Model-Based Measurement of Actual Volatility in High-Frequency Data

B. Jungbacker[1]
S.J. Koopman[1,2]

[1] *Faculty of Economics and Business Administration, Vrije Universiteit Amsterdam,*
[2] *Tinbergen Institute.*

Model-based measurement of actual volatility in high-frequency data

B. Jungbacker   and   S.J. Koopman

Department of Econometrics, Free University Amsterdam,

De Boelelaan 1105, NL-1081 HV Amsterdam, Netherlands

December 23, 2004

Abstract

In this paper we aim to measure actual volatility within a model-based framework using high-frequency data. In the empirical finance literature it is known that tick-by-tick prices are subject to market micro-structure such as bid-ask bounces and trade information. Such market micro-structure effects become more and more apparent as prices or returns are sampled at smaller and smaller time intervals. High-frequency returns can be used for the computation of realised volatility. Recent theoretical results show that realised volatility is a consistent estimator of actual volatility but when it is subject to micro-structure noise, the estimator diverges. Nonparametric methods can be adopted to account for the micro-structure bias. Here we measure actual volatility using a model that takes account of micro-structure noise together with intra-daily volatility patterns and stochastic volatility. The coefficients of this model are estimated by maximum likelihood methods that are based on importance sampling techniques. It is shown that such Monte Carlo techniques can be employed successfully for our purpose in a feasible way. As far as we know, this is a first serious attempt to model the basic components of the mean and variance of high-frequency prices simultaneously. An illustration is given for three months of tick-by-tick transaction prices of the IBM stock traded at the New York Stock Exchange.

Keywords: Importance sampling; Maximum likelihood estimation; Micro-structure noise; Realised variance; Stochastic volatility model.
JEL classification: C22, C53, G15.

1 Introduction

The price of a financial asset is denoted by P_t. A common assumption in the finance literature is that the log of P_t can be represented by a stochastic differential equation (SDE) of the form

$$d \log P_t = \mu_t(\psi)dt + \sigma_t(\psi)dB_t, \qquad t > 0, \tag{1}$$

where $\mu_t(\psi)$ is the drift function representing expected return, $\sigma_t(\psi)$ is a stochastic process representing the spot volatility, B_t is a standard Brownian motion and ψ is a vector of unknown parameters, see Campbell, Lo, and MacKinlay (1997) for more background. In many contexts the financial economist is interested in measuring and predicting the variability of the asset price. This variability is mainly determined by what is called integrated volatility

$$\sigma^{*2}(0, t) = \int_0^t \sigma_t^2(\psi)dt, \tag{2}$$

where the dependence of ψ is implied. The related concept of actual volatility for the interval $[t_1, t_2]$ is defined as $\sigma^{*2}(t_1, t_2)$ where

$$\sigma^{*2}(t_1, t_2) = \sigma^{*2}(0, t_2) - \sigma^{*2}(0, t_1). \tag{3}$$

It should be noted that integrated and actual variance would be the more precise names for integrated and actual volatily, respectively. However we choose to follow the convention in much of the financial econometrics literature and refer to these quantities as volatilities.

The price of an asset can be observed when a trade has taken place. Trades ocur continuously at international financial markets. The Trades and Quotes (TAQ) database of the New York Stock Exchange (NYSE) contains all equity transactions reported on the so-called Consolidated Tape and it includes transactions from the well-known NYSE, AMEX and NASDAQ markets but also from various important other exchange markets. When such trade prices within a certain period are collected, we obtain a so-called high-frequency dataset. We refer to high-frequency data when observations are sampled at very small time intervals. In the finance literature, this usually means that observations are taken at the intra-daily interval of five minutes or 1 minute (calendar time sampling) or that observations are recorded trade-by-trade (business time sampling). The trade-by-trade data is regarded as the ultimate high-frequency collection of prices. Note that time in such databases is usually measured in seconds. In calendar time, we may therefore have multiple trades within the same time-interval although this is unlikely. It is more likely that on a time-scale of seconds many prices will be missing since trades do not take place every second.

The observed log price at the discrete time point t_n (in seconds) is denoted by $Y_n = \log P_{t_n}$ for observation index n. The number of time points (seconds) in one trading day is denoted by

N_d. We therefore have potentially N_d observations $Y_1, Y_2, \ldots, Y_{N_d}$ of the log price of a trade in a particular day d. The index t_0 refers to the start of the period while the distance $t_n - t_{n-1}$ is constant for $n = 1, \ldots, N_d$. The value Y_n will not be available when no trade has taken place at time t_n. Such values will be treated as missing. The number of trades is denoted by $N \leq N_d$ so that we have $N_d - N$ missing values in day d.

A natural estimator of actual volatility is given by what is called realised volatility denoted by $\tilde{\sigma}^{*2}(t_0, t_{N_d})$ and given by

$$\tilde{\sigma}^{*2}(t_0, t_{N_d}) = \sum_{j=2}^{N_d/m} (Y_{mj} - Y_{mj-m})^2, \tag{4}$$

where m is the sampling frequency, see Andersen, Bollerslev, Diebold, and Labys (2001). For example, when the sampling frequency is 5 minutes, m equals 300 assuming that the index of Y_n refers to the nth second. In the case a transaction has not taken place at time n, so that Y_n is missing in (4), it can be approximated via an interpolation method using observed values in the neighbourhood of Y_n, see Malliavin and Mancino (2002) and Hansen and Lunde (2004) for discussions of different filtering methods. Novel asymptotic theory is developed for the realised volatility estimator (4) as the number of observations in the fixed interval $[t_0, t_n]$ increases (or as m decreases), see Barndorff-Nielsen and Shephard (2001). Specifically it is shown that $\tilde{\sigma}^{*2}(t_0, t_{N_d})$ is a consistent estimator of actual volatility. This result suggests that if we sample the log price process $\log P_t$ more often in a fixed time interval by taking m small, the efficiency of the estimator will be increased. Empirical work on the subject however indicates the complete opposite, see, in particular, Andreou and Ghysels (2001) and Bai, Russell, and Tiao (2000). If the realised volatility is computed using ever more observations, the estimate seems to diverge. The cause of this phenomenon is the fact that the efficient price is not observed directly. Observed trading prices Y_n are contaminated by so-called micro-structure noise that has various causes such as bid-ask bounces, discrete price observations and irregular trading, see Campbell, Lo, and MacKinlay (1997) for a further discussion with references.

It is therefore argued that micro-structure noise ruins the reliability of realised volatily as an estimator. Recently non-parametric methods have been proposed by Barndorff-Nielsen, Hansen, Lunde, and Shephard (2004) and Aït-Sahalia, Mykland, and Zhang (2004) that produce consistent estimates of volatility even in the presence of micro-structure noise.

In this paper we take a model-based approach to measure volatility using high-frequency prices that are observed with micro-structure noise. Standard models for price and volatility from the finance literature will be considered. Further, the model allows for an intra-daily volatility pattern and stochastic volatility. The details of the model are described in section 2. In this way, the salient features of high-frequency prices are described and efficient estimates of actual volatility can be produced. However, the estimation of parameters in this class of models

is nonstandard and simulation-based methods need to be employed. This also applies to the methods of volatility measurement. We propose importance sampling techniques in section 3 and it is shown that such methods can work effectively and efficiently for our purposes. This is illustrated in section 4 in which daily volatilities are measured from high-frequency IBM prices recorded at the New York Stock Exchange for a period of three months.

2 Models for high-frequency prices

2.1 Model for price with micro-structure noise

Many possible specifications for the drift and diffusion components of model (1) have been proposed in the financial literature. Throughout the paper we assume that the drift term equals zero so that $\text{var}(P_{t+\tau}|P_t)$ only depends on the diffusion term $\sigma_t(\psi)$ in (1), see Andersen, Bollerslev, and Diebold (2002). For the volatility process $\sigma_t(\psi)$ we consider various specifications. The first and most basic specification is where the volatility is kept constant over time, that is $\sigma_t(\psi) = \sigma(\psi)$. These assumptions lead us to the following model for the efficient price process,

$$\mathrm{d}\log P_t = \sigma(\psi)\mathrm{d}B_t, \tag{5}$$

where B_t is standard Brownian motion.

Further we assume that the observed trade prices Y_n is a noisy observation of the efficient price and we denote the micro-structure noise by U_n. We therefore have $Y_n = \log P_t + U_n$ for $t = t_n$. It is assumed that the micro-structure noise has zero mean and variance σ_U^2. The discrete time model then becomes

$$Y_n = p_n + \sigma_U U_n, \qquad\qquad U_n \sim \text{IID}(0,1), \tag{6}$$

$$p_{n+1} = p_n + \sigma_\varepsilon \varepsilon_n, \qquad\qquad \varepsilon_n \sim \text{NID}(0,1), \tag{7}$$

where $p_n = \log P_t$ is the unobserved price (in logs) at time $t = t_n$ for $n = 1, \ldots, N_d$. It is noted that n refers to an index of seconds with equidistances $t_n - t_{n-1}$.

In this setting we have a simple expression for actual volatility

$$\sigma^{*2}(t_n, t_{n+1}) = (t_{n+1} - t_n)\sigma_\varepsilon^2.$$

The model implies that the observed return

$$R_n = \Delta Y_{n+1} = \Delta p_{n+1} + \sigma_U \Delta U_{n+1} = \sigma_\varepsilon \varepsilon_n + \sigma_U U_{n+1} - \sigma_U U_n,$$

follows an moving average (MA) process of order one, that is $R_n \sim \text{MA}(1)$, see Harvey (1989) for a further discussion of the local level model.

4

The assumption of constant volatility is too strong for a relative long period, say one day. However, this simple framework allows us to obtain a preliminary estimate of daily volatility using high-frequency data. The estimation of the local level model (7) is explored in detail by Durbin and Koopman (2001, Chapter 1) and is based on the standard Kalman filter equations. The possibly many missing values in the series Y_n can be accounted for within the Kalman filter straightforwardly. When it is assumed that micro-structure noise U_n and price innovation ε_n are Gaussian distributed error terms, exact maximum likelihood (ML) estimates of σ_U^2 and σ_ε^2 are obtained by numerically maximising the Gaussian likelihood function that can be evaluated by the Kalman filter. When the Gaussian assumption is not wanted, these estimates can be referred to as quasi maximum likelihood (QML) estimates.

Aït-Sahalia, Mykland, and Zhang (2004) also consider the local level model framework to describe the true process of the observed log prices and also observe that the returns therefore follow an MA(1) process. In their theoretical analysis it is argued that distributional properties of U_n do not matter asymptotically. Their main conclusions of their analysis are that *modelling the noise explicitly restores the first order statistical effect that sampling as often as possible is optimal* and *this remains the case if one misspecifies the assumed distribution of the noise term.* We regard this as an endorsement of our modelling approach. They further discuss possible extensions of the local level model by modelling U_n as a stationary autoregressive process and by allowing for contemporaneous correlation between U_n and ε_n. In our modelling framework, the former extension can be incorporated straightforwardly although the estimation of autoregressive moving average (ARMA) processes for financial data may be hard in practice. The latter proposed extension is more difficult from an inference point of view since the correlation coefficient between U_n and ε_n is not identified when both variances are unrestricted, see the discussion in Harvey and Koopman (2000).

2.2 Intra-daily seasonal patterns in volatility

In empirical work it is often found that estimates of actual volatility for different intervals within the day show a strong seasonal pattern. At the opening and closure of financial markets, price changes are more volatile than at other times during the trading session. In 24-hour markets, such different volatile periods within the day can be found too. Discussions of this phenomenon and empirical evidence are given by, amongst many others, Dacarogna, Müller, Nagler, Olsen, and Pictet (1993) and Andersen and Bollerslev (1997). To account for the intra-daily variation of integrated volatility we replace the constant spot volatility σ^2 in (5) by an intra-daily seasonal specification in the volatility, that is

$$\sigma_t^2 = \sigma^2 \exp g(t), \quad \text{or} \quad \log \sigma_t^2 = \log \sigma^2 + g(t),$$

where $g(t)$ is a deterministic function that can represent a diurnal pattern and starts at zero, that is $g(0) = 0$. An example of an appropriate function $g(t)$ is given in Appendix A. The integrated volatility becomes

$$\sigma^{*2}(0,t) = \int_0^t \sigma_s^2 \mathrm{d}s = \sigma^2 \int_0^t \exp g(s) ds. \tag{8}$$

The actual volatility can be analytically derived from (8) or it can be approximated by

$$\sigma^{*2}(t_n, t_{n+1}) \approx \sigma^2 \sum_{s=t_n}^{t_{n+1}} \exp g(s),$$

where the index step length can be chosen to be very small. As a result, σ_ε^2 in (7) is replaced by $\sigma_{\varepsilon,n}^2 = \sigma^{*2}(t_n, t_{n+1})$ with σ^2 representing the constant variance part that is still denoted by σ_ε^2. The function $g(t) = g(t; \psi)$ depends on parameters that are collected in vector ψ, together with the variances σ_ε^2 and σ_U^2, that will be estimated by maximum likelihood methods. As a result, the model (7) is unchanged except that the state variance has become dependent of a deterministic function of time. The Kalman filter can incorporate time-varying coefficients and therefore the estimation methodology remains straightforward.

2.3   Stochastic volatility model

Various specifications for stochastic volatility models have been proposed. To keep the analysis and estimation simple, we will assume one of the most basic, non-trivial specifications. The efficient price process (5) is extended as follows. The constant volatility σ is replaced by a stochastic time-varying process to obtain the system of SDE's

$$\begin{aligned}
\mathrm{d} \log P_t &= \sigma_t \mathrm{d} B_t^{(1)}, \\
\log \sigma_t^2 &= \log \sigma_t'^2 + \xi, \\
\mathrm{d} \log \sigma_t'^2 &= -\lambda \log \sigma_t'^2 \mathrm{d}t + \sigma_\eta \mathrm{d} B_t^{(2)},
\end{aligned} \tag{9}$$

where $B_t^{(1)}$ and $B_t^{(2)}$ are independent Brownian motions while $\log \sigma_t'^2$ represents an Ornstein-Uhlenbeck process. Here ξ represents the fixed mean of log volatility. The vector of unknown parameters is $\psi = (\lambda \ \xi \ \sigma_\eta^2)'$. Using the Euler-Maruyama method, see Kloeden and Platen (1999) for details, we obtain an approximation to the solution of the system of SDE's (9) as given by the discrete representation

$$\begin{aligned}
\log P_{t_{n+1}} &= \log P_{t_n} + \sigma_n \varepsilon_n, & \varepsilon_n &\sim \mathrm{NID}(0,1), \\
\log \sigma_n^2 &= \log \sigma_n'^2 + \xi, & & \\
\log \sigma_{n+1}'^2 &= (1-\lambda) \log \sigma_n'^2 + \sigma_\eta \eta_n, & \eta_n &\sim \mathrm{NID}(0,1).
\end{aligned} \tag{10}$$

6

Note that $\lambda = \sigma_\eta = 0$ implies constant volatility with $\log \sigma_n^2 = \xi$. The process (10) represents the standard discrete stochastic volatility (SV) model, see Ghysels, Harvey, and Renault (1996) for an overview. It follows that the actual volatility is approximated by

$$\sigma^{*2}(t_n, t_{n+1}) \equiv \int_{t_n}^{t_{n+1}} \sigma_s^2 \mathrm{d}s \approx (t_{n+1} - t_n)\sigma_n^2.$$

Finally, assuming that a particular day d consists of N_d intraday intervals, the actual volatility of day d is approximated by

$$\sigma^{*2}(t_0, t_{N_d}) \approx \sum_{n=1}^{N_d} (t_{n+1} - t_n)\sigma_n^2.$$

To analyse the stochastic log prices (mean) and the stochastic volatility (variance) simultaneously, it is more convenient to represent the model in terms of returns $\log(P_{t_{n+1}} / P_{t_n})$. It follows from the discussion in section 2.1 that when the model for log prices accounts for micro-structure noise, the observed returns R_n follows an MA(1) process. By further allowing for stochastic volatility, we obtain

$$R_n = \sigma_n \varepsilon_n + \sigma_U W_n, \tag{11}$$

where $\log \sigma_n^2$ is modelled as in (10) and $W_n = U_{n+1} - U_n$ such that $W_n \sim$ MA(1). From an estimation point of view, it will be argued in the next section that maximum likelihood estimation of this model is intricate. We therefore consider a white noise process for W_n in the empirical part of this paper.

The final model that we consider in this paper is the price model with SV that also accounts for the intra-daily seasonal pattern. In the previous section we have introduced the flexible deterministic function $g(t)$ for this purpose. The final model is therefore based on the system of SDE's

$$\mathrm{d} \log P_t = \sigma_t \mathrm{d}B_t^{(1)},$$
$$\log \sigma_t^2 = \log \sigma_t'^2 + g(t),$$
$$\mathrm{d} \log \sigma_t'^2 = -\lambda \log \sigma_t'^2 \mathrm{d}t + \sigma_\eta \mathrm{d}B_t^{(2)},$$

The flexible function $g(t)$ is incorporated in the SV specification (10) in the same way as described in section 2.2. In particular, $\log \sigma_n^2$ in (10) is replaced by

$$\log \sigma_n^2 = \log \sigma_n'^2 + \xi + g(t).$$

Borus: this only works because in our case $t_{n+1} - t_n = 1$ second.

7

3 Estimation methods

3.1 The problem of estimation

It is already argued in sections 2.1 and 2.2 that the model for prices with constant or deterministic time-varying volatilities is relatively straightforward to estimate by using Kalman filter methods. However, estimating the model with stochastic volatility is known to be much more intricate. Various methods have been developed for the estimation of the SV model without micro-structure noise. Such methods have been based on quasi-maximum likelihood, Bayesian Markov chain Monte Carlo procedures, importance sampling techniques, numerical integration, method of moments and others. The presentation of an overview of all these methods is beyond the scope of this paper but the interested reader can be referred to the collection of articles in Shephard (2005). Most of these methods have not considered the existence of micro-structure noise in the returns since the empirical applications in this work have been concerned with returns data measured at a low frequency such as months, weeks and days. The issue of micro-structure noise is irrelevant in these cases. This section discusses feasible methods for the estimation of the model for returns with SV and noise since this is relevant for high-frequency data. We limit ourselves to approximate and maximum likelihood methods. Bayesian and (efficient and/or simulated) method of moments can be considered as well and in fact we believe that such methods will be applicable too. However, given our favourable experiences with maximum likelihood estimation using importance sampling techniques for standard SV models, we have been encouraged to generalise these methods for the models described in the previous section.

To focus the discussion on estimation, the model for returns with stochastic volatility, intra-daily seasonality and micro-structure noise is represented as the nonlinear state space model

$$R_n = \exp(\frac{1}{2}h_n)\varepsilon_n + \sigma_U W_n, \tag{12}$$

$$h_n = \xi(n) + \phi \{h_{n-1} - \xi(n-1)\} + \sigma_\eta \eta_n, \tag{13}$$

where $h_n = \log \sigma_n'^2$ and $\phi = 1 - \lambda$. The log-volatily h_n follows an autoregressive process of order one and the micro-structure noise W_n follows an moving average process of order one. These processes can be generalised to other stationary time series processes. The disturbances driving the time series processes for h_n and W_n together with ε_n are assumed Gaussian and independent of each other, contemporaneously and at all time lags. These assumptions can be relaxed, see the discussion in section 3.3. The model for the returns (12) is nonlinear and depending on a state vector with the log variance $\log \sigma_n'^2$ modelled as a linear autoregressive process and with coefficients to measure the intra-daily volatility pattern $g(t)$. The nonlinearity is caused by the term $\exp(\frac{1}{2}h_n)\varepsilon_n$ in (12) since both h_n and ε_n are stochastic. Conditional on

the unobservable h_n, the model (12) is a linear Gaussian ARMA model (for the micro-structure noise $\sigma_U U_t$) with additive heteroskedastic noise (for the returns $\log P_{t_{n+1}} - \log P_{t_n}$).

Different approximate methods for the estimation of the unknown parameters in model (12) and (13) can be considered. For example, the multiplicative term $\exp(h_n / 2)\varepsilon_n$ can be linearised by a first-order Taylor expansion in h_n. The resulting linearised model can be considered by the Kalman filter. This approach is referred to as the Extended Kalman filter. The details will not be discussed here since we believe that this approach will provide a poor approximation especially when the volatility is relatively large or small, that is, when $|h_n|$ is large. Some improvements may be obtained when the resulting estimate of h_n is inserted in the model so that a linear model is obtained which can be treated using standard methods. Such a mix of approximate methods does not lead to a satisfactory estimation strategy and therefore we aim to provide a maximum likelihood estimation method in the next section.

3.2   Estimation using importance sampling techniques

The evaluation of the likelihood function using importance sampling techniques has been considered for the model (12) and (13) with $\sigma_U = 0$ by Shephard and Pitt (1997) and Durbin and Koopman (1997). Further details of this approach have been explored in Part II of the monograph of Durbin and Koopman (2001). The basic ingredients of this approach are as follows.

- The approximate linear Gaussian model

$$y = \theta + u, \qquad u \sim \text{NID}(c, V), \tag{14}$$

  is considered with its conditional density denoted by $g(y|\theta)$ where y is the vector of observations and θ is the associated unobserved signal. In the SV model without noise, we have $y = (R_1, \ldots, R_{N_d})'$ and $\theta = (h_1, \ldots, h_{N_d})'$. The approximate conditional Gaussian density $g(y|\theta)$ depends on mean vector c and diagonal variance matrix V which are chosen such that

$$\dot{g}(y|\theta) = \dot{p}(y|\theta), \qquad \ddot{g}(y|\theta) = \ddot{p}(y|\theta),$$

  where $\dot{q}(\cdot)$ and $\ddot{q}(\cdot)$ are the first and second derivatives, respectively, of the density $q(\cdot)$ with respect to θ. Further, $p(\cdot)$ refers to the density of the model (12) and (13), here with $\sigma_U = 0$. To obtain the mean and variance of $g(y|\theta)$, we require to estimate θ from the approximate linear Gaussian model (14) that also depends on θ. Therefore an iterative method involving Kalman filtering and smoothing needs to be carried out.

- Given the importance density associated with the approximate model (14), simulations from density $g(\theta|y)$ can be obtained using simulation smoothing algorithms such as the

9

recent ones of de Jong and Shephard (1995) and Durbin and Koopman (2002). The resulting simulated θ's are denoted by $\theta^{(i)} \sim g(\theta|y)$.

- The importance sampling estimator of the likelihood is based on

$$L(\psi) = p(y; \psi) = \int p(y, \theta) \mathrm{d}\theta = \int \frac{p(y, \theta)}{g(\theta|y)} g(\theta|y) \mathrm{d}\theta = g(y; \psi) \int \frac{p(y, \theta)}{g(y, \theta)} g(\theta|y) \mathrm{d}\theta,$$

and since $p(\theta) = g(\theta)$, we obtain the convenient expression

$$L(\psi) = L_g(\psi) \int \frac{p(y|\theta)}{g(y|\theta)} g(\theta|y) \mathrm{d}\theta,$$

where $L_g(\psi) = g(y; \psi)$ is the likelihood function of the approximating model. All densities $p(\cdot)$ and $g(\cdot)$ depend on parameter vector ψ even when this is not made explicit. The importance sampling estimator of the likelihood function $L(\psi)$ is therefore given by

$$\widehat{L}(\psi) = L_g(\psi) \sum_{i=1}^{M} \frac{p(y|\theta^{(i)})}{g(y|\theta^{(i)})},$$

where $\theta^{(i)} \sim g(\theta|y)$ for $i = 1, \ldots, M$. It is noted that the densities $p(y|\theta)$ and $g(y|\theta)$ are relatively easy to evaluate. The likelihood function evaluated by importance sampling is exact but subject to Monte Carlo error. Further, to ensure that the importance estimator of the likelihood is reliable, the importance density needs to have a variance, see Geweke (1989).

The last items are general and do not depend on the particular model specification. Finding an approximate linear Gaussian model from which we can generate simulation samples from $g(\theta|y)$, does obviously depend on the model in question. The details of obtaining an approximate model for the standard SV model for importance sampling can be found in Shephard and Pitt (1997) and Durbin and Koopman (2001, page 195). The values for c_n and V_n, the n-th element of c and the n-th diagonal element of V, respectively, in this case are obtained by

$$V_n = 2 \frac{\exp(h_n)}{R_n^2}, \qquad c_n = \frac{1}{2} V_n + R_n - h_n - 1. \tag{15}$$

For the case with micro-structure noise the values for c and V need to be derived as hinted in the first item. The details of the derivations are given in Appendix B. For the case of IID noise, that is $W_n \sim \mathrm{IID}(0, 1)$, the actual values are given by

$$V_n^{-1} = \frac{1}{2} \left(b_n - b_n^2 \right) + \left(b_n - \frac{1}{2} \right) \frac{b_n}{a_n} R_n^2, \qquad c_n = R_n - h_n - \frac{1}{2} V_n b_n \left(\frac{R_n^2}{a_n} - 1 \right), \tag{16}$$

where $a_n = \exp(h_n) + \sigma_U^2$ and $b_n = \exp(h_n) / a_n$. We note that $a_n > 0$ and $0 < b_n \leq 1$. For this methodology to work, we obviously require $V_n > 0$ for all n. This can only be guaranteed

10

when $\sigma_U^2 < \exp(h_n)$ since this implies that $b_n > \frac{1}{2}$. In other words, the variation due to microstructure noise must be smaller than the variation due to volatility in returns. This condition refers unexpectedly to the debate whether realised volatility merely reflects microstructure noise variation or underlying price changes of assets, see Bandi and Russell (2004) and Zhang, Mykland, and Aït-Sahalia (2004). If $V_n \leq 0$, importance sampling has a problem. We take a practical approach on this issue. If V_n is negative, it is replaced by the V_n obtained when no micro-structure noise is considered, that is $V_n = 2 \exp(h_n) / R_n^2$.

3.3 Discussion of estimation methods

Details of importance sampling methods for estimating the general model are presented in the previous section. It is assumed that W_n is IID while the basic modelling framework for the prices in section 2.1 insists that W_n should be at least modelled as a MA(1) process or possibly as an ARMA process. The consideration of an ARMA disturbance term in the measurement equation requires multivariate sampling devices which are intricate and will slow down the computations considerably. Other methods need to be considered when we want to allow for ARMA structures in W_n and currently research is devoted on this matter.

For estimation purposes the price model with stochastic volatility is reformulated in terms of returns. The ultimate aim however is to estimate models as specified in (10). This is not an easy task and various methods can be considered. In this paper we have considered Kalman filter and importance sampling techniques. This may lead to feasible methods but it is not yet clear how they can be utilised more effectively for the task at hand.

Other estimation techniques can also be considered such as numerical integration, simulated method of moment and Bayesian methods. It should be noted that the number of transactions in one trading day can be as big as $23,400$ but is usually between 1000 and 5000 for a liquid stock. As a consequence, the integral of the likelihood function is of a very high dimension and therefore numerical integration is not feasible.

As far as we know, effective methods of moments and Bayesian methods are not developed as yet for models such as (10). For example, the Markov chain Monte Carlo (MCMC) method of Kim, Shephard, and Chib (1998), in which candidate samples are generated by approximate densities based on mixture of normals, can not be used straightforwardly.

4 Empirical results for three months of IBM prices

4.1 Data

A small subset of the the Trades and Quotes (TAQ) database for the New York Stock Exchange (NYSE) is available to us. We have considered the IBM equity transactions reported on Consolidated Tape. The IBM stock is regarded by many as a heavily traded or liquid stock. The NYSE market opens at 9:30 AM and closes at 4 PM. Prices of transactions made outside these official trading hours have been deleted. The resulting database consists of prices and times (measured in seconds) of transactions realised in the three months of November 2002, December 2002 and January 2003. No further manipulations have been carried out on this dataset. The prices for each trading day are considered as a time series with the time index in seconds. Such time series have possibly many missing observations. For example, when no trade has taken place in the last two minutes, we have at least 120 consecutive missing values in the series.

4.2 Measuring actual volatility for one day

As a first illustration we consider tick-by-tick prices and returns of IBM realised in the NYSE trading day of November 1, 2002. In Figure 1 the prices and returns are presented for the hourly intervals of the trading day. The number of trades that has taken place on this day is $3,321$. Given that a trading day consists of $23,400$ seconds, that is 6.5 trading hours times 3600 seconds in one hour, the average duration between trades is 7.05 seconds. In other words, on average, 511 trades in one hour and 8.5 trades in one minute has been realised. However, approximately, the first 300 trades took place before 10 am and the last 600 trades took place after 3 pm. The time series of prices and returns presented in Figure 1 are against an index of seconds. This means that $23,400$ observations can be displayed but only $3,321$ transactions have been realised, resulting in $20,079$ missing values on this day. We note that no multiple trades occurred in the same second. These facts aim to put the plots of Figure 1 into some perspective. Due to the lack of resolution in our graphs, the majority of missing values go almost unnoticed.

In Figure 2 we present prices, returns and squared log returns of the IBM stock for November 1, 2002. Here the index is trade by trade. Nevertheless, the series of prices in Figures 1 and 2 appear to be very similar. This is again due to the limited resolution that can be provided in these graphs. In any case, both plots of returns show that volatility is substantially higher at the beginning of the trading day and somewhat higher at the end of the trade session. The small price variation in the middle of the trading day is probably due to the fact that no relevant information has arrived in these hours.
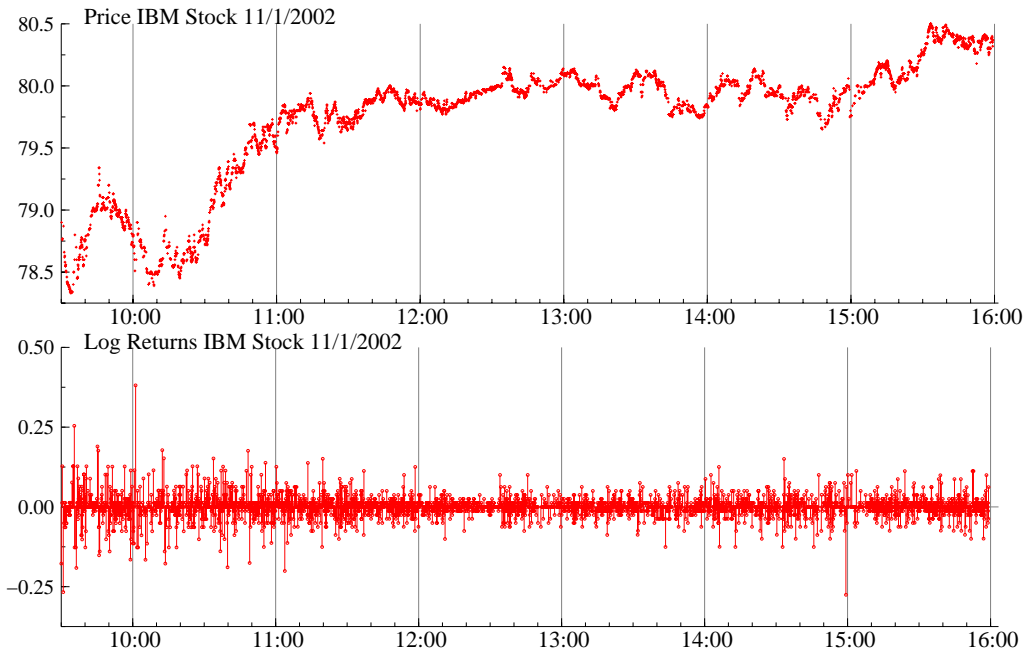
Figure 1: IBM stock prices and returns for all trades on November 1, 2002. The tick-by-tick data is presented against time in seconds.

To analyse the trade prices on this day, we first consider the model for prices (5) with constant volatility σ and intra-daily pattern $g(t)$ for spot volatility. For the function of $g(t)$ we adopted the cubic spline as described in Appendix A with three knots (two at either end of the trading day and one in the middle of the day, the first knot value is restricted to be zero so that $g(t_0) = 0$. It is argued in section 2.1 that the standard Kalman filter can be used for the estimation of the coefficients of this model. The Kalman filter as implemented in the `SsfPack` package of Koopman, Shephard, and Doornik (1999) allows for missing values and deterministic time-varying variances. We have implemented the calculations in the `Ox` package of Doornik (2001) using the state space library of `SsfPack`, version 3. The estimation results are as follows with

$$\log \widehat{\sigma} = -5.112, \quad \widehat{\gamma}_2 = -1.747, \quad \widehat{\gamma}_3 = -1.135.$$

These results give some initial indication of results that can be obtained from a high-frequency dataset.

More interestingly from theoretical and empirical perspectives are the results for the returns model with stochastic volatility and intra-daily seasonality. In particular, we are focussed on the differences in the estimation results for this model with or without micro-structure noise. The estimation method for the model with noise requires importance sampling techniques as
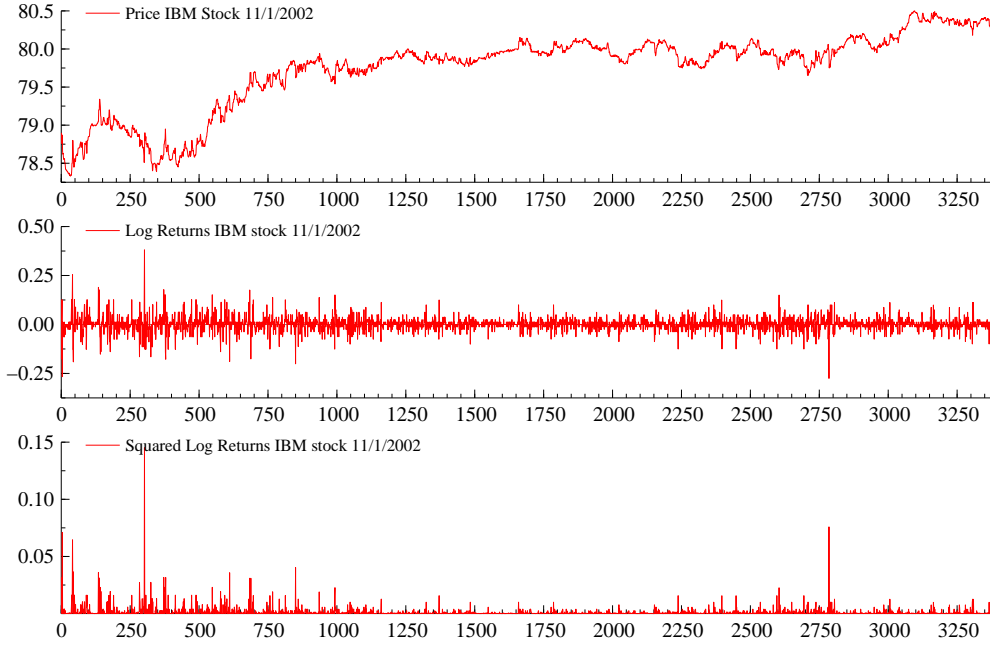
13

Figure 2: IBM stock prices, returns and squared log returns for all trades on November 1, 2002. The data is presented against the trade index so that every tick is one trade.

discussed in Section 3.2. The calculations are implemented in `Ox` with intensive use of the `SsfPack` library of state space functions to obtain an approximating model and to simulate conditional random samples of log volatility.

The estimates for the parameters are as follows. For the model (12) and (13) without micro-structure noise ($\sigma_U = 0$), we have

$$\widehat{\phi} = 0.961, \quad \widehat{\sigma}_\eta^2 = 0.0619,$$
$$\log \widehat{\sigma} = -7.977 \quad \widehat{\gamma}_2 = -1.654, \quad \widehat{\gamma}_3 = -1.135.$$

For the model with micro-structure noise, we have

$$\widehat{\sigma}_U^2 = 0.00003985, \quad \widehat{\sigma}_U = 0.00631,$$
$$\widehat{\phi} = 0.955, \quad \widehat{\sigma}_\eta^2 = 0.0821,$$
$$\log \widehat{\sigma} = -8.033, \quad \widehat{\gamma}_2 = -1.629, \quad \widehat{\gamma}_3 = -1.065.$$

When these estimation results are compared with the constant plus spline volatility model, the constant variation drops to make place for the stochastic part of log volatility. The persistence of log volatility is of the same order when model is estimated with noise or without noise. Although apparently the micro-structure noise seems low, it has a big impact on the estimate
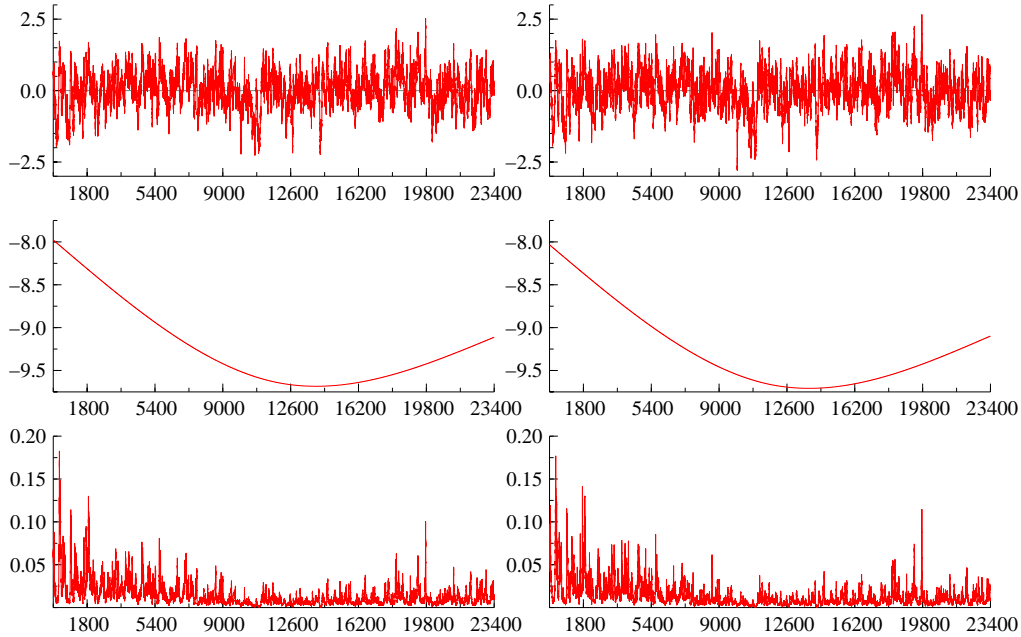
14

Figure 3: Estimated intra-day volatility: (i) log-volatility component $\log \widehat{\sigma}'$; (ii) intra-daily volatility pattern $\widehat{g}(t)$; (iii) integrated volatility $\sigma^{*2}(t_{n-1}, t_n)$.

$\widehat{\sigma}_\eta$. The estimate increases when micro noise is accounted for. Relative more variation is attributed to the stochastic part rather than the constant part when micro-noise is excluded from the observed returns.

In Figure 3 we present the estimated volatility components for this day. The time series length is $23,400$ seconds for which $20,079$ seconds have recorded no price. During the model estimation process, these $20,079$ non-available prices are treated as missing observations. The estimated prices and returns are obtained using the importance sampling methods for filtering and smoothing, see Durbin and Koopman (2001, Chapter 11) and Appendix B for further details. As a result we obtained $23,400$ estimates for which the vast majority are interpolations implied by the estimated model, see Figure 3. To provide a somewhat more detailed insight, we also present estimates of $\log \sigma'_t$ for a smaller interval of 30 minutes and four intervals of 5 minutes in Figures 4 and 5, respectively.

4.3 Measuring actual volatility for three months

We convinced ourselves that the model-based methods for estimating coefficients and for measuring volatility were implemented satisfactory given the results of the previous section. Also several limited simulation studies have been carried out and the results confirmed the reliability
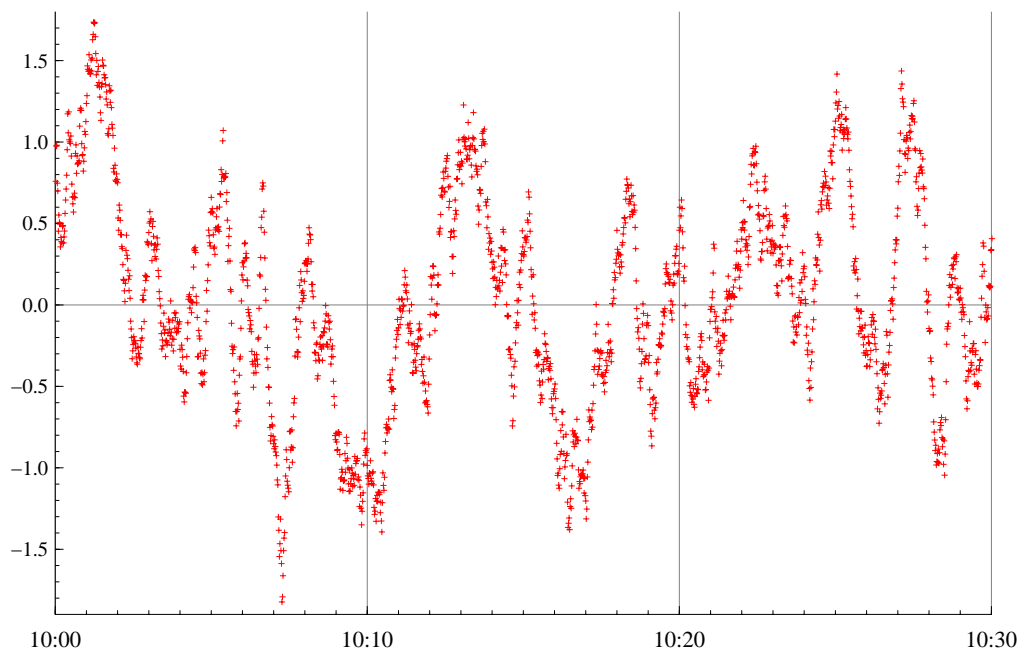
Figure 4: Estimated log-volatility component $\log \widehat{\sigma}'$ for an interval of 30 minutes.
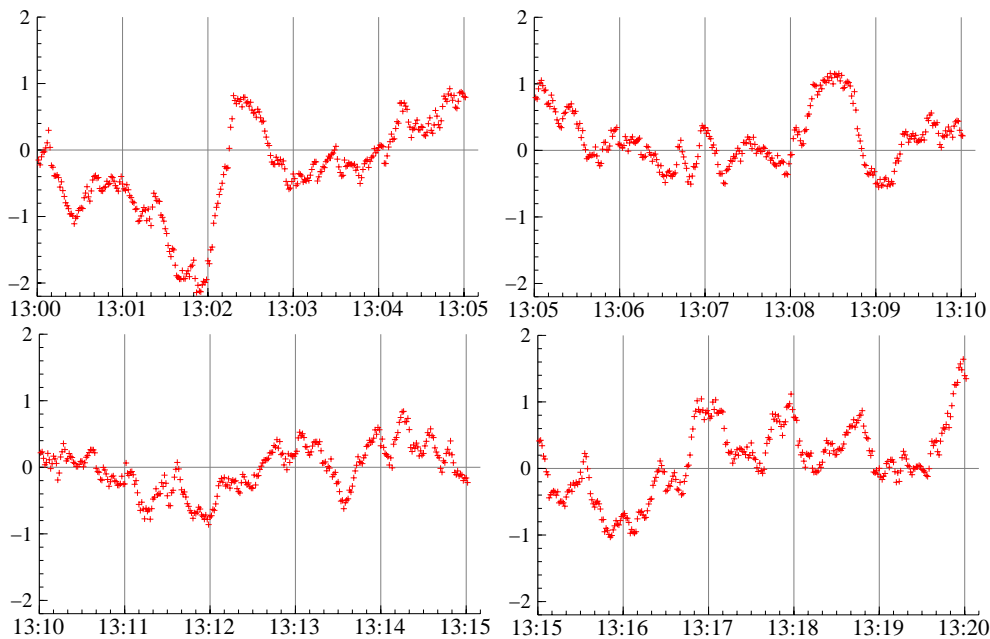

Figure 5: Estimated log-volatility component $\log \widehat{\sigma}'$ for four intervals of 5 minutes.
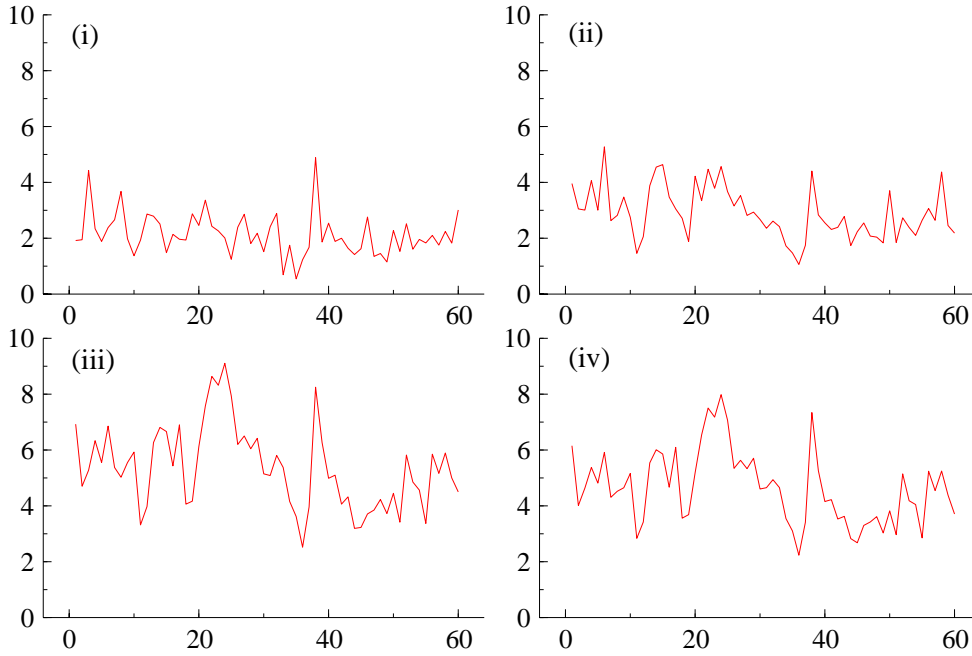
16

Figure 6: Volatility measures: (i) realised volatility; (ii) estimates from constant volatility model; (iii) estimates from constant plus spline volatility model; (iv) estimates from stochastic volatility model with spline. The volatility estimates are for the 61 trading days of November 2002, December 2002 and January 2003.

of the implemented procedures. Subsequently we repeated the analysis for a large dataset of IBM stock returns for 61 consecutive trading days in the months of November 2002, December 2002 and January 2003. We present in Figure 6 the measures obtained from standard realised volatility calculations, model with constant volatility, model with constant plus spline volatility and model with constant, spline and stochastic volatility.

The patterns of the volatility measures are similar. The overall level of measures obtained by models with SV are higher compared to realised volatility and the measure obtained from model with constant volatility. For the case of realised volatility, this can be explained by the fact that variations between, say, 5 minutes are not considered whereas in our SV modelling framework all variations within the day are used. This clearly leads to higher estimates of volatility. In the case of the model with constant volatility, the estimates are lower since they are close to a mean of squared log returns which means that excessive variations are averaged out.

The difference between the models with SV, one with micro-structure noise and another without noise, seems relatively small. However it should be bared in mind that the number
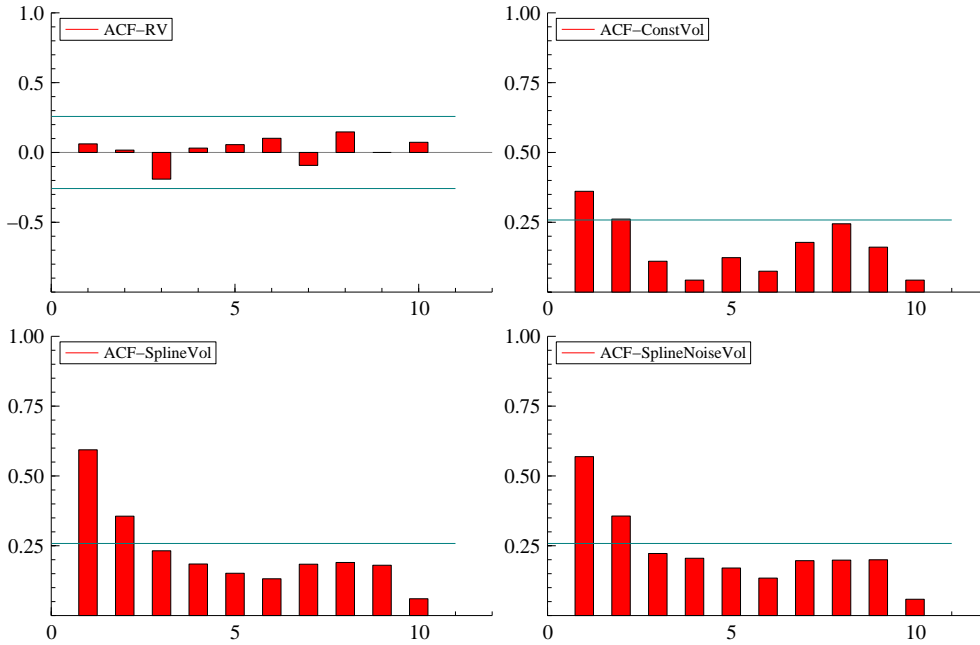
Figure 7: Sample autocorrelation functions of the volatility measures from Figure 6. The correlation coefficients are therefore based on 61 datapoints.

of trades in each day are between 2000 and 5000. It is clear that the volatility estimates for models with SV and noise are somewhat lower compared to models with SV but without noise as the former model attributes part of the noise to micro-structure effects.

Finally, we display the sample autocorrelation functions for the daily volatility measures in Figure 7. Although it is somewhat surprising that the correlogram for realised volatility is not significant at any lag despite the widely accepted view that realised volatility is serially correlated and can effectively modelled as an autoregressive fractional integrated moving average (ARFIMA) process, see, for example, Andersen, Bollerslev, Diebold, and Labys (2003). However, for the realised volatility series analysed in Koopman, Jungbacker, and Hol (2005), many instances are encountered where the correlogram is also not significant when random subsamples of length 100 are considered. Note that for the full sample of 1500 daily realised volatilies, a significantly persistent correlogram is present. In the analysis of this paper, it appears that model-based measures of volatility are persistent over days, especially when stochastic volatility is modelled explicitly. The daily time series of model-based volatility measures are relatively smooth and show to contain some level of persistency.

18

5 Discussion and conclusion

We have proposed to measure volatility from high-frequency prices using a model-based framework. A standard basic model is considered that captures the salient features of prices and volatilities in financial markets. In particular, it accounts for micro-structure noise, an intra-daily volatility pattern and stochastic volatility. Feasible estimation methods have been implemented for this class of models and the illustration shows that this approach can work effectively in determining the volatility in financial markets using tick-by-tick data. As a result, no information is lost as opposed to realised volatility for which prices are sampled at a low frequency, say 1 or 5 minutes. Therefore a part of the variation in prices is lost in realised volatility. When more detailed comparisons are made between realised volatility and the high-frequency measures, it is shown that the supposed long memory property of realised volatility may not be identified from a realtive small number of days whereas for high-frequency measures the persistence of daily volatility estimates remain. However, more empirical investigation is needed to obtain further insights on this issue. Nonparametric methods have also been proposed recently to tackle the problem of micro-structure noise. However, as far as we know, this paper presents a first attempt to analyse ultra high-frequency prices using a model that simultaneously accounts for micro-structure noise and stochastic volatility.

Appendix A: Cubic spline for intra-daily volatility pattern

The intra-daily pattern of volatility is captured by a flexible function $g(t)$. In this paper we take $g(t)$ as a cubic spline function. We follow Poirier (1976) in developing a cubic spline. Given a mesh of, say 3, x values ($\{x_0, x_1, x_2\}$) and a set of corresponding y values ($\{y_0, y_1, y_2\}$, respectively), the y values for $x_{j-1} \leq x \leq x_j$ can be interpolated by

$$y = g(x) = \frac{(x_j - x)^3}{6(x_j - x_{j-1})} z_{1,j-1} + \frac{(x - x_{j-1})^3}{6(x_j - x_{j-1})} z_{1,j} + (x_j - x) z_{2,j} + (x - x_{j-1}) z_{3,j} + z_{4,j}, \qquad j = 1, 2,$$

where $z_{i,j}$ are unknown coefficients for $i = 1, 2, 3, 4$ and $j = 1, 2$. The coefficients $z_{i,j}$ are determined by restricting smoothness conditions on $g(x)$ such as continuity at x_j ($j = 1, 2$) of the spline itself and its first and second derivatives. The resulting set of equations can be solved in $z_{i,j}$ via standard matrix algebra. Given this solution for $z_{i,j}$, the spline function can be expressed as

$$g(x) = \sum_{j=0}^{2} w_j y_j, \qquad \sum_{j=0}^{2} w_j = 1,$$

where weights w_j depend on x and the mesh $\{x_0, x_1, x_2\}$.

Appendix B: Approximating model for SV with noise

Consider a non-linear state-space model where the state equation is linear and Gaussian, but the distribution of the observations $Y = (Y_1, \ldots, Y_N)$ conditional on the states $h = (h_1, \ldots, h_N)$ is determined by a family of probability densities $p(Y_n|h_n)$, $n = 1, \ldots, N$. It is evident that the SV models considered in the main text are special cases of this class of models, the interested reader is referred to Durbin and Koopman (2001) for further examples. For the importance sampling procedure a linear Gaussian approximating model is chosen with the same state equation as the true model but with an observation equation given by

$$Y_n = c_n + h_n + u_n, \qquad u_n \sim \mathrm{NID}(0, V_n), \tag{17}$$

where the constants c_n and V_n have to be chosen in a suitable manner. The approach advocated in Durbin and Koopman (2001) consists of choosing c_n and V_n for $n = 1, \ldots, N$ such that the true smoothing density, $p(h|Y)$, and the smoothing density of the approximating model, $g(h|Y; V; c)$, have the same mode and have equal curvature around this mode. This means, denoting $V = (V_1, \ldots, V_n)'$ and $c = (c_1, \ldots, c_n)'$, that V and c are solutions to the system of equations defined by

$$\frac{\partial p(Y, h)}{\partial h_n} = \frac{\partial g(Y, h; c, V)}{\partial h_n} = 0,$$

and

$$\frac{\partial^2 p(Y, h)}{\partial h_n^2} = \frac{\partial^2 g(Y, h; c, V)}{\partial h_n^2},$$

for $n = 1, \ldots, N$. The key to solving these equations and finding c and V lies in two observations. First of all, it is a well known fact that for a Gaussian distribution the mode simply equals the mean. This means that, conditional on c and V, the mode, $\hat{h} = (\hat{h}_1, \ldots, \hat{h}_n)$, can be obtained by computing the mean of $g(h|V; c)$, a problem that is routinely handled by the Kalman filter and smoother. On the other hand the fact that the marginal distribution of h is equal for both the true as well as the approximating model, combined with the monotonicity of the log transformation implies that the system of equations is equivalent to

$$\frac{\partial \log p(Y|h)}{\partial h_i} = \frac{\partial \log g(Y|h; c, V)}{\partial h_i} = 0,$$

and

$$\frac{\partial^2 \log p(Y|h)}{\partial h_i^2} = \frac{\partial^2 \log g(Y|h, c, V)}{\partial h_i^2},$$

implying that conditional on the mode $\hat{h}$ a solution to this set of equations is given by the vectors V and c satisfying

$$\left. \frac{\partial \log p(Y_n|h_n)}{\partial h_n} \right|_{h_n = \hat{h}_n} = \left. \frac{\partial \log g(Y_n|h_n; c, V)}{\partial h_n} \right|_{h_n = \hat{h}_n}$$

20

and

$$\left.\frac{\partial^2 \log p(Y_n|h_n)}{\partial h_n^2}\right|_{h_n=\hat{h}_n} = \left.\frac{\partial^2 \log g(Y_n|h_n; c, V)}{\partial h_n^2}\right|_{h_n=\hat{h}_n},$$

for $n = 1, \ldots, N$. If we now use

$$\frac{\partial \log g(Y_n|h_n; c, V)}{\partial h_i} = \frac{Y_n - h_n - c_n}{V_n},$$

and

$$\frac{\partial^2 \log g(Y_n|h_n; c, V)}{\partial h_n^2} = \frac{1}{V_n},$$

then these expressions imply

$$V_i = \left(\frac{\partial^2 \log p(Y_i|h_i)}{\partial h_i^2}\right)^{-1}, \tag{18}$$

and

$$c_i = Y_i - h_i - V_i\frac{\partial \log p(Y_i|h_i)}{\partial h_i}. \tag{19}$$

These two observations suggest the following algorithm

1. Choose a starting value h^1 for $\hat{h}$.

2. For $i = 1, 2, \ldots$ use h_i to obtain c^i and V^i using (18) and (19). Create a new proposal for $\hat{h}$, h^{i+1}, by applying the Kalman smoother to $Y_1, \ldots, Y_n$ for the model defined by (17), with $c = c^i$ and $V = V^i$.

3. Keep repeating 2 until $\|h^{i+1} - h^i\| < \epsilon_c$, where ϵ_c is some small threshold value.

To implement this algorithm for the Stochastic Volatility models considered in this paper the only thing that remains to be done is to calculate the derivatives in (18) and (19). For the SV model defined in (10) these derivatives can easily be seen to be

$$\frac{\partial \log p(Y_n|h_n)}{\partial h_n} = \frac{1}{2}\left(\frac{Y_n^2}{\exp h_n} - 1\right), \qquad \frac{\partial^2 p(Y_n|h_n)}{\partial h_n^2} = -\frac{Y_n^2}{2\exp h_n},$$

whereas for the SV model with micro-structure noise defined in (11) we get

$$\frac{\partial \log p(Y_n|h_n)}{\partial h_n} = \frac{1}{2}b_n\left(\frac{Y_n^2}{a_n} - 1\right), \qquad \frac{\partial^2 \log p(Y_n|h_n)}{\partial h_n^2} = \left(\frac{1}{2} - b_n\right)\frac{b_n Y_n^2}{a_n} - \frac{1}{2}\left(b_n - b_n^2\right),$$

where $a_n = \exp(h_n) + \sigma_U^2$ and $b_n = \exp(h_n) / a_n$.

References

Aït-Sahalia, Y., P. A. Mykland, and L. Zhang (2004). How often to sample a continuous-time proces in the presence of market microstructure noise. *Review of Financial Studies 17*, forthcoming.

Andersen, T. G. and T. Bollerslev (1997). Intraday periodicity and volatility persistence in financial markets. *J. Empirical Finance 4*, 115–58.

Andersen, T. G., T. Bollerslev, and F. X. Diebold (2002). Parametric and nonparametric volatility measurement. In *Handbook of Financial Econometrics*, pp. forthcoming. Amsterdam: North-Holland.

Andersen, T. G., T. Bollerslev, F. X. Diebold, and P. Labys (2001). The distribution of exchange rate volatility. *Journal of the American Statistical Association 96*, 42–55.

Andersen, T. G., T. Bollerslev, F. X. Diebold, and P. Labys (2003). Modeling and forecasting realized volatility. *Econometrica 71*, 579–625.

Andreou, E. and E. Ghysels (2001). Rolling sample volatility estimators: Some new theoretical, simulation and empirical results. *Journal of Business and Economic Statistics 20*, 363–376.

Bai, X., J. R. Russell, and G. C. Tiao (2000). Beyond mertons utopia: Effects of nonnormality and dependence on the precision of variance estimates using high-frequency financial data. Manuscript, Graduate School of Business, University of Chicago.

Bandi, F. M. and J. R. Russell (2004). Microstructure noise, realized volatility, and optimal sampling. Manuscript, Graduate School of Business, University of Chicago.

Barndorff-Nielsen, O. E., P. R. Hansen, A. Lunde, and N. Shephard (2004). Regular and modified kernel-based estimators of integrated variance: The case with independent noise. Working paper, Nuffield College.

Barndorff-Nielsen, O. E. and N. Shephard (2001). Econometric analysis of realised volatility and its use in estimating stochastic volatility models. *Journal of Royal Statistical Society B 63*, 253–280.

Campbell, J. Y., A. W. Lo, and A. C. MacKinlay (1997). *The Econometrics of Financial Markets*. Princeton, New Jersey: Princeton University Press.

Dacarogna, M. M., U. A. Müller, R. J. Nagler, R. B. Olsen, and O. V. Pictet (1993). A geographical model for the daily and weekly seasonal volatility in the foreign exchange market. *Journal of International Money and Financial Economics 12*, 413–438.

de Jong, P. and N. Shephard (1995). The simulation smoother for time series models. *Biometrika 82*, 339–50.

Doornik, J. A. (2001). *Object-Oriented Matrix Programming using Ox 3.0* (4th ed.). London: Timberlake Consultants Ltd. `http://www.nuff.ox.ac.uk/Users/Doornik`.

Durbin, J. and S. J. Koopman (1997). Monte Carlo maximum likelihood estimation of non-Gaussian state space model. *Biometrika 84*, 669–84.

Durbin, J. and S. J. Koopman (2001). *Time Series Analysis by State Space Methods*. Oxford: Oxford University Press.

Durbin, J. and S. J. Koopman (2002). A simple and efficient simulation smoother for state space time series analysis. *Biometrika 89*, 603–16.

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica 57*, 1317–39.

Ghysels, E., A. Harvey, and E. Renault (1996). Stochastic volatility. In *Handbook of Statistics Vol. 14, Statistical Methods in Finance*, pp. 119–191. Amsterdam: North-Holland.

Hansen, P. R. and A. Lunde (2004). A realized variance for the whole dat based on intermittent high-frequency data. Working paper, Stanford University.

Harvey, A. C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge: Cambridge University Press.

Harvey, A. C. and S. J. Koopman (2000). Signal extraction and the formulation of unobserved components models. *Econometrics Journal 3*, 84–107.

Kim, S., N. Shephard, and S. Chib (1998). Stochastic volatility: likelihood inference and comparison with ARCH models. *Rev. Economic Studies 65*, 361–93.

Kloeden, P. and E. Platen (1999). *Numerical Solution of Stochastic Differential Equations*. Berlin: Springer-Verlag.

Koopman, S. J., B. Jungbacker, and E. Hol (2005). Forecasting daily variability of the S&P 100 stock index using historical, realised and implied volatility measurements. *Journal of Empirical Finance 12*, forthcoming.

Koopman, S. J., N. Shephard, and J. A. Doornik (1999). Statistical algorithms for models in state space form using SsfPack 2.2. *Econometrics Journal 2*, 113–66. `http://www.ssfpack.com/`.

Malliavin, P. and M. Mancino (2002). Fourier series method for measurment of multivariate volatilities. *Finance and Stochastics 6*, 49–61.

Poirier, D. J. (1976). *The Econometrics of Structural Change: with Special Emphasis on Spline Functions.* Amsterdam: North-Holland.

Shephard, N. (2005). *Stochastic Volatility.* Oxford: Oxford University Press.

Shephard, N. and M. K. Pitt (1997). Likelihood analysis of non-Gaussian measurement time series. *Biometrika 84*, 653–67.

Zhang, L., P. A. Mykland, and Y. Aït-Sahalia (2004). A tale of two time scales: Determining integrated volatility with noisy high-frequency data. Manuscript, Graduate School of Business, University of Chicago.