Autoregressive Conditional Duration (ACD) models is directly tailored to the irregular spacing of the data (Engle and Russell, 1998). If \( X_i \) is the duration between two observations at times \( t_{i-1} \) and \( t_i \), Engle and Russell proposed the ACD\((q,p)\) model

\[
X_i = \phi(t_i) \Psi_i \epsilon_i \\
\Psi_i = \omega + \sum_{j=1}^{q} \alpha_j x_{i-j} + \sum_{j=1}^{p} \beta_j \Psi_{i-j}.
\]

where \( \phi(t_i) \) is a deterministic daily seasonal component and \( x_i = X_i / \phi(t_i) \) is the \( i \)-th seasonally adjusted duration. Assuming \( \epsilon_i \) identically and independently distributed with \( E(\epsilon_i) = 1 \), it is easy to show \( E(x_i | F_{i-1}) = \Psi_i \), where \( F_{i-1} \) is

\[
V_{t+h|t} = \mu_h + \phi_h \sum_{k=0}^{k_{\text{max}}} b_h(k; \theta) \tilde{X}_{t-k|t-k+1} + \epsilon_{ht}
\]

where \( V_{t+h|t} \) is a measure of volatility at time \( t+h \) given the information at time \( t \) and \( \tilde{X}_{t-k|t-k+1} \) contains the explanatory variables, not necessarily lags of the left-hand side variable. The index \( m \) specifies the sampling frequency.

The context is very promising, in that there is the possibility of exploring the predictive power of high-frequency (intradaily) regressors versus a daily forecast volatility. The conclusions of GSV are cautious, as they maintain that the use of high-frequency data does not necessarily lead to better volatility forecast. In our opinion, a deeper analysis (e.g. the analysis of new data) could shed more light on a very interesting and flexible statistical tool.

Other studies on the possible incremental information from high-frequency data are present in literature (e.g. Blair et al., 2001).

5. CAN ULTRA-HIGH-FREQUENCY DATA HELP TO DISCOVER FEATURES OTHERWISE NOT CAPTURABLE?

The ultimate step of the increase of the sampling frequency is reached when the data are recorded each time a market event occurs. In this context, instead of sampling prices at shorter and shorter intervals, we record the price whenever a change occurs. In this case the time between two events (price change) is not constant. We have irregularly spaced time series for which the traditional tools of financial econometrics do not apply. In this case we have to model the time between two events (duration) (Engle, 2000).

The class of Autoregressive Conditional Duration (ACD) models is directly tailored to the irregular spacing of the data (Engle and Russell, 1998). If \( X_i \) is the duration between two observations at times \( t_{i-1} \) and \( t_i \), Engle and Russell proposed the ACD\((q,p)\) model

\[
X_i = \phi(t_i) \Psi_i \epsilon_i \\
\Psi_i = \omega + \sum_{j=1}^{q} \alpha_j x_{i-j} + \sum_{j=1}^{p} \beta_j \Psi_{i-j}.
\]

where \( \phi(t_i) \) is a deterministic daily seasonal component and \( x_i = X_i / \phi(t_i) \) is the \( i \)-th seasonally adjusted duration. Assuming \( \epsilon_i \) identically and independently distributed with \( E(\epsilon_i) = 1 \), it is easy to show \( E(x_i | F_{i-1}) = \Psi_i \), where \( F_{i-1} \) is
the information at time \( t_{i-1} \). As a result, \( \Psi_i \) can be interpreted as the expected (deseasonalized) duration conditionally on the information at time \( t_{i-1} \).

A rich production has extended the original approach in several directions. However, very few articles deal with the effect (possibly the gain) of the use of these data on the volatility estimation and forecasting. We will mention two of them.

Ghysels and Jasiak (1998), henceforth GJ, developed a class of ARCH models for time series characterized by unequal time intervals. The idea of GJ is that of modelling both the durations and the volatility. For the former they use the class of ACD models (Engle and Russell, 1998), for the latter the popular GARCH(1,1).

The ACD-GARCH is a bivariate model, so the issue of the possible interdependence between the variables is relevant. Do past volatilities affect durations? Do past durations affect volatilities?

It is assumed that

\[
\begin{align*}
\Psi_i & = \omega^d + \alpha^d (t_{i-1} - t_{i-2}) + \beta^d \Psi_{i-1} \\
\sigma_i^2 & = \omega + \alpha e_{i-1}^2 + \beta \sigma_{i-1}^2
\end{align*}
\]

where the parameters \( \omega, \alpha_i \) and \( \beta_i \) are computed applying the formulas given in DN (see formulas (3)) with \( m \) replaced by \( \Psi_i \) as provided by the first equation of (10). Actually this type of modelling has not been very successful.

Engle (2000) maintains that prices convey information on the volatility of the market. His aim is to determine a measure of price volatility using transaction data and focus on the relevance of the timing of trades on the volatility itself. The proposed model is denoted as Ultra-High-Frequency GARCH (UHF-GARCH). The return \( r_i \) is defined as the return from the \((i-1)\)-th to the \(i\)-th transaction. The conditional (on the past and on the current duration \( x_i \)) variance (per transaction) is

\[
Var(r_i|x_i, I_{i-1}) = h_i.
\]

The conditional volatility per unit of time is given by

\[
Var \left( \frac{r_i}{\sqrt{x_i}} | x_i, I_{i-1} \right) = \sigma_i^2,
\]

and the relationship between the two variances is simply \( h_i = x_i \sigma_i^2 \).

A simple formulation of the UHF-GARCH model is obtained through a GARCH(1,1) specification for \( \sigma_i^2 \),

\[
\sigma_i^2 = \omega + \alpha e_{i-1}^2 + \beta \sigma_{i-1}^2 + \gamma x_i^{-1},
\]
where \( e_i \) is the innovation of returns per square root of time following an ARMA(1,1) process. The expected sign of \( \gamma \) is positive, implying that long durations lower volatility.\(^{(2)}\) When \( \gamma = 0 \), no effect of durations on volatility is accounted for. This is a GARCH model estimated in transaction time.

A more extended formulation considers

\[
\sigma_i^2 = \omega + \alpha e_{i-1}^2 + \beta \sigma_{i-1}^2 + \gamma_1 x_{i-1} + \gamma_2 x_i \Psi_i^{-1} + \gamma_3 \xi_i^{-1} + \gamma_4 \Psi_i^{-1}.
\]

Volatility is affected by current duration \( x_i \), the excess duration \( x_i \Psi_i^{-1} \), the long run volatility \( \xi_i \) computed by exponentially smoothing \( r^2/x \) and the expected trade arrival rate \( \Psi_i^{-1} \). Because of the presence of terms containing \( \Psi_i \), a previous estimation of an ACD model is needed. The estimates reveal a sum \( \alpha + \beta \) usually lower than 0.5, that is a very small persistence.

Starting from the UHF-GARCH model and taking into account the autocorrelation caused by microstructure effect, Engle and Sun (2005) deal with the one-day volatility forecasting using tick-by-tick data.

6. CONCLUSIONS

The analysis of volatility of financial time series is a research area for a lot of academics and a very high number of contributions witnesses the relevance of the financial and statistical issues. In recent years one of the most important developments has been represented by the use of high-frequency data (regularly and irregularly spaced). A review of the most innovative papers has been presented. In particular, we believe that a more subtle effort could be done in order to explore the potentiality of ultra-high-frequency data, obtained when the limit of the increasing of the sampling frequency is reached.

REFERENCES


\(^2\) In the financial market microstructure theory, the hypothesis by Easley and O’Hara (1992) support this point of view.


**LA PREVISIONE DELLA VOLATILITÀ CON DATI AD ALTA FREQUENZA**

**Riassunto**

Il lavoro si focalizza sull’analisi e la previsione della volatilità utilizzando dati con una frequenza maggiore di quella giornaliera. L’obiettivo consiste nel verificare l’adeguatezza della tradizionale modellistica di tipo ARCH per l’analisi dei rendimenti ad alta frequenza e l’ausilio degli stessi per ottenere una migliore previsione della volatilità giornaliera. Infine l’attenzione viene posta sui cosiddetti dati tick-by-tick caratterizzati da intervalli di tempo non equispaziati.