Chapter 13
Volatility Estimation Based on High-Frequency Data

Christian Pigorsch, Uta Pigorsch, and Ivaylo Popov

Abstract With the availability of high-frequency data ex post daily (or lower frequency) nonparametric volatility measures have been developed, that are more precise than conventionally used volatility estimators, such as squared or absolute daily returns. The consistency of these estimators hinges on increasingly finer sampled high-frequency returns. In practice, however, the prices recorded at the very high frequency are contaminated by market microstructure noise. We provide a theoretical review and comparison of high-frequency based volatility estimators and the impact of different types of noise. In doing so we pay special focus on volatility estimators that explore different facets of high-frequency data, such as the price range, return quantiles or durations between specific levels of price changes.The various volatility estimators are applied to transaction and quotes data of the S&P500 E-mini and of one stock of Microsoft using different sampling frequencies and schemes. We further discuss potential sources of the market microstructure noise and test for its type and magnitude. Moreover, due to the volume of high-frequency financial data we focus also on computational aspects, such as data storage and retrieval.

C. Pigorsch
Department of Economics, University of Bonn, Adenauerallee 24-42, 53113 Bonn, Germany
e-mail: christian.pigorsch@uni-bonn.de

U. Pigorsch (✉)
Department of Economics, University of Mannheim, L7, 3-5, 68131 Mannheim, Germany
e-mail: uta.pigorsch@vwl.uni-mannheim.de

I.Popov
Business School of the University of Mannheim, L5, 5, 68131 Mannheim, Germany
e-mail: ipopov@mail.uni-mannheim.de

J.-C. Duan et al. (eds.), *Handbook of Computational Finance*, Springer Handbooks of Computational Statistics, DOI 10.1007/978-3-642-17254-0_13,

13.1 Introduction

This chapter presents a review and empirical illustration of nonparametric volatility estimators that exploit the information contained in high-frequency financial data. Such ex-post volatility measures can be directly used for the modelling and forecasting of the (future) volatility dynamics, which in turn may be essential for an adequate risk management or hedging decisions. Moreover, volatility constitutes the main ingredient in asset pricing and the knowledge of this quantity therefore plays a major role in most financial applications.

One of the most recent milestones in financial econometrics is therefore probably the introduction of the concept of realized volatility, which allows to consistently estimate the price variation accumulated over some time interval, such as 1 day, by summing over squared (intraday) high-frequency returns. The consistency of this estimator hinges on increasingly finer sampled high-frequency returns. In practice, however, the sampling frequency is limited by the actual quotation or transaction frequency and prices are contaminated by market microstructure effects, so-called noise. We discuss different types and potential sources of the noise and its impact on realized volatility. We further review two of the probably most popular approaches to estimate volatility based on squares or products of high-frequency returns, i.e. the two time scales estimators and kernel-based approaches. However, our main focus in this chapter is on volatility estimators that explore different facets of high-frequency data, such as the price range, return quantiles or durations between specific levels of price changes. Our review thus differs from the one provided in McAleer and Medeiros (2008). A theoretical summary and comparison of the estimators is given. Moreover, as the high-frequency financial data exceeds the amount of data usually encountered by financial econometricians we provide a discussion on data storage and retrieval, i.e. computational aspects that may be of interest to anybody dealing with such high-frequency data. In our empirical application we estimate and illustrate realized volatility over various frequencies for different sampling schemes and price series of one future (S&P500 E-mini) and one stock (Microsoft). We test for the magnitude and type of market microstructure noise and implement the discussed volatility estimators.

13.2 Realized Volatility

Assume that the logarithmic price of a financial asset is given by the following diffusion process

$$p_t = \int_0^t \mu(s)\mathrm{d}s + \int_0^t \sigma(s)\mathrm{d}W(s), \tag{13.1}$$

where the mean process μ is continuous and of finite variation, $\sigma(t) > 0$ denotes the càdlàg instantaneous volatility and W is a standard Brownian motion. The object of interest is the *integrated variance* (IV), i.e. the amount of variation at time point t

accumulated over a past time interval Δ:

$$IV_t = \int_{t-\Delta}^{t} \sigma^2(s)\mathrm{d}s.$$

In the sequel, our focus is on the estimation of *IV* over one period, e.g. 1 day. For the ease of exposition we, thus, normalize $\Delta = 1$ and drop the time subscript. Suppose there exist m intraday returns, the ith intraday return is then defined as:

$$r_i^{(m)} = p_{i/m} - p_{(i-1)/m},\ i = 1, 2, \ldots, m.$$

The sum of the squared intraday returns:

$$RV^{(m)} = \sum_{i=1}^{m} r_i^{(m)2} \tag{13.2}$$

provides a natural estimator of *IV*. In fact, based on the theory of quadratic variation, Andersen et al. (2003) show that $RV^{(m)} \xrightarrow{p} IV$ as $m \to \infty$. Following the recent literature we will refer to this ex-post measure of *IV* as the *realized volatility*, see e.g. Andersen and Bollerslev (1998).

Barndorff-Nielsen and Shephard (2002a) show the consistency of this estimator and that its asymptotic distribution is normal:

$$\frac{\sqrt{m}\left(RV^{(m)} - IV\right)}{\sqrt{2IQ}} \xrightarrow{d} \mathcal{N}(0, 1),$$

with $IQ = \int_0^1 \sigma^4(s)\mathrm{d}s$ denoting the *integrated quarticity*. An application of this asymptotic result, e.g. the construction of confidence intervals, however, is complicated by the unobservability of *IQ*. A solution is offered in Barndorff-Nielsen and Shephard (2004), who propose the concept of realized power variation, that allows to estimate *IQ* via the *realized quarticity*:

$$RQ^{(m)} = \frac{m}{3} \sum_{i=1}^{m} r_i^{(m)4},$$

such that

$$\frac{RV^{(m)} - IV}{\sqrt{\frac{2}{3}\sum_{i=1}^{m} r_i^{(m)4}}} \xrightarrow{d} \mathcal{N}(0, 1)$$

can be used for large m. In practice, however, the sampling frequency is limited by the actual transaction or quotation frequency. Moreover, the very high-frequency prices are contaminated by market microstructure effects (noise), such as bid-ask

bounce effects, price discreteness etc., leading to biases in realized volatility, see e.g. Andersen et al. (2001) and Barndorff-Nielsen and Shephard (2002b). The next section discusses typical assumptions on the structure of the market microstructure noise and its implications for realized volatility. Section 13.4 presents modifications of the realized volatility estimator, while Sect. 13.5 focuses on estimators that exploit other data characteristics for measuring *IV*. Section 13.6 provides a comparison of the various estimators and discusses the situation where the price process (13.1) additionally exhibits finite active jumps.

13.3 Market Microstructure Noise: Assumptions and Implications

Assume that the observed (log) price is contaminated by market microstructure noise u (or measurement error), i.e.:

$$p_{i/m} = p^*_{i/m} + u_{i/m}, \quad i = 1, 2, \ldots m,$$

where $p^*_{i/m}$ is the latent true, or so-called efficient, price that follows the semimartingale given in (13.1). In this case, the observed intraday return is given by:

$$r_i^{(m)} = r_i^{*(m)} + \epsilon_i^{(m)}, \quad i = 1, 2, \ldots m,$$

i.e. by the efficient intraday return $r_i^{*(m)} = p^*_{i/m} - p^*_{(i-1)/m}$ and the intraday noise increment $\epsilon_i^{(m)} = u_{i/m} - u_{(i-1)/m}$. As a consequence, the observed RV can be decomposed as:

$$RV^{(m)} = RV^{*(m)} + 2\sum_{i=1}^{m} r_i^{*(m)} \epsilon_i{}^{(m)} + \sum_{j=1}^{m} \epsilon_j^{(m)2},$$

where the last term on the right-hand side can be interpreted as the (unobservable) realized variance of the noise process, while the second term is induced by potential dependence between the efficient price and the noise. Based on this decomposition and the assumption of covariance stationary noise with mean zero, Hansen and Lunde (2006) show that *RV* is a biased estimator of *IV*. Interestingly, this bias is positive if the noise increments and the returns are uncorrelated, but may become negative in the case of negative correlation. One possible explanation for such negative correlation is given in Hansen and Lunde (2006), who show that in price series compiled from mid-quotes (see Sect. 13.7 for a definition), this can be caused by non-synchronous revisions of the bid and the ask prices, leading to a temporary widening of the spread. Another source of negative correlation may be the staleness of the mid-quote prices.

Obviously, the precise implications of the presence of noise for the properties of the *RV* estimator depend on the assumed structure of the noise process. In the following we focus on the most popular noise assumption.

Assumption 1: Independent noise.

(a) The noise process u is independent and identically distributed with mean zero and finite variance ω^2 and finite fourth moment.
(b) The noise is independent of the efficient price.

The independent noise assumption implies that the intraday returns have an MA(1) component. Such a return specification is well established in the market microstructure literature and is usually justified by the existence of the bid-ask bounce effect, see e.g. Roll (1984). However, as shown in Hansen and Lunde (2006) and Zhang et al. (2005) the iid noise introduces a bias into the RV estimator:

$$E\left[RV^{(m)}\right] = IV + 2m\omega^2 \tag{13.3}$$

that diverges to infinity as $m \to \infty$. Moreover, the asymptotic distribution of *RV* is given by:

$$\frac{\left(RV^{(m)} - IV - 2m\omega^2\right)}{2\sqrt{mE(u^4)}} \xrightarrow{d} \mathcal{N}(0,1)\,.$$

Sampling at lower frequencies, i.e. sparse sampling, reduces the bias but leads to an increase in the variance (see e.g. Barndorff-Nielsen and Shephard 2002b), which is usually referred to as the bias-variance trade-off.

The independent noise assumption seems restrictive. In fact, Hansen and Lunde (2006) provide some evidence of serial dependence in the noise process and correlation with the efficient price, i.e. *time-dependent* and *endogenous* noise, respectively. Alternative estimators of *IV* have been developed and are shown to be robust to some dependence in the noise process, but they are in no way developed around a universally accepted dependence specification like Assumption 1. The next section discusses the probably most popular alternatives to the *RV* estimator that are asymptotically unbiased and consistent under iid and under dependent noise types.

13.4 Subsampling and Realized Kernels

In the following we briefly present two more elaborate, but under specific noise assumptions consistent procedures for estimating *IV*.

13.4.1 Averaging and Subsampling

The subsampling approach originally suggested by Zhang et al. (2005) builds on the idea of averaging over various *RV*s constructed by sampling sparsely over high-frequency subsamples. To this end the intraday observations are allocated to K subsamples. Using a regular allocation, 5 min returns can for example be sampled at the time points 9:30, 9:35, 9:40, ...; and at the time points 9:31, 9:36, 9:41, ... and so forth. Averaging over the subsample *RV*s yields the so-called *average RV* estimator: $(1/K)\sum_{k=1}^{K} RV^{(k,m_k)}$ with m_k denoting the sampling frequency used in the RV computation for subsample k. Usually, m_k is equal across all subsamples. The average RV estimator is still biased, but the bias now depends on the average size of the subsamples rather than on the total number of observations. RV constructed from all observations, $RV^{(all)}$ can be used for bias correction yielding the estimator:

$$TTSRV^{(m,m_1,\ldots,m_K,K)} = \frac{1}{K}\sum_{k=1}^{K} RV^{(k,m_k)} - \frac{\bar{m}}{m} RV^{(all)}, \tag{13.4}$$

where $\bar{m} = (1/K)\sum_{k=1}^{K} m_k$. As the estimator (13.4) consists of a component based on sparsely sampled data and one based on the full grid of price observations, the estimator is also called *two time scales estimator*.

Under the independent noise assumption, the estimator is consistent. Furthermore, under equidistant observations and under regular allocation to the grids, the asymptotic distribution is given by:

$$\frac{m^{1/6}(TTSRV^{(m,m_1,\ldots,m_K,K)} - IV)}{\sqrt{\frac{8}{c^2}(\omega^2)^2 + c\frac{4}{3}IQ}} \xrightarrow{d} \mathcal{N}(0,1).$$

for $K = cm^{2/3}$. The optimal value of K, i.e. minimizing the expected asymptotic variance, can be obtained by estimating $c_{opt} = \left(12\omega^2/IQ\right)^{1/3}$ based on data prior to the day under consideration (see Zhang et al. 2005).

A generalization of *TTSRV* was introduced by Aït-Sahalia et al. (2010) and Zhang (2006), which is consistent and asymptotically unbiased also under time-dependent noise. To account for serial correlation in the noise, the *RV*s are based on overlapping J-period intraday returns. Using these so-called *average-lag-J RV*s the estimator becomes:

$$TTSRV_{adj}^{(m,K,J)} = s\left(\frac{1}{K}\sum_{i=0}^{m-K}(p_{(i+K)/m} - p_{i/m})^2 - \frac{\bar{m}^{(K)}}{\bar{m}^{(J)}}\frac{1}{J}\sum_{l=0}^{m-J}(p_{(l+J)/m} - p_{l/m})^2\right) \tag{13.5}$$

with $\bar{m}^{(K)} = (m - K + 1)/K$, $\bar{m}^{(J)} = (m - J + 1)/J$, $1 \leq J < K < m$ and the small sample adjustment factor $s = \left(1 - \bar{m}^{(K)}/\bar{m}^{(J)}\right)^{-1}$. Note that K and J now basically denote the slow and fast time scales, respectively. The asymptotic distribution is given by:

$$\frac{m^{1/6}\left(TTSRV_{adj}^{(m,K,J)} - IV\right)}{\sqrt{\frac{1}{c^2}\xi^2 + c\frac{4}{3}IQ}} \overset{d}{\approx} \mathcal{N}(0,1)$$

with $\xi^2 = 16\left(\omega^2\right)^2 + 32\sum_{l=1}^{\infty}\left(E(u_0, u_l)\right)^2$ and $\overset{d}{\approx}$ denotes that when multiplied by a suitable factor, then the convergence is in distribution.

Obviously, $TTSRV_{adj}$ converges to IV at rate $m^{1/6}$, which is below the rate of $m^{1/4}$, established as optimal in the fully parametric case in Aït-Sahalia et al. (2005). As a consequence, Aït-Sahalia et al. (2010) introduced the multiple time scale estimator, *MTSRV*, which is based on the weighted average of average-lag-J *RV*s computed over different multiple scales. It is computationally more complex, but for suitably selected weights it attains the optimal convergence rate $m^{1/4}$.

13.4.2 Kernel-Based Estimators

Given the similarity to the problem of estimating the long-run variance of a stationary time series in the presence of autocorrelation, it is not surprising that kernel-based methods have been developed for the estimation of *IV*. Such an approach was first adopted in Zhou (1996) and generalized in Hansen and Lunde (2006), who propose to estimate *IV* by:

$$KRV_{Z\&HL}^{(m,H)} = RV^{(m)} + 2\sum_{h=1}^{H}\frac{m}{m-h}\gamma_h$$

with $\gamma_h = \sum_{i=1}^{m} r_i^{(m)} r_{i+h}^{(m)}$. As the bias correction factor $m/(m-h)$ increases the variance of the estimator, Hansen and Lunde (2006) replaced it by the Bartlett kernel. Nevertheless, all three estimators are inconsistent.

Recently, Barndorff-Nielsen et al. (2008) proposed a class of consistent kernel based estimators, *realized kernels*. The *flat-top realized kernel*:

$$KRV_{FT}^{(m,H)} = RV^{(m)} + \sum_{h=1}^{H} k\left(\frac{h-1}{H}\right)\left(\gamma_h + \gamma_{-h}\right),$$

where $k(x)$ for $x \in [0,1]$ is a deterministic weight function. If $k(0) = 1$, $k(1) = 0$ and $H = cm^{2/3}$ the estimator is asymptotically mixed normal and converges at

rate $m^{1/6}$. The constant c is a function of the kernel and the integrated quarticity, and is chosen such that the asymptotic variance of the estimator is minimized. Note that for the flat-top Bartlett kernel, where $k(x) = 1 - x$, and the cubic kernel, $k = 1 - 3x^2 + 2x^3$, $KRV_{FT}^{(m,H)}$ has the same asymptotic distribution as the TTSRV and the MTSRV estimators, respectively.

Furthermore, if $H = cm^{1/2}$, $k'(0) = 0$ and $k'(1) = 0$ (called *smooth* kernel functions), the convergence rate becomes $m^{1/4}$ and the asymptotic distribution is given by:

$$\frac{m^{1/4}\left(KRV_{FT}^{(m,H)} - IV\right)}{\sqrt{4ck_\circ IQ + \frac{8}{c}k'_\circ\omega^2 IV + \frac{4}{c^3}k''_\circ\omega^4}} \xrightarrow{d} \mathcal{N}(0,1)$$

with $k_\circ = \int_0^1 k(x)^2 \mathrm{d}x$, $k'_\circ = \int_0^1 k'(x)^2 \mathrm{d}x$ and $k''_\circ = \int_0^1 k''(x)^2 \mathrm{d}x$.

For practical applications, Barndorff-Nielsen et al. (2009) consider the *non-flat-top realized kernels*, which are robust to serial dependent noise and to dependence between noise and efficient price. The estimator is defined as:

$$KRV_{NFT}^{(m,H)} = RV^{(m)} + \sum_{h=1}^{H} k\left(\frac{h}{H}\right)(\gamma_h + \gamma_{-h}). \tag{13.6}$$

However, the above mentioned advantages of this estimator come at the cost of a lower convergence rate, i.e. $m^{1/5}$, and a small asymptotic bias:

$$m^{1/5}\left(KRV_{NFT}^{(m,H)} - IV\right) \xrightarrow{ds} \mathcal{MN}\left(c^{-2}\left|k''(0)\right|\omega^2,\ 4ck_\circ IQ\right),$$

where ds denotes stable convergence and $\mathcal{MN}$ a mixed normal distribution. Barndorff-Nielsen et al. (2009) recommend the use of the Parzen kernel as it is smooth and always produces non-negative estimates. The kernel is given by:

$$k(x) = \begin{cases} 1 - 6x^2 + 6x^3 & \text{for } 0 \le x < 1/2 \\ 2(1-x)^3 & \text{for } 1/2 \le x \le 1 \\ 0 & \text{for } x > 1 \end{cases}. \tag{13.7}$$

For non-flat-top realized kernels, the bandwidth H can be optimally selected as:

$$H^* = c^*\xi^{4/5}m^{3/5}, \quad c^* = \left(\frac{k''(0)^2}{k_\circ}\right)^{1/5} \quad \text{and} \quad \xi^2 = \frac{\omega^2}{\sqrt{IQ}}.$$

For the Parzen kernel $c^* = 3.5134$. Obviously, the optimal value of H is larger if the variance of the microstructure noise is large in comparison to the integrated

quarticity. The estimation of this signal-to-noise ratio ξ^2 is discussed in Barndorff-Nielsen et al. (2008, 2009), see also Sect. 13.8.

Realized kernels are subject to the so-called *end effects*, caused by the missing sample size adjustment of the autocovariance terms. This can be accounted for by using local averages of returns in the beginning and the end of the sample. However, Barndorff-Nielsen et al. (2009) argue that for actively traded assets these effects can be ignored in practice.

Further refinements of the realized kernels in the spirit of the subsampling approach adopted in the *TTSRV* and *MTSRV* estimators are considered in Barndorff-Nielsen et al. (2010) by using averaged covariance terms in the realized kernel estimators.

13.5 Alternative Volatility Estimators

All of the realized variance measures discussed so far are based on squared intraday returns. In the following we present estimators of the quadratic variation that exploit other aspects of high-frequency financial data.

13.5.1 Range-Based Estimation

In volatility estimation, the usage of the range, i.e. the difference between high and low (log) prices, is appealing, as it is based on extremes from the entire price path and, thus, provides more information than returns sampled at fixed time intervals. The range-based estimator has therefore attracted researcher's interest, see e.g. Feller (1951), Garman and Klass (1980), Parkinson (1980), and it has been found that using the squared range based on the daily high and low is about five times more efficient than the daily squared return. Nevertheless, it is less efficient than *RV* based on a sampling frequency higher than two hours.

Recently, Christensen and Podolskij (2007) proposed a *realized range-based estimator*, that replaces the squared intraday returns by normalized squared ranges. Assume that the (log) price process follows a continuous semimartingale and that $m_K K + 1$ equidistant prices are observed discretely over a day. Decomposing the daily time interval into K *non-overlapping* intervals of size m_K, the estimator is given by:

$$RRV^{(m_K,K)} = \frac{1}{\lambda_{2,m_K}} \sum_{i=1}^{K} s_i^{(m_K)^2}, \tag{13.8}$$

where the range of the price process over the ith interval is defined as:

$$s_i^{(m_K)} = \max_{0 \le h,l \le m_K} \left(p_{\frac{i-1+h/m_K}{K}} - p_{\frac{i-1+l/m_K}{K}} \right), \quad i = 1, \dots K,$$

and $\lambda_{r,m_K} = E\left[\max_{0\leq h,l\leq m_K}\left(W_{h/m_K} - W_{l/m_K}\right)^r\right]$. I.e. λ_{2,m_K} is the second moment of the range of a standard Brownian motion over the unit interval with m_K observed increments. This factor corrects for the downward bias arising from discretely observed data. In particular, the observed high and low prices may under- and overestimate the true ones, respectively, such that the true range is underestimated.

The estimator is asymptotically distributed according to:

$$\frac{\sqrt{K}\left(RRV^{(m_K,K)} - IV\right)}{\sqrt{\Lambda_c IQ}} \xrightarrow{d} \mathcal{N}(0,1)$$

as $K \to \infty$, where it is sufficient that m_K converges to a natural number c, i.e. $m_K \to c \in \mathbf{N} \cup \infty$, $\Lambda_c = \lim_{m_K\to c} \Lambda_{m_K}$ and $\Lambda_{m_K} = \left(\lambda_{4,m_K} - \lambda^2_{2,m_K}\right)/\lambda^2_{2,m_K}$. The efficiency of the *RRV* estimator obviously depends on the variance factor Λ. Christensen and Podolskij (2007) illustrate that for $m_K = 10$, which is a reasonable choice for moderately liquid assets, the factor is about 0.7. Its asymptotic value, i.e. for continuously observed prices, the factor is 0.4, such that *RRV* is five times more efficient than *RV*. For $m_K = 1$ the efficiency of *RV* is obtained. Notably, *IQ* can also be estimated based on the range, i.e. via the so-called *realized range-based quarticity* $RRQ^{(m_K,K)} = (1/\lambda_{4,m_K})\sum_{i=1}^{K} s_i^{(m_K)^4}$.

Market microstructure noise corrections of range-based volatility estimators have been proposed by Martens and van Dijk (2007), who focus particularly on the effect of the bid-ask bounce, and by Christensen et al. (2009a). The latter address bias correction under iid noise. However, bias correction is not as straightforward as in the case of using squared returns as the extreme value theory of *RRV* depends on the distribution of the noise. Moreover, *RRV* is more sensitive towards price outliers than squared returns. Nevertheless, the empirical results reported in Christensen et al. (2009a) indicate that bias reduction can be achieved by imposing simple parametric assumptions on the distribution of the noise process and sampling at a 1–2 min frequency.

13.5.2 Quantile-Based Estimation

An approach that is very similar to the range-based estimators is to consider quantiles of the return rather than of the (log) price. We refer to these estimators as the *quantile-based estimators*. This idea dates back at least to David (1970) and Mosteller (1946) and was generalized in Christensen et al. (2009b) by combining multiple quantiles for each of the m_K intraday subintervals yielding the so-called *quantile-based realized variance* (*QRV*) estimator.

The setup is similar to the one used in *RRV*, i.e. the sample is again split into K non-overlapping blocks with m_K returns, where we denote the set of returns contained in block j by $r_{[(j-1)m_K+1:jm_K]}$. For a vector of p return quantiles $\bar{\lambda} = (\lambda_1, \ldots \lambda_p)'$ the QRV estimator is given by:

$$QRV^{(m_K,K,\bar{\lambda})} = \frac{1}{K}\sum_{i=1}^{p}\alpha_i \sum_{j=0}^{K}\frac{q_j^{(m_K,\lambda_i)}}{\nu_1^{(m_K,\lambda_i)}} \quad \text{for } \lambda_i \in (1/2,1) \tag{13.9}$$

with the realized squared symmetric λ_i-quantile

$$\begin{aligned} q_j^{(m_K,\lambda_i)} &= g^2_{\lambda_i m_K}\left(\sqrt{m_K K} r_{[(j-1)m_K+1:jm_K]}\right) \\ &\quad + g^2_{m_K-\lambda_i m_K+1}\left(\sqrt{m_K K} r_{[(j-1)m_K+1:jm_K]}\right), \end{aligned} \tag{13.10}$$

where the function $g_l(x)$ extracts the lth order statistic from a vector x, $\alpha = (\alpha_1, \ldots \alpha_p)'$ is a non-negative vector of quantile weights, summing to unity, and

$$\nu_r^{(m_K,\lambda)} = E\left[\left(\left|U_{(\lambda m_K)}\right|^2 + \left|U_{(m_K-\lambda m_K+1)}\right|^2\right)^r\right]$$

with $U_{(\lambda m_K)}$ denoting the (λm_K)th order statistic of an independent standard normal sample $\{U_i\}_{i=1}^{m_K}$. For m_K fixed and as the number of blocks is increasing, i.e. $m = m_K K \to \infty$, $q_j^{(m_K,\lambda_i)}/\nu_1^{(m_K,\lambda_i)}$ is an estimator of the (scaled) return variance over the jth block. Summing across all blocks naturally yields a consistent estimator of the integrated variance. Christensen et al. (2009b) derive the asymptotic distribution of *QRV*:

$$\frac{\sqrt{m}\left(QRV^{(m_K,K,\bar{\lambda})} - IV\right)}{\sqrt{\theta^{(m_K,\bar{\lambda},\alpha)} IQ}} \xrightarrow{d} \mathcal{N}(0,1),$$

where $\theta^{(m_K,\bar{\lambda},\alpha)} = \alpha' \Theta^{(m_K,\bar{\lambda})}\alpha$ and the i,jth element of the $p \times p$ matrix $\Theta^{(m_K,\bar{\lambda})}$ is given by

$$\Theta_{i,j}^{(m_K,\bar{\lambda})} = m_K \frac{\nu_1^{(m_K,\lambda_i\lambda_j)} - \nu_1^{(m_K,\lambda_i)}\nu_1^{(m_K,\lambda_j)}}{\nu_1^{(m_K,\lambda_i)}\nu_1^{(m_K,\lambda_j)}}$$

with

$$\begin{aligned} \nu_1^{(m_K,\lambda_i\lambda_j)} &= E\left[\left(|U_{(\lambda_i m_K)}|^2 + |U_{(m_K-\lambda_i m_K+1)}|^2\right)\right. \\ &\quad \left.\times\left(\left|U_{(\lambda_j m_K)}\right|^2 + \left|U_{(m_K-\lambda_j m_K+1)}\right|^2\right)\right]. \end{aligned}$$

The fourth power of the realized quantiles can be used to construct a quantile-based estimator of *IQ*.

Christensen et al. (2009b) further propose a subsampled version of the *QRV* estimator that yields improvements in the efficiency of the above estimator by using overlapping subintervals.

The implementation of the estimator involves the choice of several hyperparameters, i.e. the selection of the quantiles λ, the block length m_K, and the assignment of the optimal weights α. For a fixed set of quantiles and a fixed block size, the weights α can be chosen to minimize the asymptotic variance of *QRV* estimators, i.e.minimizing θ yields the optimal weights:

$$\alpha^* = \frac{\Theta^{(m,\bar{\lambda})^{-1}}\iota}{\iota'\Theta^{(m,\bar{\lambda})^{-1}}\iota},$$

where ι is a $(p \times 1)$ vector of ones. Comparing the efficiency of the estimator, Christensen et al. (2009b) conclude that the gains from optimizing α for finite samples, instead of using the asymptotic optimal values, are only minor.

For the quantile selection, Christensen et al. (2009b) find that the 90–95% quantiles are most informative. The quantiles around the median are uninformative and those around the extremes are too erratic and less robust to potential jumps in the price process or to outliers. Nevertheless, quantiles outside the most informative region may be used to exploit the covariances structure of the order statistics for $p > 1$. Smaller block sizes deliver slightly more efficient estimators, as they achieve better locality of volatility. Also, the subsampled version is shown to be slightly more efficient than the blocked version for multiple quantiles. Finally, the efficiency constant θ can be reduced to around 2.5 for one quantile and is close to 2 for multiple quantiles, achieving the efficiency constant of *RV*.

Christensen et al. (2009b) propose a modification of the *QRV* estimator that makes it robust to iid noise. Based on a pre-averaging technique similar to Podolskij and Vetter (2009), the robust estimator is obtained by applying the *QRV* methodology to a weighted average of the observed returns. In particular, define the averaged data by:

$$\bar{y}_j = \sum_{i=1}^{L-1} h\left(\frac{i}{L}\right) r_{j+i}^{(m)}$$

with $L = c\sqrt{m} + o(m^{1/4})$ for some constant c and weight function h on $[0, 1]$. Further conditions of h are given in Christensen et al. (2009b), who use in their simulation and application the weight function $h(x) = \min(x, 1 - x)$. The *QRV* estimator is then given by:

$$QRV_{\bar{y}}^{(L,m_K,K,\bar{\lambda})} = \frac{1}{c\psi_2(m - m_K(L-1)+1)} \sum_{i=1}^{p} \alpha_i \sum_{j=0}^{m_K(K-L+1)} \frac{q_{\bar{y};j}^{(m_K,\lambda_i)}}{\nu_1^{(m_K,\lambda_i)}}$$

with

$$q_{\bar{y};j}^{(m_K,\lambda_i)} = g_{\lambda_i m_K}^2 \left(m^{1/4}\bar{y}_{[j:j+m_K(L-1)]}\right) + g_{m_K-\lambda_i m_K+1}^2 \left(m^{1/4}\bar{y}_{[j:j+m_K(L-1)]}\right).$$

The problem of $QRV_{\bar{y}}^{(L,m_K,K,\bar{\lambda})}$ is that it is biased. Incorporating a bias-correction finally yields the iid noise-robust estimator:

$$QRV_{iid}^{(L,m_K,K,\bar{\lambda})} = QRV_{\bar{y}}^{(L,m_K,K,\bar{\lambda})} - \frac{\psi_1}{c^2\psi_2}\omega^2, \tag{13.11}$$

where ψ_1 and ψ_2 can be computed by

$$\psi_2 = L\sum_{j=1}^{L}\left(h\left(\frac{j}{L}\right) - h\left(\frac{j-1}{L}\right)\right)^2$$

and

$$\psi_1 = \frac{1}{L}\sum_{j=1}^{L-1} h^2\left(\frac{j}{L}\right).$$

Under some further mild assumptions, Christensen et al. (2009b) show that this estimator converges at rate $m^{-1/4}$ to the *IV*. However, in contrast to the other volatility estimators its asymptotic variance has no explicit expression in terms of *IQ*. Nevertheless, it can be estimated based on the estimates of the q_i, ψ_2 and ν_1 terms. For $h(x) = \min(x, 1-x)$ and the constant volatility setting the estimator achieves a lower bound of $8.5\sigma^3\omega$. This is close to the theoretical bound of the variance of the realized kernel approach discussed in Sect. 13.4.2, which is $8\sigma^3\omega$. The behavior of the noise robust estimator will of course depend on the choice of L, which trades-off between the noise reduction and the efficiency loss due to pre-averaging. A simulation study suggests that a conservative choice, e.g. a larger value of L, such as $L = 20$, may be preferable. In applications the estimated signal-to-noise ratio can be used to determine L based on the mean-square error (MSE) criterion.

13.5.3 Duration-Based Estimation

While the return- and range-based volatility estimators make use of a functional of the price path between fixed points in time, the duration-based approach focuses on the time it takes the price process to travel between fixed price levels. Such an approach was first investigated by Cho and Frees (1988) for the constant volatility case. Recently, Andersen et al. (2009) provide a more comprehensive treatment of this concept in the case of constant volatility and for stochastic volatility evolving without drift. They consider three different *passage times*, i.e. three different ways to measure the time a Brownian motion needs to travel a given distance r:

$$\tau^{(r)} = \begin{cases} \inf\{t : t > 0 \,\&\, |W_t| > r\} & \text{(first exit time)} \\ \inf\{t : (\max_{0<s\le t} W_s - \min_{0<s^*\le t} W_{s^*}) > r\} & \text{(first range time)} \\ \inf\{t : t > 0 \,\&\, W_t = r\} & \text{(first hitting time)} \end{cases}$$

In the constant volatility case, the moments of these passage times are available in closed-form:

$$E\left[\tau^{(r)}\right] = \begin{cases} \frac{r^2}{\sigma^2} & \text{(first exit time)} \\ \frac{1}{2}\frac{r^2}{\sigma^2} & \text{(first range time)} \\ \infty & \text{(first hitting time)} \end{cases} \tag{13.12}$$

Interestingly, comparing these moments to the expected value of a squared Brownian increment over the interval τ, which is $\sigma^2\tau$, illustrates the duality between RV and the range-based volatility approaches and the duration-based one.

The moment conditions (13.12) suggest to estimate σ^2 via the method of moments using either an observed sample of first exit times or of first range times with fixed r. However, as the expected passage times are inversely proportional to the instantaneous variance, these estimators will suffer from severe small sample biases induced by Jensen's inequality. For this reason, Andersen et al. (2009) propose small sample unbiased estimators based on the reciprocal passage times:

$$E\left[\frac{r^2}{\tau^{(r)}}\right] = \mu_1\sigma^2 = \begin{cases} 2C\sigma^2 & \text{(first exit time)} \\ (4\log 2)\sigma^2 & \text{(first range time)} \\ \sigma^2 & \text{(first hitting time)} \end{cases},$$

where $C \approx 0.916$ is the Catalan constant. Interestingly, moment based estimators for the reciprocal hitting time are now also feasible.

The concept also allows to define a *local* volatility estimator for a single passage time at (intraday) time point i:

$$\left(\hat{\sigma}_i^{(r)}\right)^2 = \frac{1}{\mu_1}\frac{r^2}{\tau_i^{(r)}},$$

such that IV can also be estimated in the case of stochastic volatility by applying the Riemann sum. The resulting *duration-based realized variance* estimator is given by:

$$DRV^{(m,r)} = \sum_{i=1}^{m}\left(\hat{\sigma}_i^{(r)}\right)^2\delta_i \tag{13.13}$$

with δ_i denoting the times between the intraday observations.

Based on the time-reversibility of the Brownian motion, the local volatility estimates can be constructed using either the *previous passage time* or the *next passage time*, i.e. the time is determined by the path of the Brownian motion either prior to or after the time point i, respectively. In practice, market closures will thus induce the problem of censoring. In particular, the expected next passage time is affected by the time left until the market closes, *(right censoring)*, while the expected previous passage time is limited by the time the market opened, *(left censoring)*. Andersen et al. (2009) show that a one-directional approach may be

Table 13.1 Efficiency constant of *DRV* for different types of passage time

	Bi-directional	Uni-directional
First exit time	0.3841	0.7681
First range time	0.2037	0.4073
First hitting time	1.0000	2.0000

preferred, although combining both schemes, so-called *bi-directional* local volatility estimation, has the potential of reducing the variance of the duration-based estimator by a factor of two, see also Table 13.1. More precisely, to account for censoring effects, they suggest to construct the *DRV* estimator based on the next passage time scheme during the first half of the day and to use the previous passage time scheme over the second half of the day. The suggestion is motivated by their simulation results for exit and range passage times showing that left and right censoring can be ignored, if the difference in time to the market opening and closing is 2–3 times longer than the expected passage times.

The duration-based approach can also be used to construct estimators of the integrated quarticity:

$$DRQ^{(m,r)} = \sum_{i=1}^{m} \left(\hat{\sigma}^{(r)}\right)^4 \delta_i ,$$

which allows the construction of confidence intervals using the asymptotic result for *DRV*:

$$\frac{\sqrt{m}\left(DRV^{(m,r)} - IV\right)}{\sqrt{\nu IQ}} \xrightarrow{d} \mathcal{N}(0,1) ,$$

where ν is a constant that is specific to the type of passage time used in the estimation and that is independent of the choice of r. Table 13.1 presents the respective values of this efficiency constant.

The asymptotic efficiency is much higher compared to the return-based estimators, especially if the dataset allows the usage of bi-directional passage times through non-interrupted trading, suggesting the use of trade data from FOREX or GLOBEX. However, similarly to the other estimators, the *DRV* not only suffers from the problem that the price process is observed only at m discrete time points, but also that the number of observed price changes is even less, see Sect. 13.7. Andersen et al. (2009) therefore suggest to sample sparsely in order to avoid this potentially more pronounced discreetness effect. Moreover, similarly to the range-based estimator the *DRV* based on first range time and on first exit time may be biased, as the observed times may not coincide with the true ones.

A formal noise-robust *DRV* estimator has not been developed so far, however, Andersen et al. (2009) investigate the impact of market microstructure noise on the *DRV* estimator within a simulation study with independent and serial dependent noise assumptions. The results indicate that the estimator is nevertheless sufficiently

robust to independent noise with moderate levels of noise-to-signal ratio even in the case of first range and first exit times. Also, as may be naturally expected, higher threshold values r make the estimator more robust to noise.seems to be very robust to the higher persistent levels typically encountered for quote data as argued in Andersen et al. (2009).

13.6 Theoretical Comparison of Volatility Estimators and Price Jumps

So far we have discussed and presented the most popular and the most recent approaches to estimate *IV* based on various characteristics of high-frequency financial data. In the following we provide a brief summary of the main large sample properties of these estimators in order to facilitate their comparison. An empirical evaluation and illustration of the estimators is given in Sect. 13.8.

Table 13.2 summarizes the estimators, which are grouped according to the underlying assumption on the market microstructure noise under which they achieve consistency. We further report the asymptotic variances (based on rounded and optimally chosen parameter values) and the convergence rates of the various estimators. Note that due to the unavailability of a closed-form expression of the asymptotic variance of QRV_{iid} we report here only its lower bound in the setting of constant volatility. The reported asymptotic variance of KRV_{NFT} is based on the Parzen kernel. Moreover, the complexity and the performance of the estimators often depend on the choice of hyperparameters as is indicated in the table. The exact impact of those parameters and their determination have been discussed in the previous sections. Note that with the exception of the non-flat-top realized kernel all the estimators are unbiased in large samples and we, thus do not comment on this property in the table.

The table also reports the robustness of the various estimators to the presence of jumps. So far we have assumed that the log price follows a pure diffusion process. However, recent empirical evidence suggests that jumps may have a non-trivial contribution to the overall daily price variation, see e.g. Andersen et al. (2007), Eraker et al. (2003) and Huang and Tauchen (2005). Suppose the log price follows in fact a continuous-time jump diffusion process:

$$p_t = \int_0^t \mu(s)\mathrm{d}s + \int_0^t \sigma(s)\mathrm{d}W(s) + \sum_{j=1}^{N(t)} \kappa(s_j),$$

where the $N(t)$ process counts the number of jumps occurring with possibly time-varying intensity $\lambda(t)$ and jump size $\kappa(s_j)$. Given the presence of jumps in the price process, the question arises, whether the proposed approaches still deliver estimators of the integrated variance, i.e. the object of interest in many financial applications?

From the theory of quadratic variation it follows that the basic *RV* estimator converges uniformly in probability to the quadratic variation as the sampling

frequency of the underlying returns approaches infinity:

$$RV \xrightarrow{p} IV + \sum_{j=N(t-1)+1}^{N(t)} \kappa^2(s_j).$$

In other words, the realized variance provides an ex-post measure of the true *total* price variation, i.e. including the discontinuous jump part.

In order to distinguish the continuous variation from the jump component, Barndorff-Nielsen and Shephard (2004) first proposed the so-called *Bipower variation* measure, defined by:

$$BPV^{(m)} = \frac{\pi}{2} \sum_{j=2}^{m} |r_j||r_{j-1}|,$$

which becomes immune to jumps and consistently estimates the integrated variance as $m \to \infty$. A central limit theory for Bipower variation has just recently been derived in Vetter (2010). He also provides a brief review of alternative jump-robust estimators of *IV* including multipower variations, see Barndorff-Nielsen et al. (2006), and a threshold-based realized variance estimator, see Mancini (2009).

Table 13.2 shows, that only a few of the previously discussed approaches deliver consistent estimators of *IV* in the presence of (finite active) jumps. In the quantile-based estimation, the jump robustness is due to the exclusion of the extreme quantiles in the construction of the estimator. Similarly, the *DRV* estimator can be made robust by the choice of the price threshold r, i.e. limiting the impact of

Table 13.2 Asymptotic properties of the IV estimators

Estimator	Equation	Asymptotic variance	Convergence rate	Jump robust	Parameters
No microstructure noise					
RV	(13.2)	$2IQ$	$m^{1/2}$	No	m
RRV	(13.8)	$0.4IQ$	$K^{1/2}$	No	m_K, K
QRV	(13.9)	$2.3IQ$	$m^{1/2}$	Yes	$m_K, K, \bar{\lambda}$
DRV first exit	(13.13)	$0.77IQ$	$m^{1/2}$	Yes[a]	m, r
iid noise					
TTSRV	(13.4)	$1.33\frac{K}{m^{2/3}}IQ + 8\frac{m^{4/3}}{K^2}\omega^2$	$m^{1/6}$	No	$m, m_1, \ldots, m_K, K$
QRV_{iid}	(13.11)	$8.5\sigma^3\omega$	$m^{1/4}$	Yes	$m_K, K, L, \bar{\lambda}$
Time-dependent noise					
$TTSRV_{adj}$	(13.5)	$1.33\frac{K}{m^{2/3}}IQ + \frac{m^{4/3}}{K^2}\xi^2$	$m^{1/6}$	No	m, K, J
Time-dependent and endogenous noise					
KRV_{NFT}	(13.6)+(13.7)	$3.78IQ$	$m^{1/5}$	No	m, k, H

[a]Explanation is given in the text

jumps that exceed this threshold. The asterisk in Table 13.2 indicates that the jump robustness of this estimator has been shown only within a Monte Carlo simulation for a modified version of the estimator that utilizes the threshold corresponding to the observed log price at the tick time prior to the crossing of the target threshold. A jump robust range-based estimator is proposed by Klößner (2009).

13.7 High-Frequency Financial Data: Characteristics and Computational Aspects

In the following we briefly review some of the main characteristics of high-frequency financial data and of the existing sampling schemes. Moreover, as the volume of the high-frequency dataset exceeds the one usually encountered in financial statistics or econometrics, we discuss also the computational aspects concerning data storage and retrieving, which will be useful not only for the reader interested in implementing volatility estimators, but also to those planning to work with high-frequency financial data in general.

13.7.1 Price Series and Sampling Schemes

Electronic trading systems have lead to the availability of detailed price and trade information at the ultrahigh frequency. In particular, information on the arrival and volume of the sell and buy orders is stored along with the ask and bid quotes. A trade takes place if buy and sell orders could be matched and the corresponding price of this transaction, i.e. the transaction price, is recorded. As the underlying type of trading mechanism differs across exchanges, we refer the interested reader to Hasbrouck (2007) and Gouriéroux and Jasiak (2001) for a more detailed discussion on order books and existing types of markets.

An important feature of an exchange market is that prices at which one can send buy (bid) and sell (ask) quotations and at which transactions take place must be multiples of a predetermined number, called *tick size*. As a consequence, markets with a tick size relatively large in comparison to the price level of the asset, *large tick markets*, often exhibit a *spread*, i.e. the difference between the price of the highest available bid and the lowest available ask quote, that equals most of the time exactly one tick. The S&P500 future is an example for such a market, see Hasbrouck (2007). Obviously, such price discreteness or round-off errors represent one source of market microstructure noise that will affect the performance of the *IV* estimators, especially of *DRV*.

Given the availability of transaction, bid and ask prices, the question arises on which of these price series should be used in the construction of the estimators. Financial theory of course suggests to use the price at which the asset trades. However, assuming a random flow of alternating buying and selling market orders,

the trading mechanism and the discrete prices will cause transaction prices to randomly fluctuate between the best bid and ask price. This effect is called *bid-ask bounce* and was first described in Roll (1984). It induces a strong negative autocorrelation in the returns and, thus, violates the assumption of a semimartingale for the price process.

This has lead to the consideration of the *mid-quotes*, i.e. the average of the best bid and ask price. However, the mid-quotes are also not immune to microstructure effects. In fact, they suffer from the so-called *price staleness*. They change rather rarely, and are subject to non-synchronous adjustments of the bid and ask quotes. Alternatively, the bid and ask prices can be used, which in large tick markets contain a similar amount of information as the mid-quotes, but do not suffer from non-synchronous adjustment effects.

Apart from deciding upon the price series used in the empirical implementation of the *IV* estimators, one also has to choose the scheme at which prices are sampled. The literature basically distinguishes between four types of sampling schemes, see Oomen (2006): calendar time sampling, transaction time sampling, business time sampling and tick time sampling. The most obvious one is *calendar time sampling*, CTS, which samples at equal intervals in physical time. As high-frequency observations are irregularly spaced in physical time, an artificial construction of CTS from the full record of prices is necessary. A natural approach is given by the *previous tick method*, see Wasserfallen and Zimmermann (1985), which uses the last record observed prior to the sampling point. The *linear interpolation* method instead interpolates between the previous and the next observed price. At ultra high-frequencies this implies, however, that $RV \to 0$ as $m \to \infty$, see Hansen and Lunde (2006).

Alternatively, one can sample whenever a transactions takes place, i.e. the so-called *transaction time sampling*, TTS, or whenever prices change, so-called *tick time sampling*, TkTS. The latter can further be distinguish according to the type of price series, yielding tick-time sampling for transactions TkTS(T), mid-quotes TkTS(MQ), and bid and ask prices, TkTS(B) and TkTS(A), respectively. A generalization of TTS is *event time sampling*, ETS, where sampling takes place at all market events including transactions and quotations. Thus, TTS, TkTS and ETS are only based on observed prices and time points. This is not the case in *business time sampling*, BTS, which samples data such that *IV* of the intraday intervals are all equal, i.e. $IV_i = \frac{IV}{m}$.

BTS is infeasible as it depends on *IV*. However, in practice it can be approximated by prior estimates of *IV* or by standard non-parametric smoothing methods using the transaction times, see Oomen (2006). The *φ-sampling* scheme introduced by Dacorogna et al. (1993) is similar to BTS, but also removes seasonalities in the volatility across days, while the BTS just operates within the day. Empirical results of Andersen and Bollerslev (1997) and Curci and Corsi (2006) suggest that BTS can be well approximated by TTS. In the setting of Oomen (2006) random transaction times are generated by a quantity related to IV such that TTS can be directly interpreted as a feasible variant of BTS. Hansen and Lunde (2006) show that BTS,

Table 13.3 Overview of the number of observations (ticks) in different sampling schemes (01/01/2008–03/31/2008)

	S&P500 E-mini (8:30–15:15)				MSFT (9:30–16:00)			
	Total ticks	Ticks/day	Δ s	Ticks/s	Total ticks	Ticks/day	Δ s	Ticks/s
CTS (1 s)	1,496,576	23,755	1.00	1.00	1,427,277	22,655	1.05	0.95
TTS	9,466,209	150,257	0.16	6.33	8,452,679	134,170	0.18	5.65
ETS	44,646,176	708,669	0.03	29.83	–	–	–	–
QTS	–	–	–	–	22,342,994	354,651	0.07	14.93
TkTS (T)	2,772,594	44,009	0.54	1.85	1,191,310	18,910	1.26	0.80
TkTS (MQ)	1,935,415	30,721	0.77	1.29	1,893,741	30,059	0.79	1.27
TkTS (B)	968,666	15,376	1.54	0.65	831,659	13,201	1.80	0.56

by construction, minimizes *IQ*. Thus, using BTS and TTS rather than CTS may reduce the variance of RV.

Moreover, the results in Griffin and Oomen (2010), who introduce a model for transaction time patterns for analyzing the effects of TkTS and TTS, suggest that TkTS is equivalent to TTS for high levels of noise and is superior for low levels. However, once a first-order bias correction is applied, TTS is preferable.

Table 13.3 illustrates the impact of the various sampling schemes on the number of ticks available for the construction of the *IV* estimators. The numbers are based on the two datasets used in our empirical illustration, see Sect. 13.8. Generally, the number of ticks as well as the time scales are quite different across the sampling schemes. For example, for the S&P500 E-mini the one minute CTS corresponds to sampling about every 380 transactions in TTS and 1,750 events in ETS. For MSFT we obtain 340 in TTS and 900 in QTS (see Sect. 13.8). For both assets the markets have become more active, i.e. there are more quotes and trades in 2008 than in 2006. As the tick number for the Bid and Ask are similar we just report here TkTS(B).

13.7.2 Computational Aspects

A unique feature of high-frequency datasets is the vast mounds of data. In comparison to datasets commonly used in financial econometrics, e.g. daily financial data, high-frequency data requires a different approach to data storage and retrieval. The full Trade and Quote, TAQ, dataset contains for example around 10 million records per day in November 2004 and around 150 million records per day in November 2008. Obviously, the mere storage, fast retrieval and processing of this amount of data requires advanced information technology, which is discussed in this paragraph. In addition to the established row-oriented database systems we also discuss column-oriented systems and perform a comparison in terms of storage requirements and query execution speed. All computations are performed on the Microsoft Windows.Net framework but can also be replicated in econometric/statistics packages, e.g. Matlab or R, given a suitable interface to the database.

Structured data is usually stored in database management systems, i.e. software packages that offer convenient, flexible and fast read and write access to it. There is a high number of mature general purpose databases, including Microsoft SQL Server, Oracle Database, IBM DB2 and others. They are all *row-oriented*, i.e. they store entire records one after another, which may be highly disadvantageous for analytical data. Only recently have *column-oriented* databases attracted more attention, see Abadi et al. (2008). column-oriented databases store all the attributes from different records belonging to the same column contiguously and densely packed, which allows for more efficient read access, when few columns but many rows are required. column-oriented storage can be traced back to the 1970s, when transposed files and vertical partitioning clustering techniques were first studied. The interest in these techniques accelerated during the 2000s, partially because of the exponentially growing data volumes, which have become increasingly hard to handle by general purpose row-oriented databases. Another factor, which necessitates a rethinking of the design of database systems, is the increasing discrepancy between processor and physical memory speeds (Boncz 2002). While over the last decades transistor density in chips, affecting processor speed and storage density, has closely followed Moore's law – postulating a doubling of transistor chip density every 18 months – external and internal memory latency have been lagging, creating a growing bottleneck. Modern column-oriented databases are designed considering this bottleneck. Each column is stored separately, typically using large disk read units to amortize head seeks. Columnar values are stored densely, in sequence, which especially on sparse data types, can deliver astonishing levels of compression (see e.g. Stonebraker et al. 2005). Consider the storage of a bid price column, for instance. There is approx. one price change per 50 quote size changes and these price changes are mostly within a narrow band. Obviously the information entropy of the column is quite low. Furthermore, to partially avoid the processing cost involved in the decompression, column-oriented databases usually have query executors which work directly on the compressed data Abadi et al. (2006). The benefits of compression are not entirely limited to column-oriented stores but are a lot bigger, considering that the information entropy of the values within a column is almost always lower than the information entropy of the values within a record.

The biggest disadvantages of column-oriented storage manifest themselves during tuple (record) reconstruction and write operations. Write operations are problematic as inserted records have to be broken into columns and stored separately and as densely packed data makes moving records almost impossible. Some of the techniques used to mitigate the write issues are in-memory buffering and partition-merging. The problem with tuple reconstruction is again that the data for a single row is scattered in different locations on the disk. Most database interface standards (e.g. ODBC) access the results of a query on a row basis, not per columns. Thus, at some point of the query plan of a column-oriented database, the data from multiple columns must be combined in records. Abadi et al. (2008) consider several techniques, which can be used to minimize this reconstruction overhead.

A list of the currently available commercial column-oriented databases includes Sybase IQ, Vertica, VectorWise, InfoBright, Exasol, ParAccel, SAP BI Accelerator,

Kickfire and others. Not all of them are general purpose databases, e.g. InfoBright is actually an MySQL storage engine and Kickfire is offered as an hardware appliance. The most mature academic system is MonetDB/X100, developed at Centrum Wiskund & Informaticas.

The MySQL/InfoBright solution can be referred to as a hybrid system, as MySQL can simultaneously handle a number of different engines, including both column- and row-oriented stores, which can be selected on a per table basis. The usage of the highly mature MySQL database platform and the fact that InfoBright is freely available in an open-source edition (InfoBright ICE), make it a good candidate for academic comparison.

In the following we compare the retrieval speed and the compression levels of the row-oriented Microsoft SQL Server, which is a mature database system, introduced in 1989 and very well integrated into the whole palette of development tools from Microsoft, and the column-oriented database MySQL/InfoBright. The dataset used for the test comprises all transactions and quote updates in the first quarter of 2008, a total of 97.9 million records.

The sizes of the sample in the form of: raw data, flat file, uncompressed Microsoft SQL database, compressed Microsoft SQL database and compressed InfoBright database are given in Fig. 13.1. Indeed, the compression rate of InfoBright is astonishing – 1–20 compared to the raw data size. The implications are huge – a raw dataset of e.g. 10 Terabyte (TB) can be stored on an off-the-shelf hard drive with 500 GB capacity. On the contrary, Microsoft SQL manages to achieve only compression ratios of 1–2, compared to the raw data size, and 1–3, compared to the uncompressed Microsoft SQL database. The benefits of these rates may become questionable in the light of the processor overhead caused by decompression.

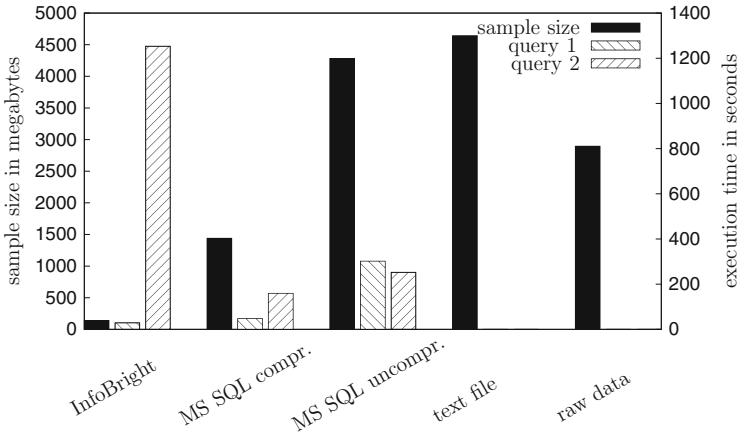

Fig. 13.1 Sample size in megabytes (*left scale*) in the form of InfoBright database, Microsoft SQL compressed database, Microsoft SQL uncompressed database, comma-separated text file and theoretical size as an in-memory structure (raw data); and query execution speed in seconds (*right scale*) for the InfoBright, Microsoft SQL compressed and Microsoft SQL uncompressed databases

The performance of the two database systems will be compared with the help of two queries, the first of which will test the speed of retrieval of aggregated and filtered information, performing an *in-database* full table scan:

```
SELECT SecurityID, DateID, MIN(Timestamp),
MAX(Timestamp), SUM(Count), SUM(Size), MIN(Price),
MAX(Price) FROM tblTickES WHERE FieldID = 3 GROUP
BY SecurityID, DateID
```

The second query will test the sequential retrieval of all the information in the table from an *external* environment:

```
SELECT SecurityID, DateID, Timestamp, FieldID, Price,
Size, Count FROM tblTickES
```

Both query types are important in analytical work (e.g. in econometrics) but the performance of the second is especially relevant, as it is used on a more regular basis and requires the transfer of huge amounts of data between applications, a process which can quickly become a bottleneck for the whole system.

It is important to note that these tests were not performed under ideal conditions and are in no way representative for the general and even optimal performance of the two database systems. The tests are designed to assess the performance which can be expected from the systems in a normal working environment by an analytical worker, who is not able or willing to spend considerable amounts of time on learning and optimizing the systems. The results of the tests on the second run are reported in Fig. 13.1. The results of the first run are ignored because they can unfairly penalize systems which make use of cache and memory to optimize their performance. The results of runs after the second one, on the other side, can be too hardware specific, since some systems could manage to cache large amount of data from hard drive media in the memory.

The speed of retrieving all data from InfoBright is low. The number in Fig. 13.1 reports the result for the general ODBC driver for MySQL. Changes of diverse settings in the ODBC driver did not improve the situation. The in-database query speed of InfoBright is satisfactory. Overall, the compressed Microsoft SQL variant offers a promising improvement over the uncompressed one – a factor of 1–6 for the in-database query and slightly less than 1–2 for the external query.

To conclude, column-oriented database systems provide a comfortable way to achieve a high comparison of the data and fast in-database queries. On the other side a sequential retrieval of all records is significantly slower than for row-oriented database systems. Thus, the preferred choice of the database system may depend whether good compression or fast sequential retrieval of all records is important.

13.8 Empirical Illustration

Our empirical application aims at illustrating the impact of market microstructure noise on the estimation of IV in finite samples and on an empirical comparison of the various estimators. To this end, we consider two high-frequency datasets over the first quarters of 2008: data on the highly liquid futures S&P500 E-mini and on an individual stock, i.e. of Microsoft, MSFT. The reason for this choice is, that the data sources and the type of asset are quite different allowing for a more detailed analysis of the market microstructure noise and the performance of the estimators.

The S&P500 E-mini, is traded on the CME Globex electronic trading platform and the dataset consists of all transaction and quotation updates in correct order and with time-stamps given in milliseconds. The quality of the data is very high and no filtering or cleaning is required, except for a trivial removal of any non-positive prices or volume. Note that in our application we roll-over between the most liquid contracts.

Such highly accurate information is not available for MSFT, which is obtained from the (monthly) TAQ dataset, disseminated by the NYSE for all listed stocks. The dataset includes quotation updates and transactions provided in separate files and the time-stamps are available only up to the precision of a second. This requires a more involved data filtering and cleaning. As the TAQ is probably the most popular high-frequency dataset, we give here a few more details on the data manipulations conducted for our analysis. In particular, we focus on the TAQ data coming from the NASDAQ, such that we filter all records with exchange identifiers being different from T, D or Q, as specified in the TAQ 3 User's Guide (2004–2008). For transactions we have additionally removed records with a CORR attribute different from 0 or 1. The resulting data contains numerous outliers, such as prices equal to 0.01$ or 2,000$ right next to regular prices varying around the usual trading price range of MSFT, i.e. around 30$. Such outliers have been removed by first dismissing records with non-positive price or size and by discarding records with a price that deviates from the last one by more than 10%. More advanced methods involve filtering based on rolling windows and a deviation threshold adapted to the current volatility of the price, see Brownlees and Gallo (2006).

One of the major problems of the TAQ dataset, however, is the separate distribution of transaction and quote data and the lack of millisecond precision in the time-stamps, such that the exact order of the generation of transaction prices and quotes over the trading day cannot be deduced. An approach to match transactions at least to the corresponding bid and ask quotes has been proposed in Lee and Ready (1991). For volatility estimation such synchronicity is only required for sampling schemes involving price type pairs. For MSFT we have, thus, limited TTS to transaction prices and introduce a modification of ETS, which only regards quote updates as events, called *quote-time sampling* (QTS). This sampling scheme can of course be applied to mid-quotes, bid and ask prices avoiding any mixing of transaction and quote series.

13.8.1 Type and Magnitude of the Noise Process

A common tool to visualize the impact of market microstructure noise on the high-frequency based volatility estimators are the so-called volatility signature plots made popular in Andersen et al. (2000). Depicted are usually the average estimates of daily volatility as a function of the sampling frequency, where the average is taken across multiple days, i.e. in our case all trading days in the first quarter of 2008. Figure 13.2 shows the volatility signature plots for RV based on different sampling schemes and different prices.

Overall, it seems that CTS is most strongly affected by market microstructure noise, compared to the alternative sampling schemes. This is important, as CTS is probably the most commonly applied sampling method. Moreover, under the assumption of a pure jump diffusion price process, Oomen (2006) shows theoretically that TTS is superior to CTS, if the in a MSE sense optimal sampling frequency is used.

Interestingly, the biases observed in the RV estimates for both of our datasets are all positive, irrespective of the sampling scheme and the employed price series. Moreover, we find across all sampling schemes that transaction prices produce the most severe bias in the case of the S&P500 E-mini (note that all quotes series yield identical RVs in both CTS and TTS and we thus only display the estimates based on the mid-quotes), but are preferable for MSFT. Using the same stock but a different sample period, Hansen and Lunde (2006) instead observe a superiority of quotation data in terms of bias reduction and a negative bias if quote data is used. The latter may be induced by the non-synchronous quote revisions or price staleness. Recall that another source of a negative bias may be given by the dependence between the efficient price and the noise. Obviously, the potential presence of these different types of market microstructure effects make it difficult to draw general statements on the expected sign and size of the biases in the RV estimator and the preferable sampling method/price series. Negative and positive biases may be present at the same time, leading to overall small biases or to non-monotone patterns like the one observed for the S&P500 E-mini under ETS. Volatility signature plots based on estimators that are robust to particular types of microstructure noise, allow to shed more light on the noise effects. Using a kernel-based approach, Hansen and Lunde (2006) for example find, that the iid robust RV based on transaction prices also exhibits a negative bias and that this may be due to endogenous noise.

Instead of using pure visual inspections to judge the presence and potential type of market microstructure noise, Awartani et al. (2009) propose statistical tests on the no noise assumption and on noise with constant variance. The no market microstructure noise test builds on the idea, that RV sampled at two different frequencies, e.g. very frequently and sparsely, should both converge in probability to IV. The test therefore evaluates, whether the difference of both estimators is zero. Asymptotically the test statistic is normally distributed. The implementation of the

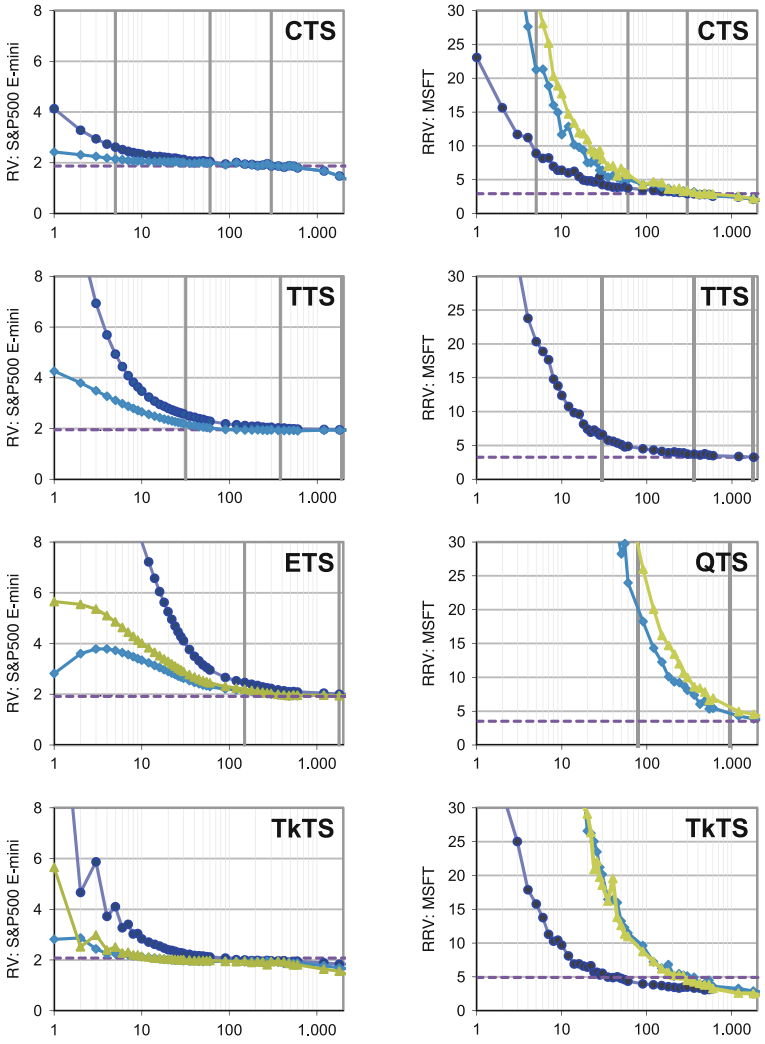

Fig. 13.2 Volatility signature plots for RV of S&P500 E-mini (*left*) and MSFT (*right*), first quarter 2008, based on different sampling schemes and different price series: transaction prices (*circles*), mid-quotes (*rhombuses*) and bid/ask prices (*triangles*). The *bold vertical lines* represent the frequency equal, on average, to 5 s, 1 and 5 min in the respective sampling scheme, the *horizontal line* refers to the average RV estimate based on a 30 min frequency

test of course depends on the choice of both sampling frequencies. As an alternative, Awartani et al. (2009) suggest to exploit the autocovariance structure of the intraday returns. Focusing on the first lag, the scaled autocovariance estimator over e.g. n days can be expressed by

$$\bar{m}^{3/2}\widehat{cov}(r_i^{(m)}, r_{i-1}^{(m)}) = \sqrt{\bar{m}}\left(\sum_{i=2}^{\bar{m}} r_i^{(m)} r_{i-1}^{(m)} + \sum_{i=2}^{\bar{m}} \epsilon_{i/m}\epsilon_{(i-1)/m} + \sum_{i=2}^{\bar{m}} r_i^{(m)}\epsilon_{(i-1)/m} + \sum_{i=2}^{\bar{m}} \epsilon_{i/m} r_{i-1}^{(m)}\right),$$

where $\bar{m} = nm$. Under the null of no noise the last three terms converge to zero almost surely. The first term therefore drives the asymptotic distribution of the test statistic which is given by:

$$\frac{\sqrt{\bar{m}}\sum_{i=2}^{\bar{m}} r_i^{(m)} r_{i-1}^{(m)}}{\sqrt{IQ}} \xrightarrow{d} \mathcal{N}(0,1).$$

After some rearrangement and for large $\bar{m}$, the feasible test statistic can also be computed in terms of the sample autocorrelation coefficient of the intraday returns $\hat{\rho}_1(r_i^{(m)}) = \sum_{i=2}^{\bar{m}} r_i^{(m)} r_{i-1}^{(m)} / \sum_{i=2}^{\bar{m}} (r_i^{(m)})^2$:

$$z_{AC,1} = \frac{\hat{\rho}_1(r_i^{(m)})}{\sqrt{\frac{1}{3}\sum_{i=2}^{\bar{m}} (r_i^{(m)})^4 / \sum_{i=2}^{\bar{m}} (r_i^{(m)})^2}} \xrightarrow{d} \mathcal{N}(0,1).$$

Figure 13.3 presents the test statistic and corresponding confidence intervals as a function of the sampling frequency over the first quarter of 2008. The results indicate that the noise "kicks in" at frequencies exceeding approximately 1 and 3 min for the S&P500 E-mini and MSFT data, respectively. Moreover, if quote data is used in the case of MSFT, then the noise robust sampling frequency should be lower than approx. every 5 min.

Most of the noise robust estimators have been derived under the assumption of iid noise, implying also that the noise has a constant noise variance, irrespective of the sampling frequency. Awartani et al. (2009) therefore propose a test for the null of constant noise variance. To this end it is instructive to first consider feasible estimators of the noise variance. Based on the bias of RV in the presence of iid noise, see (13.3), the noise variance can be estimated by $\tilde{\omega}^2 = RV/2m$ using sparse sampling. However, Hansen and Lunde (2006) show that this estimator will overestimate the true noise variance whenever $IV/2m$ is negligible. They therefore suggest the following estimator:

$$\hat{\omega}^2 = \frac{RV^{(m_K)} - RV^{(m_J)}}{2(m_K - m_J)}, \tag{13.14}$$

where m_J denotes a lower sampling frequency, such that $RV^{(m_J)}$ is an unbiased estimator of IV. However, both variance estimators may be inadequate if the iid noise assumption is not appropriate, which may be the case at very high frequencies.

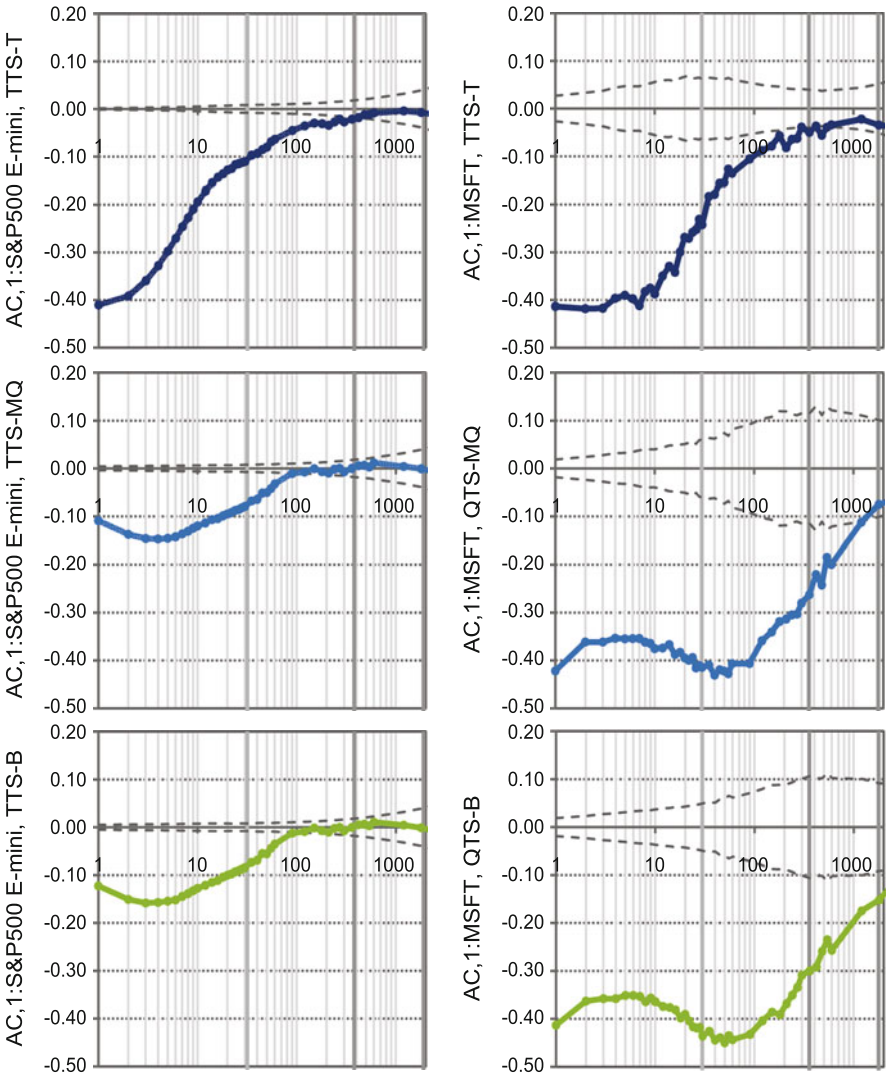

Fig. 13.3 Tests on no noise. Depicted are the $z_{AC,1}$ statistics and corresponding confidence intervals (*dashed*) based on different sampling frequencies for the S&P500 E-mini (*left*) and MSFT (*right*) using TTS/QTS and transaction (*top*), mid-quote (*middle row*) and bid/ask quote (*bottom*) price series. The *bold vertical lines* give the frequency equal, on average, to 5 s, 1 and 5 min

The constant noise variance test of Awartani et al. (2009) considers the difference between two noise variances estimated at different sampling frequencies:

$$z_{IND} = \sqrt{\bar{m}_J} \frac{\frac{RV^{(\bar{m}_K)} - RV^{(\bar{m}_L)}}{2\bar{m}_K} - \frac{RV^{(\bar{m}_J)} - RV^{(\bar{m}_L)}}{2\bar{m}_J}}{\sqrt{3\left(\frac{IQ^{(\bar{m}_J)}}{2\bar{m}_J{}^2} - \left(\frac{RV^{(\bar{m}_J)}}{2\bar{m}_J}\right)^2\right)}} \sim \mathcal{N}(0,1)\,,$$

Table 13.4 Test on constant noise variance. Reported are the test statistics z_{IND}. Asterisks denote rejections at the 5% significance level

S&P500 E-mini, TTS, B					MSFT, TTS, T				
	m_J					m_J			
m_K	40	90	120	240	m_K	90	180	360	420
60	20.71*	–	–	–	180	8.86*	–	–	–
120	15.43*	2.61*	–	–	360	9.09*	5.78*	–	–
300	12.38*	1.2	0.85	−0.61	540	6.65*	4.11*	2.73*	1.6

where the third frequency $\bar{m}_L$ should be unaffected by the noise. Moreover, $m_J < m_K$.

Table 13.4 presents the test results for some pairs of m_K and m_J, where the choice of m_L is conditional on the no noise test results. Obviously, the constant variance assumption is rejected only at the very high frequencies. The results are very much in line with those reported in Awartani et al. (2009) and we conclude that noise seems to be statistically significant at frequencies higher than 1–5 min (depending on the dataset and the price series used) and that the iid noise assumption is violated only at the ultrahigh frequencies, e.g. at approximately 0.5 min TTS, B sampling for S&P500 E-mini and at 1.5 min TTS, transaction prices for MSFT.

The noise test results may serve as a guidance for the selection of the sampling frequencies in the noise variance estimation, see (13.14). In particular, m_J should be set to a frequency where no noise is present, while m_K should correspond to a very high frequency, at which, however, the iid assumption is not violated. The procedure should produce reliable noise variance estimates. Applying this method to the S&P500 E-mini, for example, yields an estimated signal-to-noise ratio $2m_J\omega^2/IV$ of about 8% using TTS, B. In contrast, $\tilde{\omega}^2$ yields a signal-to-noise ratio of about 45%.

13.8.2 Volatility Estimates

In the following we provide an empirical illustration of the various volatility estimators. To this end we first need to determine the values of the hyperparameters, which can be partly guided by the findings of the previous section. In the computation of $TTSRV_{adj}$, for example, we can set the return horizon of the slow time scale (K) to the highest frequencies without significant noise and the horizon of the fast time scale returns (J) to the highest frequencies at which the iid assumption is not violated. For the *KRV* estimator we implement the non-flat-top Parzen kernel. The optimal bandwidth H is selected to minimize the variance of the estimator, as described in Sect. 13.4. Estimates for ω^2 and IV are derived from the full sample period to obtain constant values $H(m)$ for all days.

In the implementation of the *RRV* estimator we vary K and sample at every observation in each interval. The *DRV* estimator is implemented in its first exit time

variant. Specifically, in the first half of the day the next exit time is used, while the second half of the day is based on the previous exit time. We depict the results for 8 ticks on the original scale, which is converted to the logarithmic scale at the beginning of each day using the current price level. Note that we have also experimented with alternative numbers of ticks ranging from 1 to 15 and we found that the resulting *IV* estimates are quite stable for values of $r = 3$–14.

Following Christensen et al. (2009b), we compute the subsampled version of the *QRV* estimators for three different block lengths, $m_K = 20, 40, 100$, for a fixed set of quantiles, $\bar{\lambda} = (0.80, 0.85, 0.90, 0.95)'$, and asymptotically optimal weights. These parameters are also adopted for the QRV_{iid} estimator. While the optimal value of L can be determined by a data-driven simulation of the MSE loss function, we set here L to a conservative value of 20 and $c = 0.02$, which is motivated by the finite sample performance study of Christensen et al. (2009b). Nevertheless, note that ideally L and c should be chosen at each sampling frequency m. Thus, our volatility signature plots of *QRV* should be interpreted with care.

Figures 13.4 and 13.5 depict the resulting volatility estimates over the period from 01/01/2006 to 05/31/2008 with respect to various sampling frequencies m. Clearly, the estimators that have been derived under the no noise assumption seem to be subject to severe positive biases at high frequencies, with the exception of the *BPV* for the S&P500 E-mini, which seems to be negatively biased. Interestingly, the two estimators that are robust to time-dependent noise specifications, i.e. *TTSRV* and *KRV*, appear to be unbiased. In contrast, the QRV_{iid} is negatively biased at ultrahigh frequencies, pointing towards the presence of time-dependent noise at those frequencies. Overall, the results are thus in line with our findings from the noise specification tests. Moreover, although the *DRV* estimator has been formally derived under the no noise assumption, we find empirical support for the simulation results of Andersen et al. (2009), indicating that *DRV* is robust to iid and serial dependent noise. (Note that we do not report *DRV* for MSFT due to the coarse time stamping.)

Another aspect that should be kept in mind when interpreting the volatility estimates is that some estimators do not only measure *IV*, but additionally the variation due to jumps. From a closer, i.e. zoomed-in, look at the volatility signature plots (not presented here), however, we cannot observe systematically lower volatility estimates based on the jump robust estimators. Testing for the relative contribution of jumps to total price variation based on the ratio *BPV* to *RV*, see e.g. Huang and Tauchen (2005), we do not find significant evidence for jumps at lower frequencies, e.g. lower than 5 min for the S&P500 data, respectively. Computing the tests at various sampling frequencies, similarly to the volatility signature plots, we could, however, observe that the relative jump contribution seems to be increasing strongly at the very high frequencies (e.g. for the S&P500 E-mini we observe that 20% of the total price variation is due to jumps at a sampling frequency of 1 s and reaches up to 80% for ultrahigh frequencies). Still it is interesting to understand the behavior of the statistic at higher frequencies, which, at a first glance, points to significant presence of discontinuities. *BPV* and the jump statistic are not derived under microstructure noise. We know that bid prices are rather stale, with long series

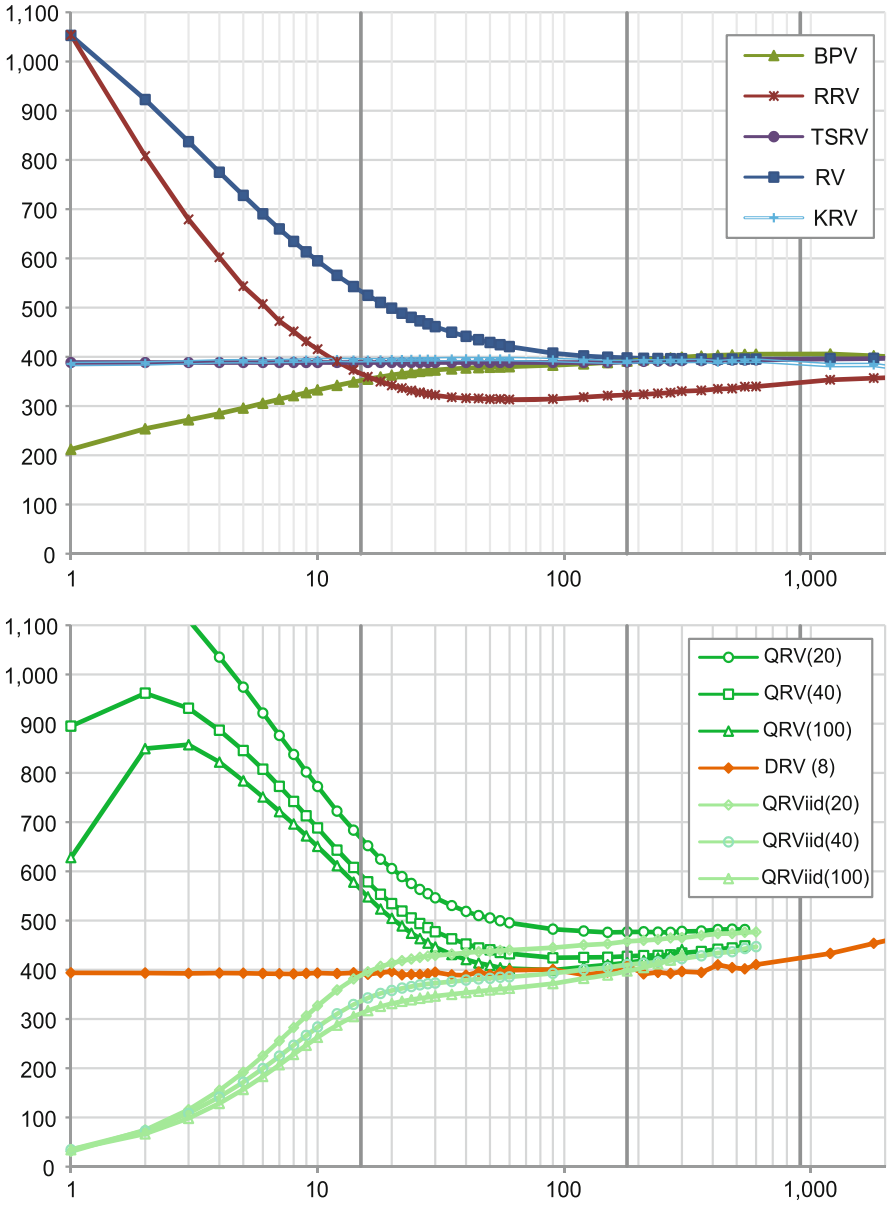

Fig. 13.4 Volatility signatures of the various high-frequency based estimators for the S&P500 E-mini based on TTS with bid prices over the period from 01/01/2006 to 05/31/2008. The *bold vertical lines* represent the frequency equal, on average, to 5 s, 1 and 5 min

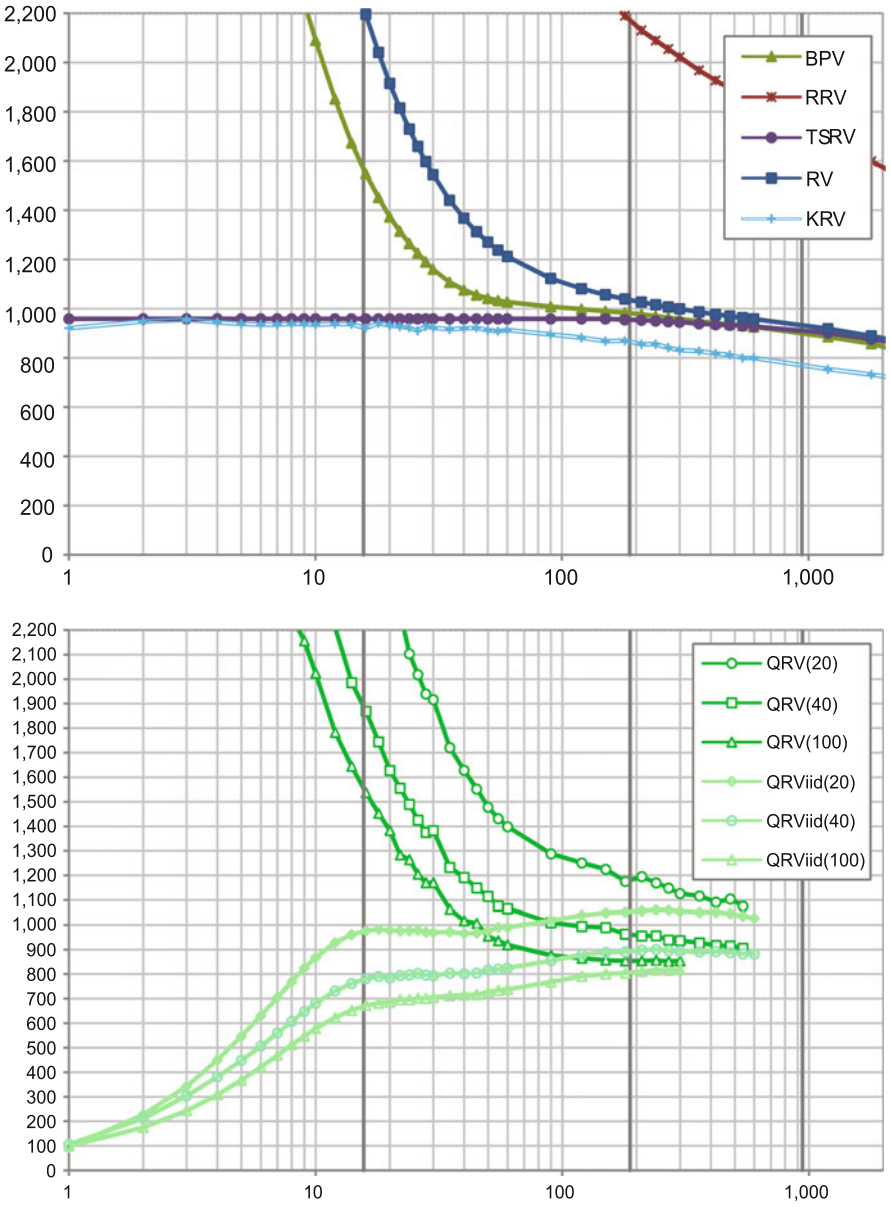

Fig. 13.5 Volatility signatures of the various high-frequency based estimators for MSFT based on QTS with transaction prices over the period from 01/01/2006 to 05/31/2008. The *bold vertical lines* represent the frequency equal, on average, to 5 s, 1 and 5 min

of zero-return observations and infrequent returns of at least one tick. If we imagine a scenario, in which we sample in TTS and there are at least two transactions between every two consecutive price moves, actually not so far from reality, then there will be no two consecutive returns, both $\neq 0$, and therefore *BPV* goes to zero. Thus, from the perspective of *BPV*, high-frequency data approaches a pure jump process as the frequency of the observations increases.

13.9 Conclusion

In this chapter we have reviewed the rapidly growing literature on volatility estimation based on high-frequency financial data. We have paid particular attention to estimators that exploit different facets of such data. We have provided a theoretical and empirical comparison of the discussed estimators. Moreover, statistical tests indicated that for the series considered in this chapter market microstructure noise can be neglected at sampling frequencies lower than 5 min, and that the common assumption of iid noise is only violated at the very high frequencies. The specific type of noise at these ultra-high frequencies is still an open question. Interestingly, estimators that are robust to serial dependent and/or endogenous noise (*TSRV*, *KRV*) seem to provide plausible estimates at all frequencies. Nevertheless, understanding the properties of estimators under different noise types could be considered in more detail within a simulation study, allowing also for a more thorough comparison of the various estimators in terms of their finite sample performance.

References

Abadi, D. J., Madden, S. R., & Ferreira, M. (2006). Integrating compression and execution in column-oriented database systems. In *Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data* (pp. 671–682).

Abadi, D. J., Madden, S. R., & Hachem, N. (2008). Column-stores vs. row-stores: How different are they really? In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data* (pp. 967–980).

Aït-Sahalia, Y., Mykland, P. A., & Zhang, L. (2005). How often to sample a continuous-time process in the presence of market microstructure noise. *Review of Financial Studies*, *18*, 351–416.

Aït-Sahalia, Y., Mykland, P. A., & Zhang, L. (2010). Ultra high frequency volatility estimation with dependent microstructure noise. *Journal of Econometrics*, *160*(1), 2011, 160–175.

Andersen, T. G., & Bollerslev, T. (1997). Intraday periodicity and volatility persistence in financial markets. *Journal of Empirical Finance*, *4*, 115–158.

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

Andersen, T. G., Bollerslev, T., Diebold, F. X., & Labys, P. (2000). Great realisations. *Risk*, *13*, 105–108.

Andersen, T. G., Bollerslev, T., Diebold, F. X., & Ebens, H. (2001). The distribution of realized stock return volatility. *Journal of Financial Economics*, *61*, 43–76.

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

Andersen, T. G., Bollerslev, T., & Diebold, F. X. (2007). Roughing it up: Including jump components in the measurement, modeling, and forecasting of return volatility. *The Review of Economics and Statistics*, *89*, 701–720.

Andersen, T. G., Dobrev, D., & Schaumburg, E. (2009). Duration-based volatility estimation. Working Paper.

Awartani, B., Corradi, V., & Distaso, W. (2009). Assessing market microstructure effects via realized volatility measures with an application to the dow jones industrial average stocks. *Journal of Business & Economic Statistics*, *27*, 251–265.

Barndorff-Nielsen, O. E., & Shephard, N. (2002a). Econometric analysis of realised volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society, Series B*, *64*, 253–280.

Barndorff-Nielsen, O. E., & Shephard, N. (2002b). Estimating quadratic variation using realized variance. *Journal of Applied Econometrics*, *17*(5), 457–477.

Barndorff-Nielsen, O. E., & Shephard, N. (2004). Power and bipower variation with stochastic volatility and jumps. *Journal of Financial Econometrics*, *2*, 1–37.

Barndorff-Nielsen, O. E., Shephard, N., & Winkel, M. (2006). Limit theorems for multipower variation in the presence of jumps. *Stochastic Processes and their Applications*, *116*, 796–806.

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., & Shephard, N. (2008). Designing realised kernels to measure the ex-post variation of equity prices in the presence of noise. *Econometrica*, *76*, 1481–1536.

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., & Shephard, N. (2009). Realized kernels in practice: Trades and quotes. *Econometrics Journal*, *12*, C1–C32.

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., & Shephard, N. (2010). Subsampled realised kernels. *Journal of Econometrics*, *160*(1), 2011, 204–219.

Boncz, P. A. (2002). *Monet: A Next-Generation DBMS Kernel for Query-Intensive Applications*. PhD thesis, Universiteit van Amsterdam: Netherlands.

Brownlees, C. T., & Gallo, G. M. (2006). Financial econometric analysis at ultra-high frequency: Data handling concerns. *Computational Statistics & Data Analysis*, *51*, 2232–2245.

Cho, C., & Frees, F. (1988). Estimating the volatility of discrete stock prices. *Journal of Finance*, *43*, 451–466.

Christensen, K., & Podolskij, M. (2007). Realized range-based estimation of integrated variance. *Journal of Econometrics*, *141*, 323–349.

Christensen, K., Podolskij, M.. & Vetter, M. (2009a). Bias-correcting the realized range-based variance in the presence of market microstructure noise. *Finance and Stochastics*, *13*, 239–268.

Christensen, K., Oomen, R., & Podolskij, M. (2009b). Realised quantile-based estimation of the integrated variance. Working Paper.

Curci, G., & Corsi, F. (2006). A discrete sine transform approach for realized volatility measurement. Working Paper.

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

David, H. A. (1970). *Order statistics*. New York: Wiley.

Eraker, B., Johannes, M., & Polson, N. (2003). The impact of jumps in volatility and returns. *Journal of Finance*, *58*, 1269–1300.

Feller, W. (1951). The asymptotic distribution of the range of sums of independent random variables. *The Annals of Mathematical Statistics*, *22*, 427–432.

Garman, M. B., & Klass, M. J. (1980). On the estimation of security price volatilities from historical data. *Journal of Business*, *53*, 67–78.

Gouriéroux, C., & Jasiak, J. (2001). *Financial Econometrics: Problems, Models, and Methods*. Princeton, NJ: Princeton University Press.

Griffin, J. E., & Oomen, R. C. A. (2010). Covariance measurement in the presence of non-synchronous trading and market microstructure noise. *Journal of Econometrics*, *160*(1), 2011, 58–68.

Hansen, P. R., & Lunde, A. (2006). Realized variance and market microstructure noise. *Journal of Business & Economic Statistics*, *24*(2), 127–161.

Hasbrouck, J. (2007). *Empirical market microstructure: The institutions, economics, and econometrics of securities trading*. UK: Oxford University Press

Huang, X., & Tauchen, G. (2005). The relative contribution of jumps to total price variance. *Journal of Financial Econometrics*, *3*(4), 456–499.

Klößner, S. (2009). Estimating volatility using intradaily highs and lows. Working Paper.

Lee, C. M. C., & Ready, M. J. (1991). Inferring trade direction from intraday data. *Journal of Finance*, *46*, 733–746.

Mancini, C. (2009). Non-parametric threshold estimation for models with stochastic diffusion coefficient and jumps. *Scandinavian Journal of Statistics*, *36*, 270–296.

Martens, M., & van Dijk, D. (2007). Measuring volatility with the realized range. *Journal of Econometrics*, *138*, 181–207.

McAleer, M., & Medeiros, M. (2008). Realized volatilty: A review. *Econometric Reviews*, *26*, 10–45.

Mosteller, F. (1946). On some useful "inefficient" statistics. *The Annals of Mathematical Statistics*, *17*, 377–408.

Oomen, R. C. (2006). Properties of realized variance under alternative sampling schemes. *Journal of Business & Economic Statistics*, *24*(2), 219–237.

Parkinson, M. (1980). The extreme value method for estimating the variance of the rate of return. *Journal of Business*, *53*, 61–65.

Podolskij, M., & Vetter, M. (2009). Estimation of volatility functionals in the simultaneous presence of microstructure noise and jumps. *Bernoulli*, *15*, 634–658.

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

Stonebraker, M., Abadi, D., Batkin, A., Chen, X., Cherniack, M., Ferreiraa, M., Lau, E., Lin, A., Madden, S., ONeil, E., ONeil, P., Rasin, A., Tran, N., & Zdonik, S. (2005). C-Store: A column-oriented DBMS. In *Proceedings of the 31st International Conference on Very Large Data Bases* (pp. 553–564). NY: ACM.

Vetter, M. (2010). Limit theorems for bipower variation of semimartingales. *Stochastic Processes and their Applications*, *120*, 22–38.

Wasserfallen, W., & Zimmermann, H. (1985). The behavior of intra-daily exchange rates. *Journal of Banking and Finance*, *9*, 55–72.

Zhang, L. (2006). Efficient estimation of stochastic volatility using noisy observations: A multi-scale approach. *Bernoulli*, *12*, 1019–1043.

Zhang, L., Mykland, P. A., & Aït-Sahalia, Y. (2005). A tale of two time scales: Determining integrated volatility with noisy high-frequency data. *Journal of the American Statistical Association*, *100*, 1394–1411.

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