

# Separating microstructure noise from volatility\*

Federico M. Bandi,<sup>†</sup> Jeffrey R. Russell<sup>‡</sup>  
*Graduate School of Business, The University of Chicago*

December 2003  
*This version: February 2005*

## Abstract

There are two variance components embedded in the returns constructed using high frequency asset prices: the time-varying variance of the *unobservable* efficient returns that would prevail in a frictionless economy and the variance of the equally *unobservable* microstructure noise. Using sample moments of high frequency return data recorded at different frequencies, we provide a simple and robust technique to identify both variance components.

In the context of a volatility-timing trading strategy, we show that careful (optimal) separation of the two volatility components of the observed stock returns yields substantial utility gains.

*Key words:* volatility, volatility timing, microstructure noise, high frequency data

*JEL Classification:* G12, C14, C22

---

\*We thank Laurence Lescourret, Benoit Perron, and seminar participants at the conference “Analysis of high frequency data and market microstructure,” Taipei (Taiwan), December 15-16, 2003, the conference “Econometric forecasting and high frequency data analysis,” Singapore, May 7-8, 2004, the European Finance Association Meetings, Maastricht (Netherlands), August 20-22, 2004, and the American Finance Association Meetings (Philadelphia), January 7-9, 2005, for discussions. We thank the editor, Bill Schwert, and a referee for comments that led to substantial improvement of the paper.

<sup>†</sup>Address: Office 408, 5807 South Woodlawn Avenue, Chicago, IL 60637. E-mail: federico.band@gsb.uchicago.edu.

<sup>‡</sup>Address: Office 452, 5807 South Woodlawn Avenue, Chicago, IL 60637. E-mail: jeffrey.russell@gsb.uchicago.edu.

# 1 Introduction

The logarithm of a recorded asset price can be written as the sum of the logarithm of the efficient price and a noise component that is induced by microstructure frictions, such as price discreteness and bid-ask bounce effects. Accordingly, the variance of continuously compounded returns based on recorded logarithmic prices depends on the variance of the underlying efficient returns and the variance of the microstructure noise components in returns. Both variance measures carry a fundamental economic significance. The variance of the efficient return process is a crucial ingredient in the practise and theory of asset valuation and risk management. The variance of the microstructure noise component reflects the market structure and the price setting behavior of market participants and thereby contains information about the market's fine-grain dynamics.

The availability of high frequency data provides researchers with an opportunity to learn about financial return volatility through robust identification methods that are simple to implement in that they are based on straightforward descriptive statistics (see the literature review of Andersen, Bollerslev, and Diebold, 2002). Nonetheless, the observation that recorded asset prices sampled at high frequencies contain a nonnegligible microstructure friction component has imposed theoretical and empirical limitations on the exploitation of the informational content of high frequency data. This paper contributes to the literature on nonparametric variance estimation through high frequency data by re-evaluating the identification potential of high frequency data. Specifically, we show that both *unobserved* components of the variance of recorded asset returns can be estimated using high frequency data sampled at different frequencies. Very high frequency asset price data can be employed to consistently estimate the microstructure noise variance. Data sampled at lower frequencies can be utilized to learn about the efficient return variance. While this latter fact is recognized in the literature, albeit not formally studied (see Andersen, Bollerslev, Diebold, and Ebens, 2001, for instance), we provide a rigorous and easily implementable procedure to purge high frequency return data of their microstructure components and extract information about the true variance dynamics by sampling at optimal frequencies. In this context, we show that the economic benefit of optimal sampling can be substantial.

Our procedure builds directly on the work of French, Schwert, and Stambaugh (1987), Schwert (1989, 1990a,b), Schwert and Seguin (1991), and more recently, Andersen, Bollerslev, Diebold, and Ebens (2001), Andersen, Bollerslev, Diebold, and Labys (2003), and Barndorff-Nielsen and Shephard

(2002, 2004). As in the early literature, as represented by French, Schwert, and Stambaugh (1987) for example, we measure variance by using sample averages of squared return data. In agreement with the recent work of Andersen, Bollerslev, Diebold, and Ebens (2001), Andersen, Bollerslev, Diebold, and Labys (2003), and Barndorff-Nielsen and Shephard (2002, 2004), we provide robust theoretical justifications for our variance estimates in the context of a continuous-time specification for the evolution of the underlying logarithmic price and the availability of high frequency return data. In contrast to both the early approaches to nonparametric variance identification and the current work on realized variance estimation, we do not simply focus on the variance dynamics of recorded stock returns; rather, we aim to identify *both* the variance of the efficient return component and the variance of the microstructure contaminations by exploiting the considerable information potential of high frequency return data.

The first stage of our analysis makes use of data sampled at the highest possible frequency. In recent work, Bandi and Russell (2004) show that sample second moments constructed using *observed* high frequency return data provide consistent estimates of the second moment of the *unobserved* microstructure frictions in a canonical model of price determination with MA(1) microstructure noise. We use this result to identify the variance of the noise component in the recorded return data. This procedure represents the substantive core of the identification of the variance of the zero-mean microstructure noise.

We then turn to the second stage of our method, namely, the identification of the genuine variance features of the efficient return process. Should the efficient price process be observable, then high sampling frequencies would yield consistent estimation through the conventional realized variance estimator (Andersen, Bollerslev, Diebold, and Labys, 2003; Barndorff-Nielsen and Shephard, 2002). If the true price process is not observable, as is the case in practise due to microstructure frictions, then the realized variance estimator is an inconsistent estimate of the integrated variance of the efficient logarithmic price process (see Bandi and Russell, 2004, and the independent work of Zhang, Mykland, and Aït-Sahalia, 2004). In effect, frequency increases provide information about the underlying integrated variance but entail noise accumulation that impacts both the bias and the variance of the realized variance estimator (Bandi and Russell, 2004; Zhang, Mykland, and Aït-Sahalia, 2004). Thus, the optimal sampling frequency will be finite and can be chosen to balance a bias/variance trade-off. Following Bandi and Russell (2004), we quantify the trade-off by writing the conditional mean-squared error (MSE) of the realized variance estimator as a function of the sampling

frequency. Subsequently, we use the estimated MSE to evaluate the optimal sampling frequency through a straightforward minimization problem. In light of this discussion, the identification of the efficient-price integrated variance is conducted at frequencies that are lower than the frequencies used to consistently estimate the second moment of the noise process.

In sum, we use sample moments of high frequency return data sampled at different frequencies to learn about two important quantities, i.e., the time-varying variance of the *unobserved* efficient return process and the variance of the *unobserved* microstructure noise contaminations. In keeping with recent approaches to model-free volatility estimation as represented by Andersen, Bollerslev, Diebold, and Ebens (2001), Andersen, Bollerslev, Diebold, and Labys (2003), and Barndorff-Nielsen and Shephard (2002, 2004), for example, little structure is required to obtain identification of the quantities of interest by virtue of robust nonparametric estimators.

Our empirical work focuses on the stocks in the S&P100 index. We employ midpoints of bid and ask prices. Using cross-sectional estimates of the standard deviations of the unobserved noise components, we find that a 1% increase in the quoted spreads translates into a 1% increase in the noise standard deviations. We also find that the median noise standard deviation is about a quarter of the median spread. Since most trades occur within the spread and the midpoints contain residual noise, the magnitude of the estimated noise standard deviations is economically plausible.

Subsequently, we employ estimated features of the noise component (namely, the second and fourth noise moments) to identify the variance of the efficient return process at frequencies that are meant to optimally balance the bias and variance of the realized variance estimator. We show that the optimal sampling frequency of the realized variance estimator depends positively on a signal-to-noise ratio, i.e., the ratio between the second moment of the noise component and the underlying integrated variance over the period. We find that the optimal frequencies are skewed to the right with a mean value of about four minutes and a median value of 3.4 minutes. The optimal frequencies vary between approximately 0.4 minutes and 13.8 minutes, with the highest frequencies being generally associated with the lowest ratios. These frequencies are potentially very different from those used and/or conjectured in the existing literature, such as the five- and the 15-minute frequencies, and deliver substantial MSE gains. In addition, the optimal frequencies vary considerably over time. We find that failing to optimally sample realized variance across periods negatively affects the dynamic and forecasting properties of the nonparametric variance estimates.

We also show that the cross-sectional relation between the estimated noise standard deviations

and the square root of the average efficient return variances is positive and significant as implied by the “operating cost” and “asymmetric information” theories of bid-ask spread determination.

By implementing the volatility-timing asset allocation strategy of Fleming, Kirby, and Ostdiek (2001, 2003), we find that there are significant utility gains to be obtained from our optimal sampling method. Specifically, we show that a risk-averse investor who is given the option of choosing volatility forecasts based on our optimal sampling method versus volatility forecasts based on the proposed frequencies in the literature would be willing to pay between roughly 25 and 300 basis points per year to achieve the level of utility that is provided by our optimal sampling procedure.

The paper proceeds as follows. Section 2 lays out the underlying price formation mechanism. In Section 3 we discuss the use of very high frequency data to identify the variance of the unobserved noise component of the recorded prices. In Section 4 we move to lower frequencies and focus on the optimal sampling of high frequency asset price data for the purpose of identifying the efficient-price integrated variance. Section 5 describes the data. Section 6 provides estimates of the variances of both unobserved components of the observed returns, i.e., microstructure noise and efficient returns. Section 7 provides simulations. Section 8 reports the economic gains of optimal sampling. Section 9 concludes.

## 2 Price formation mechanism

Let  $h$  denote a trading day. Consider  $n$  trading days and write the observed logarithmic price process at time  $ih$  as

$$\tilde{p}_{ih} = p_{ih} + \eta_{ih} \quad i = 1, 2, \dots, n, \quad (1)$$

where  $p_{ih}$  is the logarithmic efficient price, i.e., the price that would prevail in the absence of market microstructure frictions, and  $\eta_{ih}$  represents logarithmic microstructure noise. Divide each trading day into  $M$  subperiods and define the observed high frequency returns as

$$\tilde{r}_{j,i} = \tilde{p}_{(i-1)h+j\delta} - \tilde{p}_{(i-1)h+(j-1)\delta}, \quad j = 1, 2, \dots, M, \quad (2)$$

where  $\delta = h/M$ . Hence,  $\tilde{r}_{j,i}$  is the  $j$ -th intradaily observed return for day  $i$ . Such a return is defined as

$$\tilde{r}_{j,i} = r_{j,i} + \varepsilon_{j,i}, \quad (3)$$

where

$$r_{j,i} = p_{(i-1)h+j\delta} - p_{(i-1)h+(j-1)\delta}, \quad j = 1, 2, \dots, M, \quad (4)$$

and

$$\varepsilon_{j,i} = \eta_{(i-1)h+j\delta} - \eta_{(i-1)h+(j-1)\delta}, \quad j = 1, 2, \dots, M, \quad (5)$$

have obvious interpretations in terms of efficient return and microstructure contamination in the return process. For simplicity, hereafter we set  $h = 1$  without loss of generality. Below we give the assumptions that we impose on the efficient price process and microstructure noise.

**Assumption 1. (Efficient price process.)**

- (1) *The efficient logarithmic price process  $p_t$  is a continuous stochastic volatility local martingale. Specifically,  $p_t = m_t$ , where  $m_t = \int_0^t \sigma_s dW_s$  and  $\{W_t : t \geq 0\}$  is a standard Brownian motion.*
- (2) *The spot volatility process  $\sigma_t$  is càdlàg and bounded away from zero.*
- (3) *The spot volatility process  $\sigma_t$  is independent of  $W_t$  for all  $t$ .*
- (4) *The quarticity process  $Q_t = \int_0^t \sigma_s^4 ds$  is bounded almost surely for all  $t$ .*

**Assumption 2. (Microstructure noise.)**

- (1) *The random shocks  $\eta$  are i.i.d. mean zero with a bounded eighth moment.*
- (2) *The true return process  $r_{j,i}$  is independent of  $\eta_{j,i}$  for all  $i$  and for all  $j$ .*

The econometrician does not observe  $r$ , i.e., the efficient return, but rather a contaminated return series  $\tilde{r}$ , which is given by  $r$  plus an independent random shock  $\varepsilon$ . The true return process  $r$  is a stochastic volatility martingale difference sequence with bounded variance for all  $M$ . The underlying stochastic volatility is permitted to display jumps, diurnal effects, high persistence (possibly of the

long memory type), and nonstationarities.<sup>1</sup> In view of the properties of the microstructure noise component in the price process  $\eta$ , we interpret  $\varepsilon$  as being an MA(1) microstructure contamination in the return series. The MA(1) structure of the noise returns induces both a negative first-order autocovariance for the return series that is equal to  $-\sigma_\eta^2$ , i.e., the variance of the underlying i.i.d. microstructure noise  $\eta$  taken with a negative sign, as well as higher-order serial covariances that are equal to zero.

The canonical MA(1) microstructure model is known to be valid in the case of decentralized markets. In such markets, the random arrival of traders with idiosyncratic price setting behavior induces microstructure contaminations in the price process that are roughly independent, thereby providing validity to an MA(1) structure for the observed return data. The foreign exchange market is an important example (see Bai, Russell, and Tiao, 2004). The MA(1) model is more of an approximation when prices are set by a single specialist, as is the case for the NYSE, for example. However, even though the serial correlations of order higher than one can be statistically significant, their economic magnitude is often considerably smaller than the magnitude of the first-order autocorrelations. Section 5 below confirms this fact in the case of our sample of equities. We refer the reader to Bandi and Russell (2004) for a general approach to realized variance estimation in the presence of dependent microstructure noise.

Our method exploits the different orders of magnitude of the components of the returns based on recorded logarithmic prices as implied by the assumptions above. While the efficient returns are of order  $O_p(\sqrt{\delta})$  over periods of size  $\delta$ , the microstructure noise returns are  $O_p(1)$  over any period of time, however small. This is, of course, an asymptotic approximation which captures the nature of realistic price formation mechanisms and the economic difference between true and observed prices. The rounding of recorded prices to a grid *alone* makes this feature of the model compelling provided sampling does not occur between price updates. Below we expand on the economic rationale behind our modelling design.

The efficient price dynamics are modelled as being driven by a continuous process. Time is needed for market participants to acquire, digest, and react to new information. With the exception

---

<sup>1</sup>For jumps, see Bates (2000), Duffie, Pan, and Singleton (2000), Eraker, Johannes, and Polson (2003), Pan (2002), among others, and the references therein. For diurnal effects, see Andersen and Bollerslev (1997a,b, 1998), among others, and the references therein. For persistence in volatility, see Alizadeh, Brandt, and Diebold (2002), Baillie, Bollerslev, and Mikkelsen (1996), Bandi and Perron (2004), Bollerslev and Mikkelsen (1996, 1999), Chernov, Gallant, Ghysels, and Tauchen (2003), Ding, Granger, and Engle (1993), Engle and Lee (1999), Jones (2003), Meddahi (2001) and Ohanissian, Russell, and Tsay (2004), among others, and the references therein. For nonstationarities in volatility, see Comte and Renault (1998) and Bandi and Perron (2004), among others, and the references therein.

of discrete responses to important, infrequent public news announcements, the efficient price is not expected to jump from one level to another. Rather, it is expected to adjust slowly as the market comes to grips with new information. In agreement with these observations and standard approaches in the asset pricing and realized variance literature, we specify the continuously compounded return process as having an order of magnitude equal to  $O_p(\sqrt{\delta})$  over any time interval of size  $\delta$ . The characteristics of the noise process are different from the true price characteristics since recorded prices inherently reflect additional information. First, as noted, the observed prices cannot vary continuously; rather, they fall on a fixed grid of prices or ticks. The changes in the prices and midquotes are therefore discrete in nature. Furthermore, classic microstructure theory suggests that a market maker posting prices and quotes will take into consideration the nature of its operating costs and the needed reward for the provision of liquidity as well as the risks associated with asymmetric information (see the review by Madhavan, 2000, and the discussion in Section 6 below). For example, the adjustments that new limit orders induce are necessarily discrete in nature. Hence, nonnegligible adjustments can occur to the noise process regardless of how short the time interval is between price updates. It is therefore natural to consider the departures of the observed returns from the true returns as being discontinuous processes (i.e.,  $O_p(1)$ ) and, hence, consistent with our assumed structure.

In what follows we discuss the identification of the variance of the noise component,  $\sigma_\eta^2$  (Section 3), and the identification of the integrated daily variance of the underlying efficient price,  $V_i = \int_{i-1}^i \sigma_s^2 ds$  (Section 4). As indicated, the former is conducted at very high frequencies, namely, the highest frequencies at which transactions occur. The latter is performed at optimally chosen lower frequencies. Our consistency arguments rely on asymptotic increases in  $M$ , the number of high frequency return data, over a trading day. Since  $M = \frac{1}{\delta}$ , where  $\delta$  denotes the distance between intradaily observations, it is equivalent to write  $M \rightarrow \infty$  or  $\delta \rightarrow 0$ . Hereafter, we use the notation  $M \rightarrow \infty$ .

### **3 Identification at high frequencies: Volatility of the *unobserved* microstructure noise**

Sample moments of the *observed* return data can be used to consistently estimate moments of both the *unobserved* noise returns  $\varepsilon$  and, through the specification in Eq. (5) above, the price contaminations  $\eta$ . Here we focus on the variance of the noise components, i.e.,  $\mathbf{E}(\varepsilon^2)$ . Bandi and

Russell (2004, Corollary to Theorem 2) show that

$$\frac{\sum_{j=1}^M \tilde{r}_{j,i}^2}{M} \xrightarrow[M \rightarrow \infty]{p} \mathbf{E}(\varepsilon^2) \quad (6)$$

and, consequently,

$$\frac{\sum_{j=1}^M \tilde{r}_{j,i}^2}{2M} \xrightarrow[M \rightarrow \infty]{p} \mathbf{E}(\eta^2), \quad (7)$$

since  $\mathbf{E}(\varepsilon^2) = 2\mathbf{E}(\eta^2)$  by virtue of the MA(1) structure of the noise returns  $\varepsilon$ . The intuition is as follows. The sum of the squared observed returns can be written as

$$\sum_{j=1}^M \tilde{r}_{j,i}^2 = \sum_{j=1}^M r_{j,i}^2 + \sum_{j=1}^M \varepsilon_{j,i}^2 + 2 \sum_{j=1}^M r_{j,i} \varepsilon_{j,i}, \quad (8)$$

that is, as the sum of the squared true returns plus the sum of the squared noise returns and a cross-product term. The price formation mechanism that is discussed and motivated in the previous section is such that the orders of magnitude of the three terms in Eq. (8) differ since  $r_{j,i} = O_p(\sqrt{\delta})$  whereas  $\varepsilon_{j,i} = O_p(1)$ . Hence, the microstructure noise component dominates the true return process at high frequencies, i.e., for values of  $\delta$  that are small. When averaging the observed squared returns as in Eq. (6), the average of the squared noise returns constitutes the dominating term in the total average. Thus, while the remaining terms in the total average wash out due to the asymptotic order of the efficient returns, i.e.,  $O_p(\sqrt{\delta})$ , the average of the squared noise returns converges to the second moment of the noise returns as implied by Eq. (6). Finally, the result in Eq. (7) simply follows from the MA(1) structure of the return contaminations. The previous discussion suggests the following proposition.

**Proposition 1a.** *The arithmetic average of the second powers of the observed return data within the day,  $\frac{\sum_{j=1}^M \tilde{r}_{j,i}^2}{M}$ , consistently estimates the second moment of the noise returns,  $\mathbf{E}(\varepsilon^2)$ . The sampling frequency  $\delta = \frac{1}{M}$  is chosen as the highest frequency at which new information arrives (Bandi and Russell, 2004, Corollary to Theorem 2).*

If the price contaminations are i.i.d. across periods, then the following extension can be readily justified. Recall,  $n$  denotes the number of days in our sample.

**Proposition 1b.** *The arithmetic average of the second powers of the observed return data within and across days,  $\frac{\sum_{i=1}^n \sum_{j=1}^M \tilde{r}_{j,i}^2}{nM}$ , consistently estimates the second moment of the noise returns,  $\mathbf{E}(\varepsilon^2)$ . The sampling frequency  $\delta = \frac{1}{M}$  is chosen as the highest frequency at which new information arrives.*

We now turn to the identification of the variance of the efficient return process.

## 4 Identification at low frequencies: Volatility of the *unobserved* efficient return

When microstructure noise plays a role, the standard realized variance estimator loses its asymptotic validity in that the summation of an increasing (in the limit) number of contaminated squared return data entails infinite accumulation of noise (Bandi and Russell, 2004, and Zhang, Mykland, and Aït-Sahalia, 2004).

The intuition for this result comes directly from the decomposition in Eq. (8). Following Andersen, Bollerslev, Diebold, and Ebens (2001), Andersen, Bollerslev, Diebold, and Labys (2003), and Barndorff-Nielsen and Shephard (2002), one can appeal to a standard result in continuous-time process theory to show that the first term in the sum, namely  $\sum_{j=1}^M r_{j,i}^2$ , converges in probability to the daily integrated variance of the logarithmic price process, i.e.,  $\int_{i-1}^i \sigma_s^2 ds$ , as the sampling frequency increases asymptotically (i.e., as  $M \rightarrow \infty$ ). Unfortunately, the summation of an increasing number of contaminated squared returns, i.e.,  $\sum_{j=1}^M \tilde{r}_{j,i}^2$ , involves the summation of squared noise terms as well. Inevitably, the sum of the squared noise contaminations diverges to infinity almost surely. Hence, the conventional realized variance estimator cannot converge to the object of interest, i.e., integrated variance, when the return data are affected by microstructure frictions as implied by a realistic price formation mechanism. Instead, it grows without bound with increases in the sampling frequency (Bandi and Russell, 2004).

Despite this observation, one can extract information from the traditional realized variance estimator by sampling the observed return data at frequencies that optimally balance the finite sample bias and the variance of the estimator as summarized by its conditional (on the volatility path) MSE. In effect, frequency increases cause finite sample bias increases due to noise accumulation. At the same time, frequency increases cause reductions in the theoretical dispersion of the estimator. Conversely, the realized variance estimator is expected to be less biased at lower frequencies, since

noise plays a relatively smaller role when  $\delta$  is large, but considerably more volatile. Bandi and Russell (2004) quantify the trade-off between the finite sample bias and variance of the realized variance estimator at any frequency in terms of a conditional MSE. Under Assumptions 1 and 2 above, they show that the form of the MSE is

$$\mathbf{E}_\sigma \left( \sum_{j=1}^M \tilde{r}_{j,i}^2 - \int_{i-1}^i \sigma_s^2 ds \right)^2 = 2 \frac{1}{M} (Q_i + o(1)) + M\beta + M^2\alpha + \gamma, \quad (9)$$

where  $Q_i = \int_{i-1}^i \sigma_s^4 ds$  is the bounded (from Assumption 1(4)) quarticity of Barndorff-Nielsen and Shephard (2002) and the parameters  $\alpha, \beta$ , and  $\gamma$  are defined as follows:

$$\alpha = (\mathbf{E}(\varepsilon^2))^2, \quad (10)$$

$$\beta = 2\mathbf{E}(\varepsilon^4) - 3(\mathbf{E}(\varepsilon^2))^2, \quad (11)$$

and

$$\gamma = 4\mathbf{E}(\varepsilon^2) \left( \int_{i-1}^i \sigma_s^2 ds \right) - \mathbf{E}(\varepsilon^4) + 2(\mathbf{E}(\varepsilon^2))^2. \quad (12)$$

Thus, the optimal (in a conditional MSE sense) frequency for sampling high frequency observations to identify the integrated variance of the logarithmic price process through the contaminated realized variance estimator  $\sum_{j=1}^M \tilde{r}_{j,i}^2$  is given by the minimum of the MSE expansion in Eq. (9). When we specialize the analysis to an underlying price process modelled as a constant variance diffusion in the presence of Gaussian microstructure noise, the expansion in Eq. (9) reduces to the MSE expansion in Ait-Sahalia, Mykland, and Zhang (2005).

The necessary ingredients to compute the minimum of the MSE are the second moment of the noise process,  $\mathbf{E}(\varepsilon^2)$ , the fourth moment of the noise process,  $\mathbf{E}(\varepsilon^4)$ , and the quarticity,  $\int_{i-1}^i \sigma_s^4 ds$ . The second moment of the noise returns can be readily estimated by using the procedure in Proposition 1b. As for the fourth moment of the noise term, a similar argument to the one in the previous section suggests the following propositions.

**Proposition 2a.** *The arithmetic average of the fourth powers of the observed return data within the day,  $\frac{\sum_{j=1}^M \tilde{r}_{j,i}^4}{M}$ , consistently estimates the fourth moment of the noise returns,  $\mathbf{E}(\varepsilon^4)$ . The sampling frequency  $\delta = \frac{1}{M}$  is chosen as the highest frequency at which new information arrives (Bandi and Russell, 2004, Corollary to Theorem 2).*

As before, if the price contaminations are i.i.d. across periods, then the following extension can be derived. Again, recall that  $n$  denotes the number of days in our sample.

**Proposition 2b.** *The arithmetic average of the fourth powers of the observed return data within and across days,  $\frac{\sum_{i=1}^n \sum_{j=1}^M \tilde{r}_{j,i}^4}{nM}$ , consistently estimates the fourth moment of the noise returns,  $\mathbf{E}(\varepsilon^4)$ . The sampling frequency  $\delta = \frac{1}{M}$  is chosen as the highest frequency at which new information arrives.*

We are now left with the remaining ingredient of the MSE expansion, namely, the quarticity  $\int_{i-1}^i \sigma_s^4 ds$ . The traditional quarticity estimator as introduced by Barndorff-Nielsen and Shephard (2002), i.e.,  $\frac{M}{3} \sum_{j=1}^M \tilde{r}_{j,i}^4$ , cannot be a consistent estimator (as  $M \rightarrow \infty$ ) in the presence of microstructure noise. In fact, as is the case for realized variance, frequency increases cause infinite noise accumulation. Consequently, we construct quarticity estimates by sampling at low frequencies. In view of the attention that the 15-minute sampling interval has received in empirical work (see Andersen, Bollerslev, Diebold, and Labys, 2000, for instance), we choose to sample every 15 minutes. While this sampling frequency can be conservative (i.e., lower than optimal) in the case of very liquid stocks, plausible alternative sampling intervals for the quarticity can be shown to have a relatively small effect on the minimum of the conditional MSE expansion and, consequently, on the optimal sampling frequency of the realized variance estimator (see the simulations in Section 7). We summarize the previous discussion with the following remark.

**Remark 1.** *Following Barndorff-Nielsen and Shephard (2002), we employ a rescaled average of the fourth powers of the observed return data within the day,  $\frac{M}{3} \sum_{j=1}^M \tilde{r}_{j,i}^4$ , to estimate the daily quarticity of the underlying logarithmic price process,  $\int_{i-1}^i \sigma_s^4 ds$ . In our empirical work we use 15-minute sampling intervals.*

Finally, we turn to the optimal sampling of the realized variance estimator  $\sum_{j=1}^M \tilde{r}_{j,i}^2$ .

**Proposition 3.** *The optimal sampling frequency is chosen as the value  $\delta^* = \frac{1}{M^*}$  with*

$$\left\{ M^* = M : 2M^3 \hat{\alpha} + M^2 \hat{\beta} - 2\hat{Q}_i = 0 \right\}, \quad (13)$$

where

$$\hat{\alpha} = \left( \frac{\sum_{i=1}^n \sum_{j=1}^M \tilde{r}_{j,i}^2}{nM} \right)^2, \quad (14)$$

$$\hat{\beta} = 2 \frac{\sum_{i=1}^n \sum_{j=1}^M \tilde{r}_{j,i}^4}{nM} - 3 \left( \frac{\sum_{i=1}^n \sum_{j=1}^M \tilde{r}_{j,i}^2}{nM} \right)^2, \quad (15)$$

and,

$$\hat{Q}_i = \frac{M}{3} \sum_{j=1}^M \tilde{r}_{j,i}^4. \quad (16)$$

(See Bandi and Russell, 2004, Second Corollary to Theorem 3.) The sampling frequency for estimating the quantities in the terms  $\hat{\alpha}$  and  $\hat{\beta}$  follows from Propositions 1b and 2b. The relevant sampling frequency for the quarticity estimator  $\hat{Q}_i$  is discussed in Remark 1.

When the optimal sampling frequency is high, the following rule-of-thumb applies:

$$M^* \sim \left( \frac{Q_i}{(\mathbf{E}(\varepsilon^2))^2} \right)^{1/3}. \quad (17)$$

From a theoretical perspective, the approximation in Eq. (17) is useful in that it highlights the main components of the optimal frequency, namely, the underlying quarticity and the squared second moment of the noise process. The larger the squared second moment of the noise process with respect to the quarticity of the underlying efficient price, the stronger is the relative noise and the smaller  $M$  should be to avoid substantial contaminations. From an applied perspective, the rule-of-thumb represents a valid and immediate methodology for choosing the optimal frequency for a variety of stocks with different liquidity properties. Section 6 provides empirical evidence for this result. In general, the higher the true optimal frequency, the better the approximation.

**Proposition 4.** *The approximate optimal sampling frequency is chosen as the value  $\delta^* = \frac{1}{M^*}$  with*

$$M^* = \left( \frac{\hat{Q}_i}{\hat{\alpha}} \right)^{1/3}, \quad (18)$$

where

$$\hat{\alpha} = \left( \frac{\sum_{i=1}^n \sum_{j=1}^M \tilde{r}_{j,i}^2}{nM} \right)^2 \quad (19)$$

and

$$\widehat{Q}_i = \frac{M}{3} \sum_{j=1}^M \widetilde{r}_{j,i}^4. \quad (20)$$

(See Bandi and Russell, 2004, Remark 8.) The sampling frequency for estimating the term  $\widehat{\alpha}$  follows from Proposition 1b. The relevant sampling frequency for the quarticity estimator  $\widehat{Q}_i$  is discussed in Remark 1.

The conditional MSE in Eq. (9) applies to each individual period, thereby requiring repeated applications of the procedure. In our empirical work in Section 6 we obtain an estimated optimal frequency (as in Proposition 3) as well as an approximate optimal frequency (as in Proposition 4) that are valid for the entire data set by working with an integrated version of the conditional MSE in Eq. (9). This is done by minimizing the average (over  $i$ ) of the individual conditional MSEs. Specifically, we implement the procedure in Propositions 3 and 4 by simply replacing the individual  $\widehat{Q}_i$ 's (as defined in Remark 1) with  $\frac{1}{n} \sum_{i=1}^n \widehat{Q}_i$ .

Before turning to a description of the data, we should mention recent contributions that provide solutions to the issue of integrated variance estimation through the use of high frequency data in the presence of market microstructure noise. Following early work by Garman and Klass (1980), Parkinson (1980), and Beekers (1983), among others, Alizadeh, Brandt, and Diebold (2002) and Brandt and Diebold (2004) suggest the use of the so-called range, i.e., the difference between the highest and the lowest logarithmic price over a fixed sampling interval (see, also, Andersen and Bollerslev, 1998). While the estimated range does not entail infinite noise accumulation since its average deviation from the underlying “efficient range” is roughly equal to the average bid-ask bounce, it is a less efficient volatility measure than the conventional realized volatility estimator. Zhang, Mykland, and Ait-Sahalia (2004) provide a consistent estimator of the integrated variance of the logarithmic price process in Assumption 1 in the presence of market frictions as described in Assumption 2 above. Their method is based on subsampling. Specifically, it relies on a variance estimator that entails taking an arithmetic average of (bias-corrected) realized variance estimates constructed on the basis of different, appropriately chosen, sampling grids. Under an MA(1) microstructure model and ideal conditions, the estimator in Zhang, Mykland, and Ait-Sahalia (2004) is theoretically more efficient than both the range and the standard realized variance estimator. However, the finite sample properties of this estimator are unknown in practise. In addition, the subsampling approach partly

foregoes the simplicity and empirical appeal of the more conventional approaches as discussed by Andersen, Bollerslev, and Diebold in their 2002 survey paper. Hansen and Lunde (2004a) propose a Newey-West bias correction for realized variance in presence of correlated noise.

This paper remains in the confines of standard variance estimates constructed as simple averages of squared return data as recommended in early work by French, Schwert, and Stambaugh (1987), Schwert (1989, 1990a,b), and Schwert and Seguin (1991), and as more recently justified by Andersen, Bollerslev, Diebold, and Ebens (2001), Andersen, Bollerslev, Diebold, and Labys (2003), and Barndorff-Nielsen and Shephard (2002, 2004). In this context, we explicitly address an issue that the extant literature has raised but never tackled explicitly, namely, how to preserve the simplicity of the realized variance estimator while optimally trading off efficiency versus robustness to market microstructure frictions. In their 2001 paper, Andersen, Bollerslev, Diebold, and Ebens write “Following the analysis in Andersen and Bollerslev (1997a), we rely on artificially constructed five-minute returns....The five-minute horizon is short enough that the accuracy of the continuous record asymptotics underlying our realized volatility measures work well, and long enough that the confounding influences from market microstructure frictions are not overwhelming.” This paper provides theoretical content to the previous statement as well as to similar statements in the applied finance literature. We offer a straightforward methodology to optimally sample high frequency return data for the purpose of exploiting the information potential of the classical realized variance estimator. Additionally, we provide a characterization of the economic benefit of optimal sampling (see Section 8)

Hansen and Lunde (2004b) and Oomen (2004a,b) have recently provided interesting theoretical extensions of the optimal sampling methods discussed in Bandi and Russell (2004) and the present work. Hansen and Lunde (2004b) study the MSE properties of a bias-corrected estimator for realized variance in the presence of MA(1) noise and discuss its finite sample benefits. The Hansen and Lunde estimator is in the tradition of Zhou’s first-order bias-corrected estimator (Zhou, 1996). Oomen (2004a) conducts optimal sampling by considering conditional MSE expansions for biased and bias-corrected (as in Hansen and Lunde (2004b)) realized variance estimates in the presence of an underlying efficient price process modelled as a pure jump process of finite variation. The analysis in Oomen (2004a,b) is in both calendar and transaction time. We now describe the data.

## 5 The data: S&P100 stocks

It is common practise in the realized variance literature to use midpoints of bid-ask quotes as measures of the true prices. While these measures are affected by residual noise in that there is no theoretical guarantee that the midpoints coincide with the underlying efficient prices, they are generally less noisy measures of the efficient prices than are transaction prices since they do not suffer from bid-ask bounce effects. Thus, in agreement with the realized variance literature, in this paper we use midpoints of bid-ask quotes to measure prices. Accordingly, the specification in Eq. (1) should be interpreted as a model of midquote determination based on efficient price and residual microstructure noise.

We study the stocks in the S&P100 index. The data come from the Trade and Quote (TAQ) database. The observations refer to the month of February 2002 and correspond to quotes posted on two exchanges, the NYSE and the MIDWEST. Ideally, for NYSE-listed stocks we would like to use all available quotes from the consolidated market to construct the midquote return series. However, quotes from the satellite markets tend to be far more noisy than those generated by the NYSE specialist. A notable exception is the MIDWEST exchange, which delivers quotes whose variability is comparable to the variability of the NYSE quotes. We therefore construct midquote return series for the NYSE stocks by using quote updates obtained from both the NYSE and the MIDWEST exchange. Only NASDAQ quotes are available for NASDAQ stocks. We use a mild filter and remove quotes whose associated price changes and/or spreads were larger than 10%.

In what follows we estimate the moments of the noise component using quote-to-quote continuously compounded returns. The realized variance estimates and the quarticity estimates are constructed using fixed calendar time intervals. The prevailing quote method is used when there is no quote available.

Table 1 contains information on the individual stocks. We report the average duration, the average spread, the average price, the estimated variance of the noise component (from Proposition 1b), the estimated fourth moment of noise component (from Proposition 2b), the estimated approximate optimal sampling interval, the estimated true optimal sampling interval, and the average daily variance of the efficient return process as computed using the optimal sampling frequency from Proposition 3.

In Fig. 1 we represent the histogram of the return first-order autocorrelations of the 100 stocks

in our sample. In agreement with our assumed MA(1) structure, virtually all of the first-order serial correlations are negative. Furthermore, they are generally highly statistically significant. While the higher-order (up to order four) autocorrelations are sometimes significant, their economic relevance is marginal in that their absolute values are substantially smaller than the absolute values of the first-order serial correlations. The second-order autocorrelations, for example, are smaller, on average, than the first-order autocorrelations by a factor of three. Hence, the model in Section 2 captures the main economic effects in our data.

## 6 Separating microstructure noise from volatility: The cross-section of S&P100 stocks

### 6.1 *The noise variance*

We use the estimator in Proposition 1b to consistently identify the variance of the contaminations in the logarithmic midquotes of our cross-section of S&P100 stocks.

Fig. 2 contains a histogram of the estimated standard deviations,  $\sigma_\eta$ . The reported values should be roughly interpreted as standard deviations of the percentage differences between the midpoint bid-ask quotes and the corresponding efficient prices. The cross-sectional distribution of the standard deviations is skewed to the right with a mean value of 0.000732 and a median value of 0.000597. The numbers show that the bid-ask midpoints contain residual noise that needs to be taken into consideration when estimating the genuine volatility dynamics of the underlying efficient prices; we do this in the next subsection.

It is interesting to compare the standard deviations of the noise terms to the average quoted spreads, namely, the average differences between the quoted logarithmic ask prices and the corresponding logarithmic bid prices. We report the histogram of the average quoted spreads in Fig. 3. The cross-sectional distribution of the spreads is fairly symmetric with a mean of 0.002 and a standard deviation of 0.0007. The relation between the noise standard deviations and the average spreads is nonlinear and heteroskedastic (see Fig. 4). Not surprisingly, wider spreads are associated with larger market microstructure contaminations in the observed return process. A log-log regression of the estimated standard deviations on the average quoted spreads indicates that the elasticity between the standard deviations of the unobserved noise components in the recorded midquotes and the average spreads is close to one, thereby implying that a 1% increase in the latter translates into a 1% increase in the former (see Table 2). More importantly, the median noise standard deviation

is about a quarter of the median average spread. Since most trades occur within the spread and the midpoints contain residual noise, the magnitude of the estimated noise standard deviations is economically meaningful.

## 6.2 *The efficient return variance*

Figs. 5 and 6 are histograms of the optimal sampling intervals and the approximate optimal sampling intervals from Proposition 3 and 4, respectively.<sup>2</sup> The mean and median values of the optimal sampling intervals are 3.98 minutes and 3.4 minutes. The minimum value is 0.40 minutes whereas the estimated maximum value in our sample is 13.8 minutes. The mean and median values of the approximate optimal sampling intervals are 3.8 and 3.35 minutes, respectively. The minimum value is again about 0.40. The maximum value is 12.6 minutes. Hence, the rule-of-thumb has a slight tendency to understate the true optimal interval. A further comparison between the two measures is contained in the scatterplot in Fig. 7. Fig. 8 contains a scatterplot of the logarithmic values of the MSE of the quadratic variation estimator based on the optimal sampling intervals plotted against the corresponding logarithmic MSE values obtained by employing the rule-of-thumb. Virtually all estimates fall on the 45-degree line. Hence, even when the approximation that the rule-of-thumb delivers is not very accurate, in the sense that the optimal and approximate intervals do not appear to be extremely close, the MSE loss is minimal. In our sample, therefore, the rule-of-thumb gives an immediate feel for the kind of frequencies that one should utilize in order to optimally balance the bias and variance of the realized variance estimator.

It is interesting to notice that the optimal intervals are related to an obvious signal-to-noise ratio, i.e., the ratio between the variance of the noise component and the variance of the underlying efficient price (see Fig. 9). Fig. 10 represents the estimated MSEs of three stocks with different signal-to-noise ratios. Specifically, we consider GS (Goldman Sachs), SBC (SBC Communications), and XOM (Exxon Mobile Corporation). The ratio is smallest for GS (GS corresponds to the first decile of the distribution of the ratios) and highest for XOM (XOM corresponds to the ninth decile of the distribution of the ratios). The SBC ratio constitutes the median value of the ratios in our sample. The estimated MSEs unambiguously show that different noise properties induce different optimal sampling frequencies (2.2 minutes for GS, 3.42 minutes for SBC, and 6.6 minutes for XOM;

---

<sup>2</sup>As indicated in Remark 1, we use 15-minute sampling intervals to compute the quarticity. However, we find that using 10- or 20-minute intervals for the quarticity has no significant effect on the resulting optimal sampling frequencies of the realized variance estimator in our sample. This issue is considered further in the simulations in Section 7.

see also Table 1). Furthermore, they show that the potential loss that would be brought about by suboptimal sampling frequency choices changes across stocks. The loss depends on the steepness of the MSE around its minimum value.

We now turn to the loss that would be induced by employing possibly suboptimal (in an MSE sense) frequencies for the totality of the S&P100 stocks in our sample. We focus on the comparison between our optimal frequency from Proposition 3 and two sampling frequencies that have been either used or suggested in empirical work on integrated variance estimation to avoid strong contaminations induced by market microstructure frictions. These two frequencies are five minutes (Andersen, Bollerslev, Diebold, and Ebens, 2001, among others) and 15 minutes (Andersen, Bollerslev, Diebold, and Labys, 2000, among others). Specifically, we plot the ratios between the MSE values obtained by using the five-minute frequency and our optimal frequency from Proposition 3 (Fig. 11) as well as the ratios between the MSE values obtained by using the 15-minute frequency and, again, our optimal frequency (Fig. 12). Since many optimal sampling intervals are near five minutes, the loss is not substantial when a five-minute interval is used. Exactly 50% of the MSE ratios are between 1 and 1.17, thereby implying that for 50% of the stocks in our sample the upper bound on the MSE loss is 17%. The average MSE ratio is 1.53. The maximum ratio is about eight. Thus, if one had to choose one frequency for all stocks and all periods, choosing the five-minute frequency would be a reasonable approximation to invoke. Of course, substantial losses are possible for individual stocks as testified by a mean loss that is higher than 50%. Given the average magnitudes of our estimated optimal frequencies, we expect monotonically increasing losses as we move from the five-minute frequency to the 15-minute frequency. When considering the 15-minute frequency, the median value of the ratios is 2.6 whereas the mean value is 3.67. The minimum value is one while the maximum value is 24.2. Fig. 13 presents the empirical distribution of our average daily variance estimates based on the optimal sampling frequency in Proposition 3.

### **6.3 *The noise variance versus the efficient return variance***

We quantify the relation between the standard deviations of the noise components and the square root of the average daily variances of the efficient prices by running a regression of the latter on the former (see Table 3). The relation is positive and significant. The intercept and slope coefficient are equal to 0.000333 (with a  $t$ -statistic of 5.12) and 0.017 (with a  $t$ -statistic of 7.23), respectively. The  $R^2$  of the regression is 34.8%.

Interestingly, conventional theories of transaction cost determination provide a justification for the positive cross-sectional relation between noise standard deviation and efficient price volatility. The “operating cost” theory states that measures that are positively correlated with liquidity and ease of inventory adjustment have a negative impact on the magnitude of the quoted spreads. This is due to the fact that the market maker has to be compensated for providing liquidity. Furthermore, a risk-averse market maker prefers to make all profits off the bid-ask spread and avoid exposure to adverse movements in the price due to short or long positions. The market maker, therefore, adjusts the spreads to offset positions that are overly long or short with respect to a desired inventory target. The “asymmetric information” theory recognizes that the market-maker is likely to trade with investors that have superior information. Hence, the market maker modifies the spreads to extract a profit from uninformed traders in order to obtain compensation for the expected loss to informed traders. We refer the interested reader to Stoll (2000) and the reference therein for further discussions. Hence, lower liquidity and higher risk of asymmetric information have a positive impact on the magnitude of the spreads as well as on the frequency of the quote updates (Easley and O’Hara, 1992). Everything else equal, i.e., given a certain efficient price, lower liquidity and higher asymmetric information risk should have a positive impact on  $\sigma_\eta$ , the standard deviation of the noise component in the midquotes. The efficient price variance plays the same role in both theories of quoted spread determination. Higher uncertainty about the asset’s value implies higher likelihood of adverse price movements and, in turn, higher inventory risk, mostly in the presence of severe imbalances that must be offset. Equivalently, higher uncertainty about the fundamental value of the asset increases the risk of transacting with traders with superior information. Hence, high efficient price volatility should be associated with a high standard deviation of the midquote noise. Indeed, this is what we find.

## 7 Simulations

Bandi and Russell (2004) perform simulations to show that very high sampling frequencies allow one to consistently estimate the second and fourth moments of the microstructure noise by virtue of sample analogues based on continuously compounded observed returns (as implied by Proposition 1b and Proposition 2b). The remaining ingredient of the conditional MSE expansion in Eq. (9) is the quarticity  $Q_i$ . In this section we show that quarticity estimates based on the empirically appealing, but often suboptimal, 15-minute frequency deliver rather precise measurements of the

optimal frequency of the realized variance estimator as well as very reasonable sampling distributions. Specifically, by simulating processes with different noise features, we show that the 15-minute sampling interval is a valid, albeit possibly conservative, interval to use to identify the quarticity of the logarithmic price process for the purpose of variance estimation. We show that this observation is true for a variety of stocks with different noise characteristics, thereby confirming the validity of Remark 1.

We simulate a data generating process for the logarithmic efficient price process whose dynamics are driven by the stochastic differential equation

$$dp_t = \sigma_t dW_t, \quad (21)$$

with

$$d\sigma_t^2 = \kappa(\bar{v} - \sigma_t^2)dt + \varpi\sqrt{\sigma_t^2}dB_t, \quad (22)$$

where  $\{W_t, B_t : t \geq 0\}$  denotes Brownian motion on the plane. We set the persistence parameter  $\kappa$  of the time-varying spot variance equal to 0.01. We normalize the mean spot variance to one and hence set  $\bar{v}$  equal to one. The parameter  $\varpi$  is set to 0.05. We assume that the logarithmic noise  $\eta$  is normally distributed<sup>3</sup> with mean zero and variance equal to  $\xi^2$ , where  $\xi$  can take one of three possible values, as follows. We compute the average daily realized variances for the stocks in the sample ( $\bar{V}$ ) and calculate the median ratio between the variance of the noise return and  $\bar{V}$ , i.e.,  $\frac{2\widehat{\sigma}_\eta^2}{\bar{V}}$ , as well as the equivalent ratios corresponding to the first and the ninth decile of the distribution of the ratios. These ratios correspond to SBC, GS, and XOM, respectively. Finally, we choose  $\xi$  equal to  $\frac{1}{\sqrt{2}}\sqrt{\frac{2\widehat{\sigma}_\eta^2}{\bar{V}}}$ . Since the mean spot variance is normalized to one, our choices of  $\xi$  replicate extreme and median features of the data. When ordered from the smallest to the largest, the three values of the ratio  $\frac{2\widehat{\sigma}_\eta^2}{\bar{V}}$  are 0.00041, 0.00092, and 0.0024. We simulate 1,000 contaminated return series around a single realization of the spot variance over a period of 6.5 hours. More precisely, we employ the specification in Eq. (22) to simulate second-by-second a variance path given an initial value of  $\sigma_t^2$  equal to the unconditional mean of one. Holding the variance path fixed, we then simulate second-by-second true returns using Eq. (21) and second-by-second observed returns as in Eq. (2) given the normality assumption on the logarithmic noise process.

---

<sup>3</sup>Without loss of generality, we assume Gaussianity only for simplicity in the simulations. Our theory is robust to alternative distributional assumptions.

For any assumed model, the simulated series is used to find an optimal (from an MSE perspective) sampling interval for the quarticity. We then compare the distribution of the estimated optimal sampling intervals for the realized variance estimator obtained by using the 15-minute frequency for the quarticity to the corresponding distribution obtained by using the quarticity optimal frequency. We start with GS. When using the quarticity optimal sampling interval (i.e., 2.13 minutes), the empirical distribution of the realized variance optimal intervals (Fig. 14) is fairly concentrated around the true optimal value (i.e., 2.8 minutes) and is extremely informative about the types of frequencies that one should employ. The minimum value is 2.4 minutes while the maximum value is only four minutes. While there is an upward bias (the mean and median values are 3.17 and 3.2, respectively), the bias goes in the right direction in that it prevents us from sampling at frequencies that would entail substantial accumulation of noise. We now consider the same simulated distributions for a suboptimal value of sampling frequency for the quarticity, namely the 15-minute frequency. Even though the variability of the estimated intervals increases slightly, the bias increase, as represented by the mean and median values of the empirical distribution (i.e., 3.66 and 3.6), is small (Fig. 15). Furthermore, the MSE loss induced by sampling at the estimated mean and median values rather than at the true optimal value is minimal.

Using a suboptimal 15-minute frequency for the quarticity is a conservative choice for a stock such as GS whose quarticity optimal sampling interval is close to two minutes. Our findings should (and do) improve when using the 15-minute frequency for stocks whose optimal sampling frequency for the quarticity is lower than two minutes. The SBC and XOM optimal sampling intervals for the quarticity are 4.26 and 8.53 minutes, respectively. In these cases, we find that the biases induced by suboptimal sampling choices for the quarticity are smaller, in percentage terms, than in the GS case. Such biases are smaller for XOM than for SBC in percentage terms. As earlier in the case of GS, the MSE losses due to sampling at the estimated mean and median optimal intervals rather than at the true optimal intervals are immaterial.<sup>4</sup>

## 8 The economic benefit of optimal sampling

One way to assess the economic benefit of optimal sampling is to consider a volatility timing trading strategy. This procedure requires moving from a cross-sectional analysis, as in the previous sections, to a time-series perspective. In general, the optimal frequencies may vary over time. In this

---

<sup>4</sup>The corresponding figures can be provided by the authors upon request.

section we allow for time-varying optimal frequencies. Specifically, we evaluate the daily optimal frequencies in Proposition 3 by constructing estimates of the second and fourth noise moments using averages of observed returns *within* days, as in Propositions 1a and 2a. In other words, we do not average observed return data across days when computing the noise moments and the quarticity. Based on an estimated time series of optimal sampling frequencies we construct daily realized variance estimates. Out-of-sample forecasts of optimally sampled realized variance estimates are then employed in the context of a volatility timing-based asset allocation procedure to show the utility gains that are provided by optimal sampling versus suboptimal sampling choices.

In this exercise we use 11 years' worth of high frequency data for the representative stock SBC. Recall that SBC corresponds to the median value of the empirical signal-to-noise ratios for the month of February 2002, i.e., the month used in our previous cross-sectional analysis. The relevant quotes pertain to the period between January 1993 and December 2003. We use the same filter as in the previous sections. Furthermore, because in this expanded sample some days around holidays are not full trading days, we remove these days (fewer than 20). In Fig. 16 we report the time series of daily optimal sampling intervals. We notice that the magnitude and variability of the optimal intervals have been decreasing since 1993. Specifically, optimal intervals higher than 20 minutes were not uncommon at the beginning of the sample. The optimal intervals that prevail since the beginning of 1997 are smaller in magnitude and more stable than those at the beginning of the sample. Hence, using a single sampling frequency across periods might alter the statistical properties of the resulting variance estimates.

We employ the optimally sampled realized variance estimates, variance estimates obtained by using the five-minute frequency, and variance estimates obtained by using the 15-minute frequency to construct out-of-sample forecasts of daily variances based on an ARFIMA model as in Andersen, Bollerslev, Diebold, and Labys (2003). We set the orders of the autoregressive and moving average representations equal to two. While the ARMA parameters are re-estimated in real time, the fractional parameter is fixed at the estimated value obtained on the basis of the full sample. We use the traditional Geweke and Porter-Hudak (GPH) estimator to estimate the  $d$  parameter (Geweke and Porter-Hudak (1983)). The estimated  $d$  values are equal to 0.45 in the case of the optimally sampled realized variance estimates, 0.44 in the case of the realized variance estimates constructed using five-minute intervals, and 0.37 in the case of the realized variance estimates constructed using 15-minute intervals. We utilize almost a year's worth of data, i.e., 200 observations, to construct

the first forecast. The total number of out-of-sample forecasts  $m$  is equal to 1,870.

In Fig. 17 we plot the realized variance forecasts obtained on the basis of time-varying optimal sampling frequencies and the corresponding forecasts obtained by using the ubiquitous five-minute interval. Our theory suggests that using five-minute intervals results in sampling that is too frequent in periods during which microstructure noise is substantial and the corresponding optimal frequencies are considerably lower than five minutes. Not surprisingly, using five-minute intervals during the beginning of our sample results in substantial accumulation of noise and, therefore, upward-biased variance forecasts.

Before turning to asset allocation, we implement a forecasting exercise in the spirit of Andersen, Bollerslev, Diebold, and Labys (2003). Namely, we evaluate the forecasting power of our optimally sampled realized variance, of the variance constructed using five-minute intervals, and of the variance constructed using 15-minute intervals. Since true variance is unobserved, we regress each variance measure on one-day-ahead forecasts obtained by virtue of all three measures. Of course, one should expect a model built directly for a specific variance series to deliver superior forecasts of the same series. The results are reported in Tables 4 through 6. Remarkably, we find that (1) our optimally sampled realized variance helps forecast both realized variance sampled every five minutes and realized variance sampled every 15 minutes, (2) the forecasting power of our optimally sampled realized variance is only slightly smaller than the forecasting power of realized variance sampled every five (15) minutes when forecasting realized variance sampled every five (15) minutes, (3) our optimally-sampled realized variance can *only* be forecasted using optimally sampled realized variance, and (4) our optimally sampled realized variance can be forecasted more accurately using optimally sampled realized variance than the other two measures using a combination of predictors. Results (1) and (2) suggest that the information content of optimally sampled realized variance is substantial. Result (3) and (4) suggest that our optimally sampled measure is less noisy than competing measures and, hence, more predictable.

We now turn to asset allocation. Fleming, Kirby, and Ostdiek (2001, 2003) provide a methodology to evaluate the economic benefits of asset allocation strategies relying on volatility timing. We adapt their procedure to our framework and use the out-of-sample forecasts to assess the utility gains that are furnished by optimal sampling versus sampling based on five- and 15-minute intervals. We denote by  $R^f$  and  $R$  the net risk-free return and the net return on a generic risky asset (SBC, in our case), respectively. Assume the representative investor has a conditional mean-variance utility given by

$$E_t(R_{t,t+1}^p) - \frac{\lambda}{2} V_t(R_{t,t+1}^p), \quad (23)$$

where  $R_{t,t+1}^p$  is the return on a portfolio that invests a share of wealth  $\varpi_t$  in the risky asset between time  $t$  and time  $t + 1$ , i.e.,

$$R_{t,t+1}^p = R_{t,t+1}^f + \varpi_t(R_{t,t+1} - R_{t,t+1}^f). \quad (24)$$

The relevant time interval is one day. The optimal allocation to the risky asset is  $\tilde{\varpi}_t = \frac{E_t(R_{t,t+1} - R_{t,t+1}^f)}{\lambda V_t(R_{t,t+1})}$ . To abstract from complications induced by expected stock return predictability, and thus to solely focus on volatility timing, we set  $R_{t,t+1}^f$  equal to 0.06 (converted to a daily value by dividing by 365) and  $E_t(R_{t,t+1})$  equal to the unconditional average of the daily SBC returns over the forecasting horizon. As noted, for each  $t$  the conditional variance  $V_t(R_{t,t+1})$  is computed as an out-of-sample forecast from an ARFIMA(2, $d$ ,2) model. The absolute risk aversion parameter  $\lambda$  is set to three conventional values, namely, 2, 7, and 10.

All variance measures are computed over a 6.5-hour period. We use the bias-correction procedure in Fleming, Kirby, and Ostdiek (2001, 2003) and multiply the estimates by a factor that is defined as

$$\zeta = \frac{\frac{1}{n} \sum_{t=1}^n R_{t,t+1}^2}{\frac{1}{n} \sum_{t=1}^n \widehat{V}_{t,t+1}}, \quad (25)$$

where  $\widehat{V}_{t,t+1}$  is the realized variance estimate prevailing between  $t$  and  $t + 1$ . This method guarantees that the average of the bias-corrected daily realized variances coincides with the variance of the daily returns. To compensate for the lack of high frequency overnight returns, we consider two procedures. The first procedure adds the square of each overnight return to each daily variance estimate. The second procedure simply relies on the correction provided by  $\zeta$ . In this case, in fact, the multiplicative factor  $\zeta$  is larger than in the case with overnight returns since the average in the denominator of the ratio is smaller. Below we show that the two procedures deliver similar results. For simplicity, we refer to the first procedure as “With overnights” and to the second procedure as “Without overnights.”

Call  $\tilde{\varpi}_t$  the optimal (time- $t$ ) allocation determined by a certain variance forecast. We wish to evaluate the economic significance of  $\{\tilde{\varpi}_t : t = 1, \dots, m\}$ . We use the investor’s long-run utility as our relevant economic metric, that is,

$$\widetilde{AU} = \frac{1}{m} \sum_{t=1}^m \left( \widetilde{R}_{t,t+1}^p \right) - \frac{\lambda}{2} \frac{1}{m} \sum_{t=1}^m \left( \widetilde{R}_{t,t+1}^p - \widetilde{R}^p \right)^2 \quad (26)$$

with

$$\widetilde{R}_{t,t+1}^p = R_{t,t+1}^f + \widetilde{\omega}_t (R_{t,t+1} - R_{t,t+1}^f) \quad (27)$$

and  $\widetilde{R}^p = \frac{1}{m} \sum_{t=1}^m \widetilde{R}_{t,t+1}^p$ . Specifically, following Fleming, Kirby, and Ostdiek (2001, 2003), we interpret the difference between  $\widetilde{AU}$  computed on the basis of our optimally sampled realized variances and  $\widetilde{AU}$  computed on the basis of an alternative variance estimate as the maximum daily return that the investor would sacrifice to use optimally sampled variance estimates.

One should keep in mind that the mean portfolio return for a given trading strategy might be imprecisely estimated. This imprecision can translate into noisy estimates of the utility gains (see Fleming, Kirby, and Ostdiek, 2003, for further discussions of this issue). Nevertheless, the consistency of our results (as reported below) across realized variance estimates and across methods of dealing with the overnight returns is an important sign of robustness of our findings.

Tables 7 and 8 contain the results corresponding to procedure “With overnights” and procedure “Without overnights,” respectively. The aforementioned “maximum return” is reported in the tables as an annualized fee. Since the main differences between variance estimates occur in the first part of the sample, i.e., before the beginning of 1997, we split the forecasting horizon in half, where the first half ends on June 30th, 1998.

The procedure “With overnights” indicates that an investor would forego between 27 basis points (when  $\lambda = 10$ ) and 96 basis points (when  $\lambda = 2$ ) per year to use our optimally sampled realized variance versus realized variance relying on five-minute intervals. The same investor would pay twice as much to use optimally sampled realized variance rather than realized variance relying on 15-minute intervals. The highest fees would be paid during the first part of the sample. This result is hardly surprisingly when comparing optimally sampled realized variance to variance sampled every five minutes in that the latter contains a large noise components and is severely biased in the first half of the sample (see Fig. 17). The procedure “Without overnights” confirms these findings. Specifically, an investor would sacrifice between 41 basis points (when  $\lambda = 10$ ) and 211 basis points (when  $\lambda = 2$ ) per year to use our optimally sampled realized variance versus realized variance relying on five-minute intervals. As before, the same investor would pay even more to switch from

suboptimal variance estimates based on 15-minute intervals to our optimal values.

It is also interesting to compare optimal sampling to a methodology that has been shown to be consistent, namely, the subsampling methodology of Zhang, Mykland, and Aït-Sahalia (2004). We use the optimal method to select the subsamples in Zhang, Mykland, and Aït-Sahalia (2004) - Eq. (39) - page 13. In the case of our procedure “With overnights,” we find that the investor would give up between 39 basis points (when  $\lambda = 10$ ) and 148 basis points (when  $\lambda = 2$ ) per year to use our optimally sampled realized variance versus realized variance based on subsampling. In the case of our procedure “Without overnights,” we find that the investor would be willing to forego between 50 basis points (when  $\lambda = 10$ ) and 264 basis points (when  $\lambda = 2$ ) per year to use our optimally sampled realized variance versus realized variance based on subsampling. We attribute this result to the fact that Zhang, Mykland, and Aït-Sahalia’s optimal method to select the subsamples produces little averaging across subsampled variance estimates in our sample. Little averaging may be associated with a (bias-corrected) realized variance estimator whose variance is not minimized in our sample. Furthermore, the bias-correction that the estimator requires yields negative variance estimates for 10% of the days. To circumvent this problem, we replace these values with the most recent nonnegative values. The final time series of daily estimates results in less persistence than for the other series. The long-memory parameter estimate is 0.27 in this case. We believe the subsampling approach should work well in large samples. Our results suggest the need for further research on implementing this technique on data such as those used in our work.

Finite sample performance is always the appropriate empirical benchmark. We show that the methods in Bandi and Russell (2004) and the present paper perform well along this dimension.<sup>5</sup>

## 9 Conclusions

Since the early work on nonparametric variance estimation of French, Schwert, and Stambaugh (1987), Schwert (1989, 1990a,b), and Schwert and Seguin (1991), among others, substantial attention has been devoted to model-free measurements of variance through sample averages of continuously compounded return data. In particular, the recent contributions of Andersen, Bollerslev, Diebold, and Ebens (2001), Andersen, Bollerslev, Diebold, and Labys (2003), and Barndorff-Nielsen and

---

<sup>5</sup>SBC is chosen as a moderately liquid representative stock. We compare optimally sampled realized variance estimates to the same alternative estimates in the case of two other stocks, one very liquid, Merrill Lynch (MER), and one relatively illiquid, XOM. Merrill Lynch is used as a substitute for GS (previously discussed) since it has the same liquidity and, unlike GS, has data going back to 1993. The results for both these stocks are similar to the SBC results. However, the results for the most liquid stock, namely MER, appear less strong.

Shephard (2002, 2004) provide theoretical justifications for using sums of high frequency return data to consistently estimate the integrated variance of the efficient price process.

Building on recent work by Bandi and Russell (2004), this paper pushes the use of high frequency data a step forward and argues that accurately sampled high frequency data can provide valuable information about *both* the conventional object of interest, namely, the integrated variance of the efficient price process, and the variance of the market microstructure frictions generated by the trading process. In keeping with the simplicity and empirical appeal of the early work, and with recent developments as indicated above, we employ (possibly rescaled) sample averages of return data. However, we argue that observations sampled at high frequencies provide consistent estimates of the variance (and higher-order moments) of the noise process, whereas appropriately chosen low frequencies allow us to optimally balance the bias and variance of the conventional realized variance estimator for the purpose of estimating the efficient price integrated variance. Our cross-sectional results, which are based on a sample of S&P100 stock midquotes for the month of February 2002, can be summarized as follows:

- (1) The standard deviations of the unobserved midquote noise components are positively related to the quoted spreads. Furthermore, the median noise standard deviation is about a quarter of the median spread. Since most trades occur within the spread and the midpoints contain a residual noise component, both results are economically meaningful.
- (2) When choosing a single optimal sampling frequency to estimate the efficient price variance, we find optimal frequencies that are smaller than frequencies generally employed in empirical work on nonparametric variance estimation through realized variance. While in our sample the five-minute frequency can be a reasonable approximation in practise, in that the associated MSE loss is small on average, the often-conjectured 15-minute interval might render the variance estimates excessively volatile.
- (3) The cross-sectional relation between estimated noise variance and efficient price variance in our sample is positive and significant. This result is consistent with both the “operating cost” theory and the “asymmetric information” theory of execution cost determination.

Our time-series results rely on a sample of midquotes of the representative stock SBC for the period between January 1993 and December 2003. These results can be summarized as follows:

- (1) When allowing for time-varying optimal frequencies, we show the need for lower frequencies at the beginning of our sample. This finding is likely due to lower market liquidity and larger microstructure frictions in the past.
- (2) Failing to account for time variation in the optimal frequencies might result in either biased or excessively volatile variance forecasts. Variations in the noise properties translate into variations in the bias when short, fixed intervals are used. Alternatively, if long fixed intervals are used, the bias will be small but the estimates can be excessively volatile.
- (3) The economic benefit of optimal sampling can be substantial, as we show in the context of an asset allocation problem à la Fleming, Kirby, and Ostdiek (2001, 2003). Specifically, we find that a risk-averse investor is willing to pay between 25 and 300 basis points per year to employ variance forecasts based on optimal intervals versus variance forecasts based on fixed intervals. While these magnitudes depend on SBC, this result provides useful guidance about the practical benefits of our procedure. The examination of the economic benefit of optimal sampling in the presence of a large number of stocks is left for future research.

The statistical and economic importance of optimal sampling suggests numerous directions for future research. First, the natural next step is to study the dynamic properties of optimally sampled variance estimates. Second, the optimal sampling methods in Bandi and Russell (2004) and the present work can be extended to realized covariances and realized betas. We expect these extensions to provide a rich set of tools to nonparametrically estimate fundamental ingredients in the theory and practise of asset pricing, portfolio choice, and risk management. Turning to the other unobserved component of the observed return variance, i.e., the noise variance, our methods render it observable in practise. In light of the relation between noise standard deviations and transaction costs, when applied to transaction prices our procedures can be utilized to study the cross-sectional and dynamic features of central measures of market quality, namely, execution costs. Research on these subjects is being conducted by the authors and will be reported in later work.

## References

- [1] Aït-Sahalia, Y., Mykland, P., 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.
- [2] Alizadeh, S., Brandt, M.W., Diebold, F.X., 2002. Range-based estimation of stochastic volatility models. *Journal of Finance* 57, 1047-1092.
- [3] Andersen, T.G., Bollerslev, T., 1997a. Heterogeneous information arrivals and return volatility dynamics: Uncovering the long-run in high frequency returns. *Journal of Finance* 52, 975-1005.
- [4] Andersen, T.G., Bollerslev, T., 1997b. Intraday periodicity and volatility persistence in financial markets. *Journal of Empirical Finance* 4, 115-158.
- [5] Andersen, T.G., Bollerslev, T., 1998. Deutsche market volatility: Intraday activity patterns, macroeconomic announcements, and longer run dependencies. *Journal of Finance* 53, 219-265.
- [6] Andersen, T.G., Bollerslev, T., Diebold, F.X., 2002. Parametric and nonparametric measurements of volatility. In: Aït-Sahalia, Y. and Hansen, L.P. (Eds.), *Handbook of Financial Econometrics*. North Holland, Amsterdam. Forthcoming.
- [7] 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.
- [8] Andersen, T.G., Bollerslev, T., Diebold, F.X., Labys, P. (2000). Great realizations. *Risk Magazine* 71, 105-108.
- [9] Andersen, T.G., Bollerslev, T., Diebold, F.X., Labys, P. (2003). Modeling and forecasting realized volatility. *Econometrica* 71, 579-625.
- [10] Bai, X., Russell, J.R., Tiao, G., 2004. Effects of non-normality and dependence on the precision of variance estimates using high-frequency data. Unpublished working paper. University of Chicago.
- [11] Baillie, R., Bollerslev, T., Mikkelsen, H.O., 1996. Fractionally integrated generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* 74, 3-30.
- [12] Bandi, F.M., Perron, B., 2004. Long memory and the relation between realized and implied volatility. Unpublished working paper. University of Chicago and University of Montreal.

- [13] Bandi, F. M., Russell, J.R., 2004. Microstructure noise, realized variance, and optimal sampling. Unpublished working paper. University of Chicago.
- [14] Barndorff-Nielsen, O. E., Shephard, N., 2002. Econometric analysis of realized volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society, Series B*, 64, 253-280.
- [15] Barndorff-Nielsen, O. E., Shephard, N., 2004. Econometric analysis of realized covariation: High-frequency based covariance, regression and correlation in financial economics. *Econometrica* 72, 885-925.
- [16] Bates, D.S., 2000. Post-'87 crash fears in the S&P 500 futures option market. *Journal of Econometrics* 94, 181-238.
- [17] Beckers, S., 1983. Variance of security price returns based on high, low, and closing prices. *Journal of Business* 56, 97-112.
- [18] Bollerslev, T., Mikkelsen, H.O., 1996. Modeling and pricing long-memory in stock market volatility. *Journal of Econometrics* 73, 75-99.
- [19] Bollerslev, T., Mikkelsen, H.O., 1999. Long-term equity anticipation securities and stock market volatility dynamics. *Journal of Econometrics* 92, 75-99.
- [20] Brandt, M.W., Diebold, F.X., 2004. A no-arbitrage approach to range-based estimation of return covariances and correlations. *Journal of Business*, forthcoming.
- [21] Chernov, M., Gallant, A.R., Ghysels, E., Tauchen, G., 2003. Alternative models for stock price dynamics. *Journal of Econometrics* 116, 225-257.
- [22] Comte, F., Renault R., 1998. Long-memory in continuous-time stochastic volatility models. *Mathematical Finance* 8, 291-323.
- [23] Ding, Z., Granger, C.W.J., Engle, R.F., 1993. A long memory property of stock returns and a new model. *Journal of Empirical Finance* 1, 83-106.
- [24] Duffie, D., Pan, J., Singleton, K., 2000. Transform analysis and asset pricing for affine jump-diffusions. *Econometrica* 68, 1343-1376.

- [25] Easley, D., O'Hara, M., 1992. Time and the process of security price adjustment. *Journal of Finance* 47, 577-606.
- [26] Engle, R.F. Lee, G.J., 1999. A permanent and transitory component model of stock return volatility. In: Engle, R. F. and White H. (Eds.), *Cointegration, Causality, and Forecasting: A Festschrift in Honor of Clive W.J. Granger*. Oxford University Press, pp. 475-497.
- [27] Eraker, B., Johannes, M., Polson, N., 2003. The impact of jumps in volatility and returns. *Journal of Finance* 58, 1269-1300.
- [28] Fleming, J., Kirby, C., Ostdiek, B., 2001. The economic value of volatility timing. *Journal of Finance* 56, 329-352.
- [29] Fleming, J., Kirby, C., Ostdiek, B., 2003. The economic value of volatility timing using "realized volatility." *Journal of Financial Economics* 67, 473-509.
- [30] French, K.R., Schwert, G.W., Stambaugh, R.F., 1987. Expected stock returns and volatility. *Journal of Financial Economics* 19, 3-29.
- [31] Garman, M.B., Klass, M.J., 1980. On the estimation of price volatility from historical data. *Journal of Business* 53, 67-78.
- [32] Geweke, J., Porter-Hudak, S., 1983. The estimation and application of long memory time series models. *Journal of Time Series Analysis* 4, 221-238.
- [33] Hansen, P.R., Lunde A., 2004a. An unbiased measure of realized variance. Unpublished working paper. Stanford University and University of Aarhus.
- [34] Hansen, P.R., Lunde A., 2004b. Realized variance and market microstructure noise. Unpublished working paper. Stanford University and University of Aarhus.
- [35] Jones, C.S., 2003. The dynamics of stochastic volatility: Evidence from underlying and options markets. *Journal of Econometrics* 116, 181-224.
- [36] Ohanissian, A., Russell, J.R., Tsay R., 2004. Using temporal aggregation to distinguish between true and spurious long memory. Unpublished working paper. University of Chicago.
- [37] Madhavan, A., 2000. Market microstructure: A survey. *Journal of Financial Markets* 3, 205-258.

- [38] Meddahi, N., 2001. An eigenfunction approach for volatility modelling. Unpublished working paper. University of Montreal.
- [39] Oomen, R., 2004a. Properties of realized variance for pure-jump processes: calendar time sampling versus business time sampling. Unpublished working paper. Warwick Business School.
- [40] Oomen, R., 2004b. Properties of bias-corrected realized variance in calendar time and in business time. Unpublished working paper. Warwick Business School.
- [41] Pan, J., 2002. The jump-risk premia implicit in options: Evidence from an integrated time-series study. *Journal of Financial Economics* 63, 3-50.
- [42] Parkinson, M., 1980. The extreme value method for estimating the variance of the rate of return. *Journal of Business* 53, 61-65.
- [43] Schwert, G.W., 1989. Why does stock market volatility change over time? *Journal of Finance* 44, 1115-1153.
- [44] Schwert, G.W., 1990a. Stock market volatility. *Financial Analysts Journal* 46, 23-24.
- [45] Schwert, G.W., 1990b. Stock volatility and the crash of '87. *Review of Financial Studies* 3, 77-102.
- [46] Schwert, G.W., Seguin, P.J., 1991. Heteroskedasticity in stock returns. *Journal of Finance* 45, 1129-1155.
- [47] Stoll, H.R., 2000. Friction. *Journal of Finance* 55, 1479-1514.
- [48] Zhang, L., Mykland P., Ait-Sahalia, Y., 2004. A tale of two time scales: determining integrated volatility with noisy high-frequency data. *Journal of the American Statistical Association*, forthcoming.
- [49] Zhou, B., 1996. High-frequency data and volatility in foreign exchange rates. *Journal of Business and Economic Statistics* 14, 45-52.

**Table 1**  
Descriptive statistics for the S&P100 stocks<sup>a</sup>

Symbol	Avg. Dur.	Avg. Sprd.	Avg. Price	Mid. Var.	Mid. 4 <sup>th</sup>	D <sup>a</sup>	D*	V*
AA	16.19	0.00206	36.04	6.71E-07	1.36E-11	4.71	5.2	0.00044
AEP	23.57	0.00183	41.71	4.02E-07	5.63E-12	5.06	5.8	0.00023
AES	28.80	0.00421	8.27	7.46E-06	3.07E-10	2.11	2.2	0.01148
AIG	13.46	0.00148	72.82	1.71E-07	3.59E-13	2.11	2.2	0.00036
ALL	18.78	0.00248	33.95	3.74E-07	1.16E-12	2.98	3	0.00043
AMGN <sup>*</sup>	2.71	0.00053	57.82	3.07E-08	4.83E-15	0.62	0.6	0.00043
AOL	9.31	0.00290	25.24	1.38E-06	7.56E-12	3.73	3.8	0.00105
ATI	60.96	0.00308	15.64	1.08E-06	1.21E-11	5.67	6	0.00053
AVP	24.06	0.00161	49.07	2.05E-07	1.33E-12	3.87	4.2	0.00020
AXP	12.83	0.00202	34.13	4.63E-07	1.87E-12	3.30	3.4	0.00051
BA	14.04	0.00190	43.52	2.81E-07	7.91E-13	2.60	2.6	0.00037
BAC	8.13	0.00150	61.15	1.50E-07	8.88E-13	2.25	2.4	0.00030
BAX	22.30	0.00158	55.36	3.04E-07	4.93E-12	4.29	5	0.00022
BCC	29.42	0.00259	34.96	4.37E-07	3.30E-12	4.08	4.4	0.00034
BDK	20.96	0.00194	43.13	2.58E-07	7.58E-13	2.54	2.6	0.00032
BHI	17.58	0.00251	34.51	4.04E-07	1.25E-12	2.70	2.8	0.00073
BMJ	10.51	0.00201	45.11	3.02E-07	1.09E-12	3.06	3.2	0.00036
BNI	26.44	0.00261	27.90	5.36E-07	2.86E-12	4.87	5	0.00033
BUD	25.99	0.00149	48.70	2.49E-07	1.85E-12	5.44	6.2	0.00015
C	8.00	0.00194	44.48	3.44E-07	8.45E-13	2.60	2.6	0.00055
CCU	15.48	0.00202	47.11	3.14E-07	1.54E-12	2.38	2.4	0.00060
CI	17.57	0.00133	92.69	1.58E-07	1.02E-12	1.10	1.2	0.00046
CL	15.28	0.00159	55.66	1.41E-07	1.31E-12	2.33	2.6	0.00020
CPB	27.09	0.00274	27.00	1.11E-06	1.51E-11	8.23	8.8	0.00038
CSC	18.46	0.00236	47.18	4.27E-07	2.44E-12	2.99	3	0.00067
CSCO <sup>*</sup>	4.63	0.00086	16.85	1.05E-07	4.45E-14	0.82	0.8	0.00105
DAL	19.34	0.00238	32.76	3.73E-07	1.29E-12	2.68	2.8	0.00058
DD	14.13	0.00170	45.04	2.74E-07	8.05E-13	3.05	3.2	0.00034
DIS	10.18	0.00276	23.26	9.45E-07	3.98E-12	5.08	5.2	0.00055
DOW	21.45	0.00251	30.03	6.93E-07	3.27E-12	4.43	4.6	0.00040
EK	20.32	0.00253	29.17	6.47E-07	2.83E-12	3.72	3.8	0.00051
EMC	11.86	0.00286	13.49	9.87E-07	7.25E-12	2.49	2.6	0.00156
EP	21.25	0.00253	37.04	5.01E-07	2.51E-12	3.03	3	0.00070
ETR	31.37	0.00174	41.05	2.04E-07	1.41E-12	3.38	3.6	0.00021
EXC	23.66	0.00165	49.94	2.81E-07	2.44E-12	4.15	4.6	0.00031
F	19.92	0.00241	14.69	2.28E-06	2.02E-11	10.32	10.6	0.00041
FDX	12.97	0.00181	54.83	2.48E-07	1.12E-12	3.51	3.6	0.00030
G	21.44	0.00247	33.20	3.80E-07	1.43E-12	4.14	4.2	0.00031
GD	20.63	0.00127	89.12	1.87E-07	1.94E-12	2.64	3	0.00023
GE	5.79	0.00182	37.42	5.20E-07	1.10E-12	3.45	3.4	0.00050
GM	19.02	0.00135	51.81	2.35E-07	1.89E-12	2.83	3	0.00022
GS	10.41	0.00139	82.40	1.74E-07	1.25E-12	1.98	2.2	0.00042
HAL	16.25	0.00288	15.21	2.02E-06	2.11E-11	2.51	2.6	0.00153
HCA	20.04	0.00199	42.25	3.63E-07	2.29E-12	2.25	2.4	0.00040
HD	10.69	0.00142	50.37	2.54E-07	3.11E-12	2.98	3.4	0.00031
HET	25.76	0.00273	38.39	6.12E-07	3.96E-12	3.28	3.4	0.00049
HIG	19.86	0.00152	65.93	1.98E-07	7.34E-13	2.77	2.8	0.00031
HNZ	18.67	0.00219	40.93	2.75E-07	1.07E-12	4.10	4.2	0.00020
HON	12.94	0.00253	34.34	6.44E-07	3.98E-12	1.48	1.4	0.00104
HWP	16.91	0.00217	20.60	1.13E-06	1.31E-11	5.03	5.2	0.00063
IBM	7.85	0.00106	103.05	1.77E-07	1.99E-12	2.38	2.6	0.00029
INTC <sup>*</sup>	2.62	0.00053	31.89	3.90E-08	3.84E-15	0.51	0.6	0.00073
IP	15.02	0.00167	42.93	2.85E-07	1.25E-12	3.49	3.6	0.00033
JNJ	16.33	0.00134	57.91	1.88E-07	6.70E-13	3.45	3.6	0.00021
JPM	10.23	0.00254	29.76	1.10E-06	1.16E-11	3.50	3.6	0.00114

KO	16.10	0.00162	46.33	2.65E-07	2.82E-12	4.69	5.4	0.00024
LEH	12.07	0.00179	59.25	2.83E-07	3.31E-12	2.05	2.2	0.00062
LTD	31.81	0.00225	17.69	1.19E-06	1.01E-11	4.50	4.6	0.00059
LU	21.11	0.00403	5.71	5.18E-06	1.66E-10	8.88	9.2	0.00116
MAY	24.02	0.00238	35.56	3.19E-07	1.82E-12	3.22	3.4	0.00034
MCD	14.55	0.00268	26.81	5.62E-07	1.78E-12	6.96	7.2	0.00023
MDT	13.81	0.00169	46.96	2.92E-07	1.87E-12	3.34	3.6	0.00034
MEDI*	4.21	0.00121	40.90	1.10E-07	8.70E-14	0.66	0.6	0.00102
MER	9.00	0.00194	47.72	3.58E-07	3.63E-12	2.75	3	0.00067
MMM	13.28	0.00104	114.76	3.54E-07	9.98E-12	4.00	5	0.00025
MO	13.27	0.00157	51.20	1.92E-07	2.41E-13	5.09	5.2	0.00012
MRK	13.79	0.00134	59.96	2.19E-07	6.10E-13	3.69	3.8	0.00019
MSFT*	1.96	0.00035	60.19	1.74E-08	1.14E-15	0.47	0.4	0.00043
MWD	9.90	0.00171	49.87	2.93E-07	3.29E-12	2.15	2.2	0.00059
NSC	25.73	0.00263	21.76	9.40E-07	7.43E-12	5.05	5.2	0.00055
NSM	14.44	0.00306	26.63	1.05E-06	8.84E-12	4.36	4.4	0.00103
NXTL*	7.49	0.00323	5.23	1.39E-06	4.27E-12	0.41	0.4	0.01272
ONE	16.05	0.00218	35.65	3.93E-07	2.05E-12	3.28	3.4	0.00047
ORCL*	5.52	0.00097	16.00	1.19E-07	2.52E-14	0.97	1	0.00095
PEP	12.72	0.00149	49.68	3.42E-07	2.85E-12	4.71	5.2	0.00023
PFE	7.43	0.00190	41.05	2.76E-07	4.07E-13	4.81	4.8	0.00024
PG	10.36	0.00115	83.91	9.70E-08	3.38E-13	2.79	3	0.00019
ROK	42.41	0.00299	18.73	1.41E-06	2.42E-11	5.88	6.2	0.00071
RSH	27.96	0.00247	27.74	1.13E-06	1.50E-11	4.03	4.2	0.00078
RTN	23.66	0.00211	37.91	4.57E-07	1.85E-12	3.82	4	0.00032
S	15.60	0.00169	52.78	2.05E-07	1.11E-12	3.71	4	0.00025
SBC	11.55	0.00205	36.52	3.79E-07	1.33E-12	3.42	3.4	0.00041
SLB	11.31	0.00163	55.80	2.34E-07	1.51E-12	2.32	2.4	0.00046
SLE	27.75	0.00247	21.34	1.38E-06	1.94E-11	12.66	13.8	0.00021
SO	24.60	0.00270	24.99	7.09E-07	2.41E-12	11.38	11.8	0.00016
T	19.34	0.00256	15.60	2.06E-06	2.20E-11	5.54	5.6	0.00082
TOY	40.21	0.00246	17.64	1.21E-06	1.05E-11	5.07	5.2	0.00079
TXN	7.89	0.00278	30.53	5.94E-07	2.47E-12	3.01	3	0.00091
TYC	6.95	0.00288	29.32	1.41E-06	1.71E-11	1.22	1.2	0.00382
UIS	39.75	0.00304	11.69	2.12E-06	2.58E-11	9.87	10.2	0.00048
USB	24.45	0.00249	19.95	1.24E-06	8.01E-12	8.23	8.4	0.00039
UTX	15.75	0.00130	69.45	1.97E-07	1.11E-12	2.57	2.8	0.00024
VIAB	12.36	0.00237	42.26	3.04E-07	2.46E-12	2.05	2.2	0.00065
VZ	9.86	0.00188	45.85	2.93E-07	7.69E-13	3.52	3.6	0.00031
WFC	12.68	0.00161	46.14	1.82E-07	2.82E-13	4.28	4.4	0.00019
WMB	16.66	0.00319	16.03	2.91E-06	4.14E-11	3.23	3.2	0.00251
WMT	11.00	0.00108	60.01	1.33E-07	3.05E-13	2.79	2.8	0.00023
WY	17.34	0.00170	59.66	2.46E-07	1.68E-12	2.80	3	0.00033
XOM	8.91	0.00191	39.38	4.33E-07	6.75E-13	6.47	6.6	0.00018
XRX	38.93	0.00305	10.11	3.74E-06	6.60E-11	8.17	8.4	0.00088

<sup>a</sup>The sample covers the month of February 2002. For NYSE stocks we use quotes posted on two exchanges, the NYSE and the MIDWEST. Nasdaq stocks are denoted by an asterisk. The data come from the TAQ data set. The table contains average durations in seconds (Avg. Dur.), average differences between logarithmic dollar ask prices and logarithmic dollar bid prices (Avg. Sprd.), average prices in dollar values (Avg. Price), estimated variances of the return noise components from Proposition 1b in the main text (Mid. Var.), estimated fourth moments of the return noise components from Proposition 2b in the main text (Mid. 4<sup>th</sup>), estimated approximate optimal sampling intervals in minutes from Proposition 4 in the main text (D<sup>a</sup>), estimated true optimal sampling intervals in minutes from Proposition 3 in the main text (D\*), and average, optimally sampled, daily realized variances (V\*). The variances of the noise components, the fourth moments of the noise components, and the realized variance estimates are based on continuously compounded returns constructed using midpoint bid-ask quotes. The second and fourth moments of the noise components are sample averages of second and fourth powers of continuously compounded returns calculated at the highest frequency. The optimally sampled realized variances are computed by summing squared continuously compounded returns sampled every D\* minutes.

**Table 2**

Outcome of a regression of the standard deviations of the noise components of the S&P100 stocks on the corresponding average bid-ask spreads<sup>a</sup>

	Coefficients	Std. Errors	T-statistics	P-values
<i>Intercept</i>	-0.4488	0.3627	-1.237	0.2189
<i>Avg. Sprd.</i>	1.1027	0.0578	19.061	0.0000
	$R^2=78.75\%$	$adjR^2=78.54\%$		

<sup>a</sup>The table contains the result of a regression of the logarithmic standard deviations of the noise components (computed as in Proposition 1b in the main text) of the midquotes of the S&P100 stocks on the corresponding logarithmic average bid-ask spreads (Avg. Sprd.). The standard deviations of the noise components are the square root of half the sample second moments of the quote-to-quote continuously compounded returns. The average bid-ask spreads are the average differences between logarithmic dollar ask prices and logarithmic dollar bid prices. The sample of S&P100 stocks covers the month of February 2002. For NYSE stocks we use quotes posted on two exchanges, the NYSE and the MIDWEST. The data come from the TAQ database.

**Table 3**

Outcome of a regression of the standard deviations of the noise components of the S&P100 stocks on the corresponding (average) daily realized volatilities<sup>a</sup>

	Coefficients	Std. Errors	T-statistics	P-values
<i>Intercept</i>	0.000333	0.0000649	5.129	0.000
$\sqrt{V^*}$	0.016915	0.002339	7.230	0.000
	$R^2=34.79\%$	$adjR^2=34.12\%$		

<sup>a</sup>The table contains the result of a regression of the standard deviations of the noise components (computed as in Proposition 1b in the main text) of the midquotes of the S&P100 stocks on the corresponding square roots of the (average) daily realized variances ( $\sqrt{V^*}$ ). The standard deviations of the noise components are the square root of half the sample second moments of the quote-to-quote continuously compounded returns. The realized variances are optimally sampled, i.e., they are computed by summing squared continuously compounded returns sampled every  $D^*$  minutes, where  $D^*$  is the optimal sampling interval from Proposition 3 in the main text. The sample of S&P100 stocks covers the month of February 2002. For NYSE stocks we use quotes posted on two exchanges, the NYSE and the MIDWEST. The data come from the TAQ database.

**Table 4**

One-day-ahead predictive regression of daily optimally sampled realized variance using optimally sampled realized variance, five-minute realized variance, and 15-minute realized variance for SBC<sup>a</sup>

	Coefficients	Std. Errors	T-statistics	P-values
<i>Intercept</i>	-3.4e-06	6.83e-06	-0.498	0.618
<i>V*</i>	0.881	0.079	11.121	0.000
<i>5mV</i>	0.172	0.091	1.883	0.059
<i>15mV</i>	-0.036	0.070	-0.510	0.609
	$R^2=61.3\%$	$adjR^2=61.29\%$		

<sup>a</sup>The table contains the results of a regression of daily optimally sampled realized variances on one-day-ahead forecasts constructed using daily optimally sampled realized variances ( $V^*$ ), daily five-minute realized variances (5mV), and daily 15-minute realized variances (15mV). We use quote data for SBC communications (SBC) over the period between January 1993 and December 2003. The data come from the TAQ database. The optimally sampled realized variances are computed by summing squared continuously compounded returns constructed using midpoint bid-ask quotes sampled every  $D^*$  minutes, where  $D^*$  is the optimal sampling interval from Proposition 3 in the main text. The five-minute realized variances are computed by summing squared continuously compounded returns constructed by sampling every five minutes. The 15-minute realized variances are computed by summing squared continuously compounded returns constructed by sampling every 15 minutes. The one-day-ahead forecasts are obtained by virtue of an ARFIMA(2, $d$ ,2) model. We use GPH estimates of the  $d$  parameter. The estimated  $d$  values are equal to 0.45 in the case of the optimally sampled realized variances, 0.44 in the case of the five-minute realized variances, and 0.37 in the case of the 15-minute realized variances. We use 200 observations to construct the first forecast. The total number of one-day-ahead forecasts is 1,875.

**Table 5**

One-day-ahead predictive regression of daily five-minute realized variance using optimally sampled realized variance, five-minute realized variance, and 15-minute realized variance for SBC<sup>a</sup>

	Coefficients	Std. Errors	T-statistics	P-values
<i>Intercept</i>	2.48e-05	7.43e-06	3.334	0.000
<i>V*</i>	0.404	0.086	4.697	0.000
<i>5mV</i>	0.590	0.099	5.920	0.000
<i>15mV</i>	-0.001	0.076	-0.021	0.982
	$R^2=55.29\%$	$adjR^2=55.24\%$		

<sup>a</sup>The table contains the results of a regression of daily five-minute realized variances on one-day-ahead forecasts constructed using daily optimally sampled realized variances ( $V^*$ ), daily five-minute realized variances (5mV), and daily 15-minute realized variances (15mV). We use quote data for SBC communications (SBC) over the period between January 1993 and December 2003. The data come from the TAQ database. The optimally sampled realized variances are computed by summing squared continuously compounded returns constructed using midpoint bid-ask quotes sampled every  $D^*$  minutes, where  $D^*$  is the optimal sampling interval from Proposition 3 in the main text. The five-minute realized variances are computed by summing squared continuously compounded returns constructed by sampling every five minutes. The 15-minute realized variances are computed by summing squared continuously compounded returns constructed by sampling every 15 minutes. The one-day-ahead forecasts are obtained by virtue of an ARFIMA(2, $d$ ,2) model. We use GPH estimates of the  $d$  parameter. The estimated  $d$  values are equal to 0.45 in the case of the optimally sampled realized variances, 0.44 in the case of the five-minute realized variances, and 0.37 in the case of the 15-minute realized variances. We use 200 observations to construct the first forecast. The total number of one-day-ahead forecasts is 1,875.

**Table 6**

One-day-ahead predictive regression of daily 15-minute realized variance using optimally sampled realized variance, five-minute realized variance, and 15-minute realized variance for SBC<sup>a</sup>

	Coefficients	Std. Errors	T-statistics	P-values
<i>Intercept</i>	1.18e-05	7.99e-06	1.471	0.141
<i>V*</i>	0.281	0.092	3.033	0.002
<i>5mV</i>	0.085	0.107	0.798	0.424
<i>15mV</i>	0.595	0.082	7.212	0.000
	$R^2=44.96\%$	$adjR^2=44.90\%$		

<sup>a</sup>The table contains the results of a regression of daily 15-minute realized variances on one-day-ahead forecasts constructed using daily optimally sampled realized variances ( $V^*$ ), daily five-minute realized variances (5mV), and daily 15-minute realized variances (15mV). We use quote data for SBC communications (SBC) over the period between January 1993 and December 2003. The data come from the TAQ database. The optimally sampled realized variances are computed by summing squared continuously compounded returns constructed using midpoint bid-ask quotes sampled every  $D^*$  minutes, where  $D^*$  is the optimal sampling interval from Proposition 3 in the main text. The five-minute realized variances are computed by summing squared continuously compounded returns constructed by sampling every five minutes. The 15-minute realized variances are computed by summing squared continuously-compounded returns constructed by sampling every 15 minutes. The one-day-ahead forecasts are obtained by virtue of an ARFIMA(2, $d$ ,2) model. We use GPH estimates of the  $d$  parameter. The estimated  $d$  values are equal to 0.45 in the case of the optimally sampled realized variances, 0.44 in the case of the five-minute realized variances, and 0.37 in the case of the 15-minute realized variances. We use 200 observations to construct the first forecast. The total number of one-day-ahead forecasts is 1,875.

**Table 7**

Annualized percentage fees (in basis points) that a mean-variance investor would be willing to pay to perform a volatility timing trading strategy using optimally sampled realized variances versus five-minute variances, 15-minute variances, and variances obtained by subsampling (“With overnights”)<sup>a</sup>

	$\lambda = 2$			$\lambda = 7$			$\lambda = 10$		
	$V^*/5mV$	$V^*/15mV$	$V^*/SubV$	$V^*/5mV$	$V^*/15V$	$V^*/SubV$	$V^*/5mV$	$V^*/15mV$	$V^*/SubV$
<i>1<sup>st</sup> half</i>	104	264	189	52	106	83	37	75	59
<i>2<sup>nd</sup> half</i>	89	177	112	25	50	32	18	35	22
<i>Full</i>	96	218	148	38	76	56	27	54	39

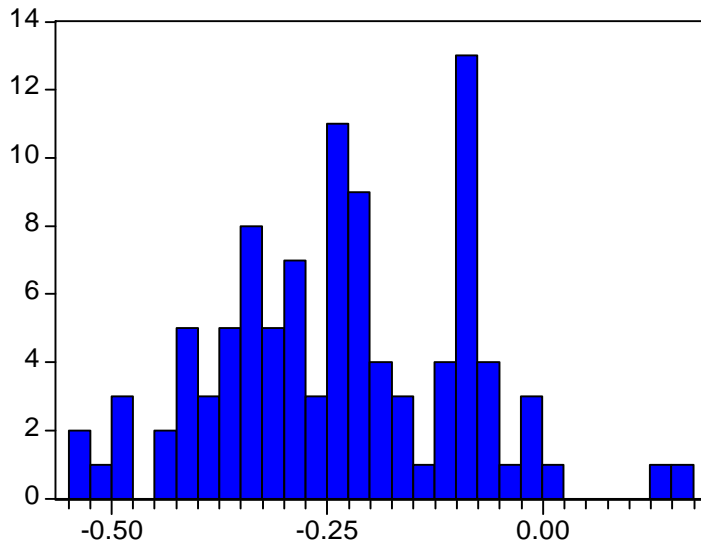
<sup>a</sup>The table contains the annualized percentage fees (in basis points) that a mean-variance investor with absolute risk-aversion parameter  $\lambda = 2, 7$ , and 10 would be willing to pay to perform a volatility timing trading strategy using optimally sampled realized variances ( $V^*$ ) versus five-minute realized variances (5mV), 15-minute realized variances (15mV), and realized variances obtained by subsampling (SubV). We use quote data for SBC communications (SBC) over the period between January 1993 and December 2003. We split the sample in half with a break date of the end of June 1998. The data come from the TAQ database. The optimally sampled realized variances are computed by summing squared continuously compounded returns constructed using midpoint bid-ask quotes sampled every  $D^*$  minutes, where  $D^*$  is the optimal sampling interval from Proposition 3 in the main text. The five-minute realized variances are computed by summing squared continuously compounded returns constructed by sampling every five minutes. The 15-minute realized variances are computed by summing squared continuously compounded returns constructed by sampling every 15 minutes. The subsampled realized variances are derived as in Zhang, Mykland, and Ait-Sahalia (2004). We follow Fleming, Kirby, and Ostdiek (2001, 2003) in performing a volatility timing trading exercise as detailed in the main text. The one-day-ahead variance forecasts are obtained by virtue of an ARFIMA(2, $d$ ,2) model. We use GPH estimates of the  $d$  parameter. The estimated  $d$  values are equal to 0.45 in the case of the optimally sampled realized variances, 0.44 in the case of the 5-minute realized variances, 0.37 in the case of the 15-minute realized variances, and 0.27 in the case of the subsampled realized variances. We employ 200 observations to construct the first forecast. The total number of one-day-ahead forecasts is equal to 1,875. We account for the fact that all variance measures are computed over a 6.5-hour period by implementing the “With overnights” procedure in the main text, namely, we add the square of the overnight returns to the daily variance estimates.

**Table 8**

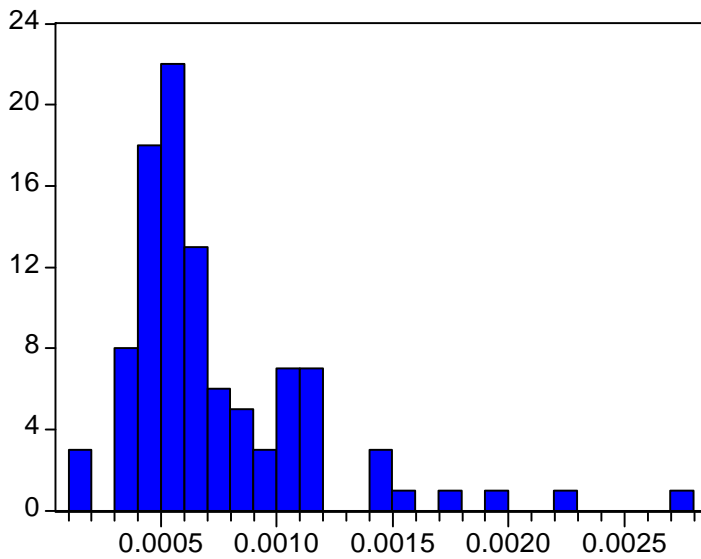
Annualized percentage fees (in basis points) that a mean-variance investor would be willing to pay to perform a volatility timing trading strategy using optimally sampled realized variances versus five-minute variances, 15-minute variances, and variances obtained by subsampling (“Without overnights”)<sup>a</sup>

	$\lambda = 2$			$\lambda = 7$			$\lambda = 10$		
	$V^*/5mV$	$V^*/15mV$	$V^*/SubV$	$V^*/5mV$	$V^*/15V$	$V^*/SubV$	$V^*/5mV$	$V^*/15mV$	$V^*/SubV$
<i>1<sup>st</sup> half</i>	363	468	231	100	125	57	70	88	40
<i>2<sup>nd</sup> half</i>	75	165	293	22	47	84	15	33	59
<i>Full</i>	211	307	264	58	84	71	41	59	50

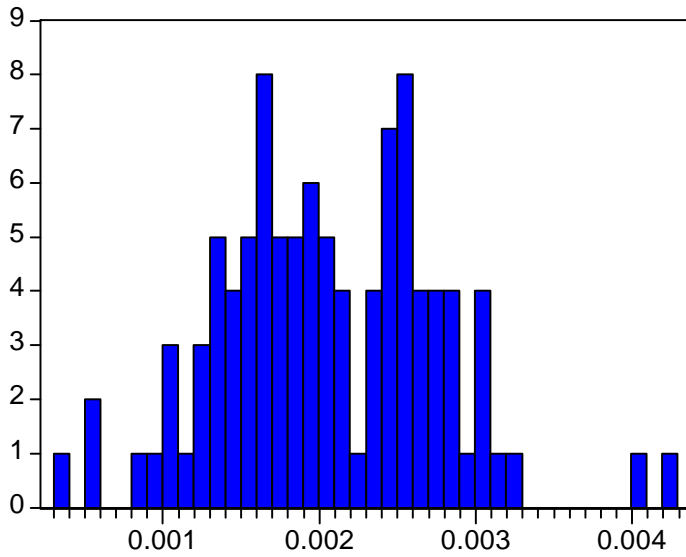
<sup>a</sup>The table contains the annualized percentage fees (in basis points) that a mean-variance investor with absolute risk-aversion parameter  $\lambda = 2, 7$ , and 10 would be willing to pay to perform a volatility timing trading strategy using optimally sampled realized variances ( $V^*$ ) versus five-minute realized variances (5mV), 15-minute realized variances (15mV), and realized variances obtained by subsampling (SubV). We use quote data for SBC communications (SBC) over the period between January 1993 and December 2003. We split the sample in half with a break date of the end of June 1998. The data come from the TAQ database. The optimally sampled realized variances are computed by summing squared continuously compounded returns constructed using midpoint bid-ask quotes sampled every  $D^*$  minutes, where  $D^*$  is the optimal sampling interval from Proposition 3 in the main text. The five-minute realized variances are computed by summing squared continuously-compounded returns constructed by sampling every five minutes. The 15-minute realized variances are computed by summing squared continuously-compounded returns constructed by sampling every 15 minutes. The subsampled realized variances are derived as in Zhang, Mykland, and Ait-Sahalia (2004). We follow Fleming, Kirby, and Ostdiek (2001, 2003) in performing a volatility timing trading exercise as detailed in the main text. The one-day-ahead variance forecasts are obtained by virtue of an ARFIMA(2, $d$ ,2) model. We use GPH estimates of the  $d$  parameter. The estimated  $d$  values are equal to 0.45 in the case of the optimally sampled realized variances, 0.44 in the case of the five-minute realized variances, 0.37 in the case of the 15-minute realized variances, and 0.27 in the case of the subsampled realized variances. We employ 200 observations to construct the first forecast. The total number of one-day-ahead forecasts is equal to 1,875. We account for the fact that all variance measures are computed over a 6.5-hour period by implementing the “Without overnights” procedure in the main text, namely, we simply multiply the daily variance estimates by a factor which guarantees that the average of the “corrected” realized variances coincides with the variance of the daily returns.



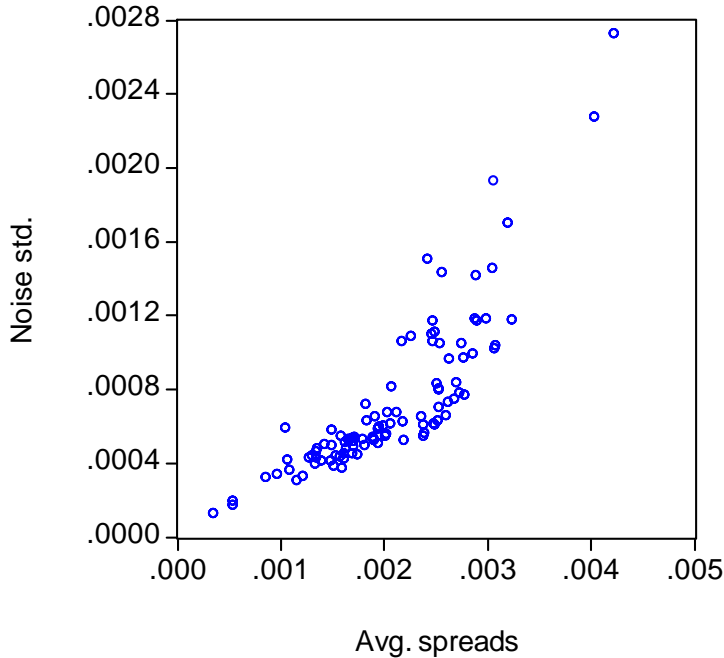
**Fig 1.** Histogram of the first-order autocorrelations of the S&P100 stock returns. The figure shows the histogram of the first-order autocorrelations of the midquote returns of the S&P100 stocks for the month of February 2002. For NYSE stocks we use quotes posted on two exchanges, the NYSE and the MIDWEST. The data come from the TAQ database.



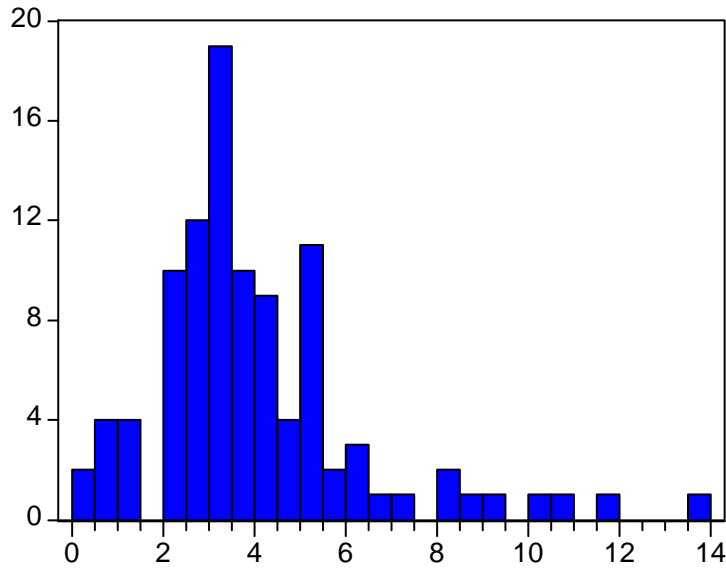
**Fig 2.** Histogram of the standard deviations of the noise components of the midquotes of the S&P100 stocks. The standard deviations of the noise components are the square root of half the sample second moment of the quote-to-quote continuously compounded returns. The sample of S&P100 stocks covers the month of February 2002. For NYSE stocks we use quotes posted on two exchanges, the NYSE and the MIDWEST. The data come from the TAQ database.



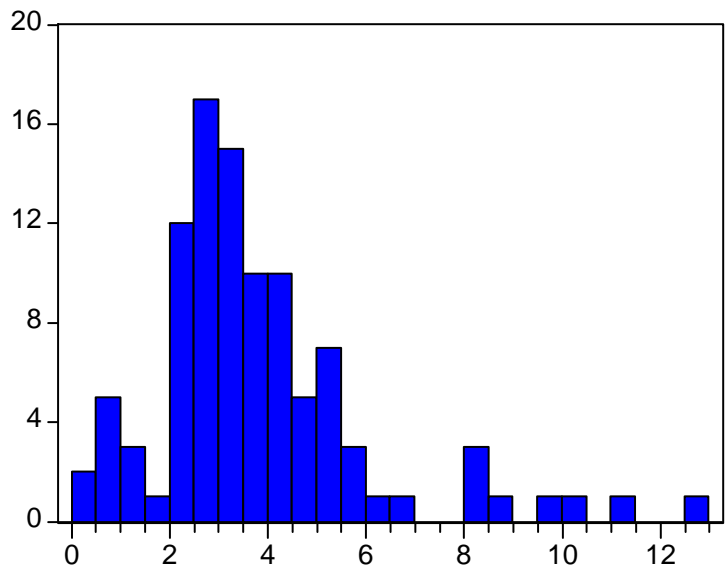
**Fig 3.** Histogram of the average bid-ask spreads of the S&P100 stocks. The average bid-ask spreads are the average differences between logarithmic dollar ask prices and logarithmic dollar bid prices. The sample of S&P100 stocks covers the month of February 2002. For NYSE stocks we use quotes posted on two exchanges, the NYSE and the MIDWEST. The data come from the TAQ database.



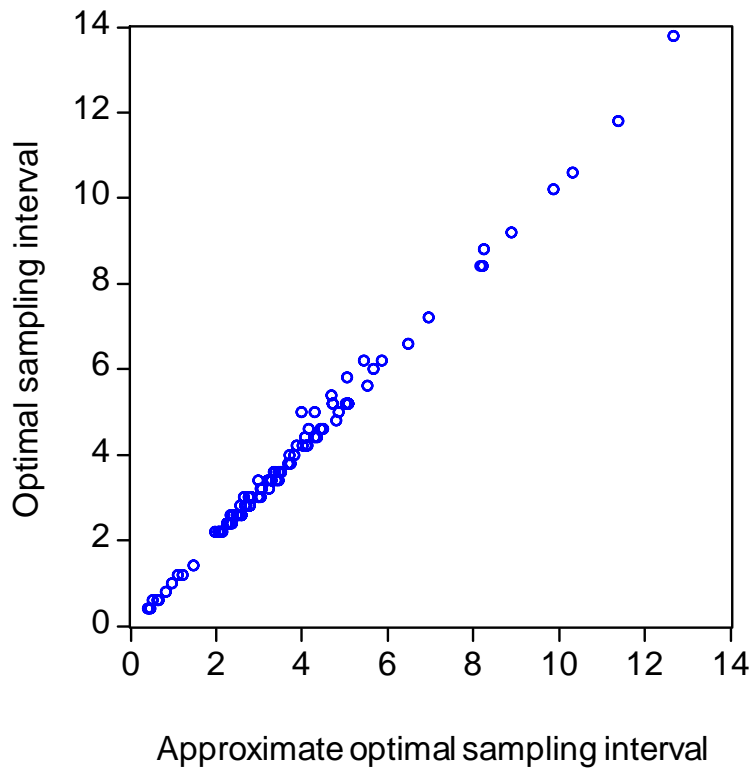
**Fig 4.** The standard deviations of the noise components of the midquotes of the S&P100 stocks versus the corresponding average bid-ask spreads. The standard deviations of the noise components are the square root of half the sample second moments of the quote-to-quote continuously compounded returns. The average bid-ask spreads are the average differences between logarithmic dollar ask prices and logarithmic dollar bid prices. The sample of S&P100 stocks covers the month of February 2002. For NYSE stocks we use quotes posted on two exchanges, the NYSE and the MIDWEST. The data come from the TAQ database.



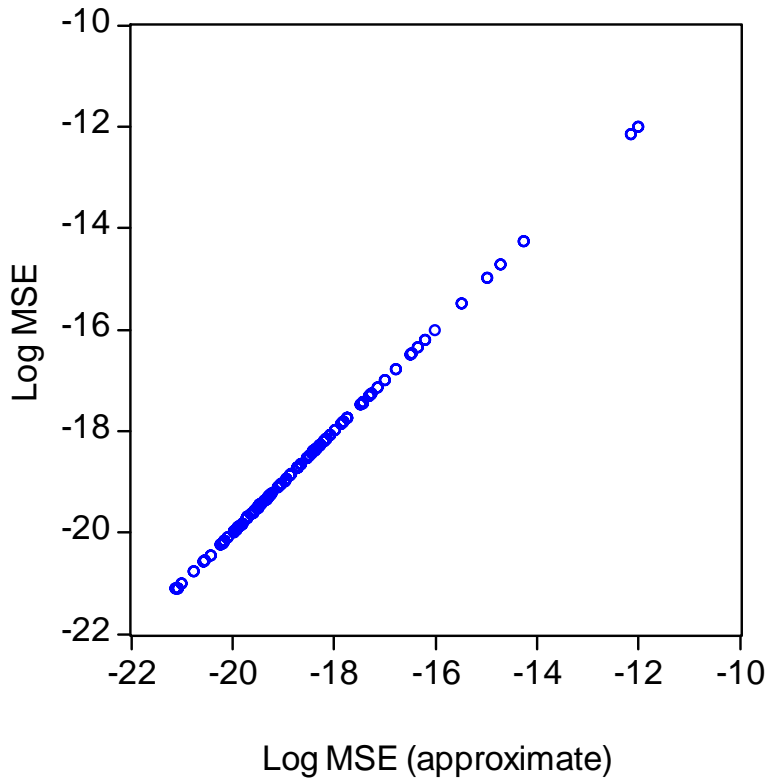
**Fig 5.** Histogram of the estimated optimal sampling intervals (in minutes) for the S&P100 stocks. The figure shows the histogram of the number of minutes that, based on Proposition 3 in the main text, should be used to construct continuously compounded returns for the purpose of realized variance estimation when using our sample of S&P100 stocks. The sample of S&P100 stocks covers the month of February 2002. For NYSE stocks we use quotes posted on two exchanges, the NYSE and the MIDWEST. The data come from the TAQ database.



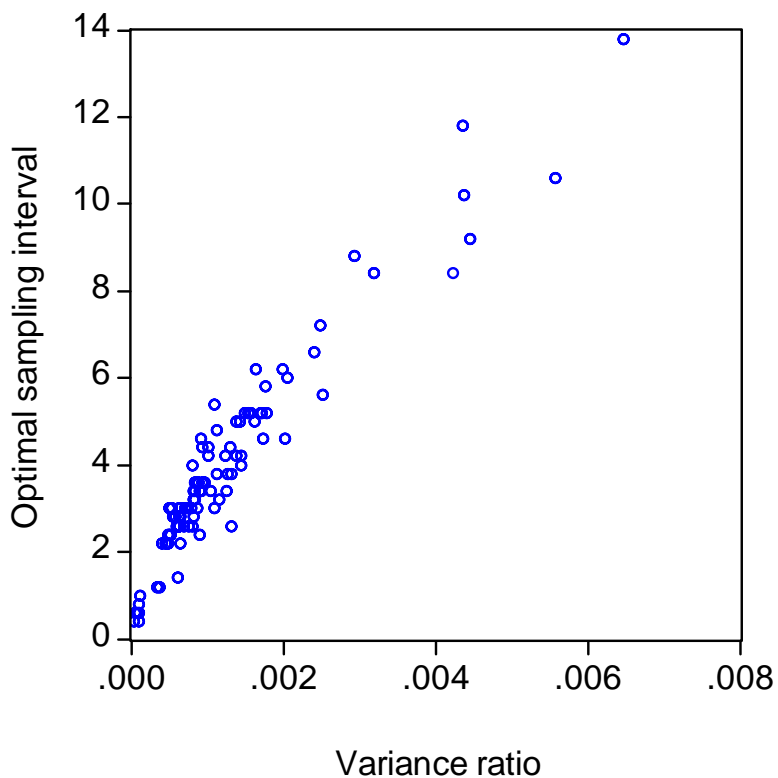
**Fig 6.** Histogram of the estimated, approximate, optimal sampling intervals (in minutes) for the S&P100 stocks. The figure shows the histogram of the approximate number of minutes that should be used, based on Proposition 4 in the main text, to optimally construct continuously compounded returns for the purpose of realized variance estimation when using our sample of S&P100 stocks. The sample of S&P100 stocks covers the month of February 2002. For NYSE stocks we use quotes posted on two exchanges, the NYSE and the MIDWEST. The data come from the TAQ database.



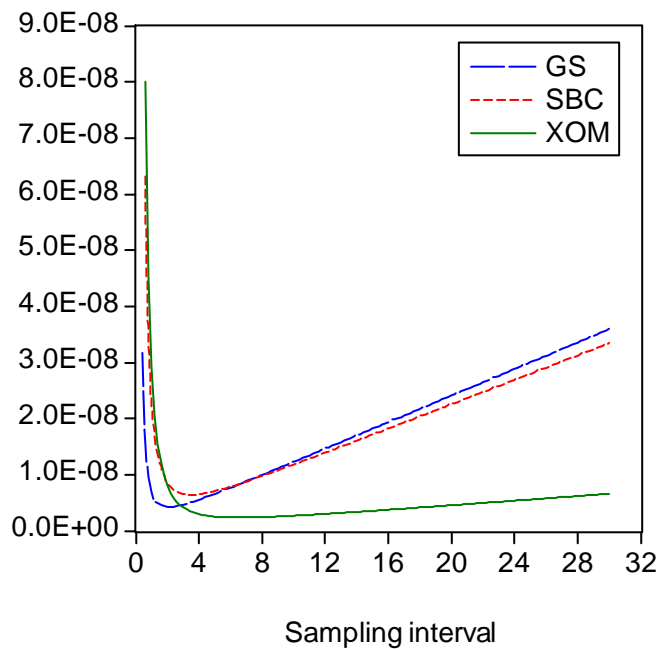
**Fig 7.** The estimated optimal sampling intervals (in minutes) for the S&P100 stocks versus the corresponding approximate optimal sampling intervals (in minutes). The figure shows the scatter-plot of the number of minutes that should be used, based on Proposition 3 in the main text, to optimally construct continuously compounded returns for the purpose of realized variance estimation when using our sample of S&P100 stocks versus the corresponding approximate sampling intervals from Proposition 4. The sample of S&P100 stocks covers the month of February 2002. For NYSE stocks we use quotes posted on two exchanges, the NYSE and the MIDWEST. The data come from the TAQ database.



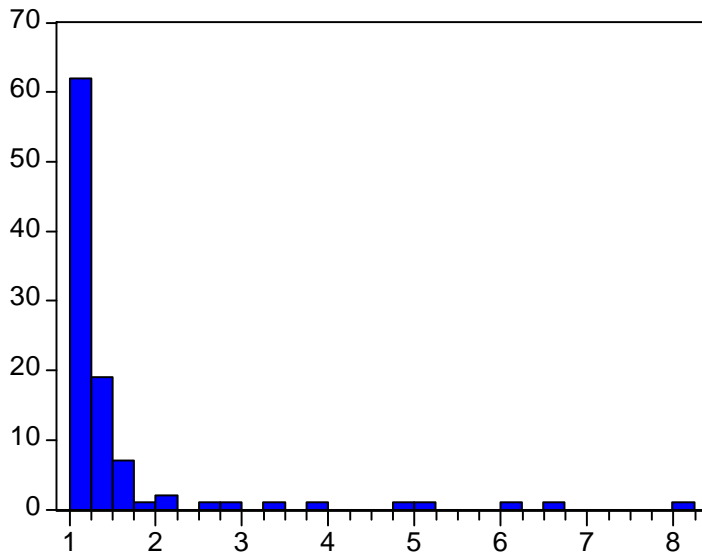
**Fig 8.** The logarithmic MSE values of the realized variance estimator based on optimal sampling intervals versus the corresponding values based on approximate optimal sampling intervals for the S&P100 stocks. The figure shows the scatter-plot of the logarithmic MSE values associated with daily realized variance estimates obtained by sampling continuously compounded returns optimally (as in Proposition 3 in the main text) versus the logarithmic MSE values of daily realized variance estimates obtained by sampling continuously compounded returns approximately optimally (as in Proposition 4 in the main text). The sample of S&P100 stocks covers the month of February 2002. For NYSE stocks we use quotes posted on two exchanges, the NYSE and the MIDWEST. The data come from the TAQ database.



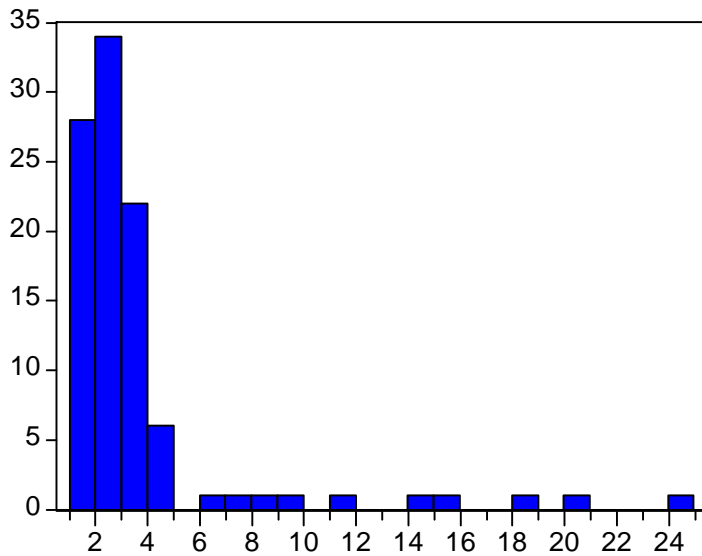
**Fig 9.** The optimal sampling intervals versus the ratios between the second moment of the noise process and the average integrated variance of the efficient price for the S&P100 stocks. The figure shows the scatter-plot of the optimal sampling intervals (in minutes) to be used to construct continuously compounded returns for the purpose of variance estimation (as detailed in Proposition 3) versus the ratios between the variances of the noise components and the average (optimally sampled) daily realized variances. The variances of the noise components are the sample second moments of the quote-to-quote continuously compounded returns. The optimally sampled realized variances are computed by summing squared continuously compounded returns constructed using midpoint bid-ask quotes sampled every  $D^*$  minutes, where  $D^*$  is the optimal sampling interval from Proposition 3 in the main text. The sample of S&P100 stocks covers the month of February 2002. For NYSE stocks we use quotes posted on two exchanges, the NYSE and the MIDWEST. The data come from the TAQ database.



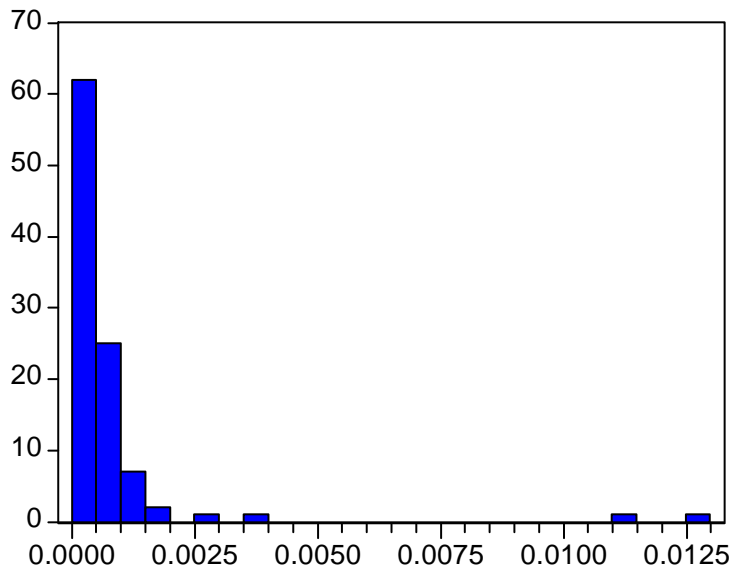
**Fig 10.** The estimated MSEs of the realized variance estimator for Goldman Sachs (GS), SBC communications (SBC), and EXXON Mobile Corporation (XOM). The realized variance estimator is constructed as the sum of squared continuously compounded returns based on midquotes. The MSEs are plotted as functions of the sampling interval (in minutes) used to compute the continuously compounded returns. The sample covers the month of February 2002. For NYSE stocks we use quotes posted on two exchanges, the NYSE and the MIDWEST. The data come from the TAQ database.



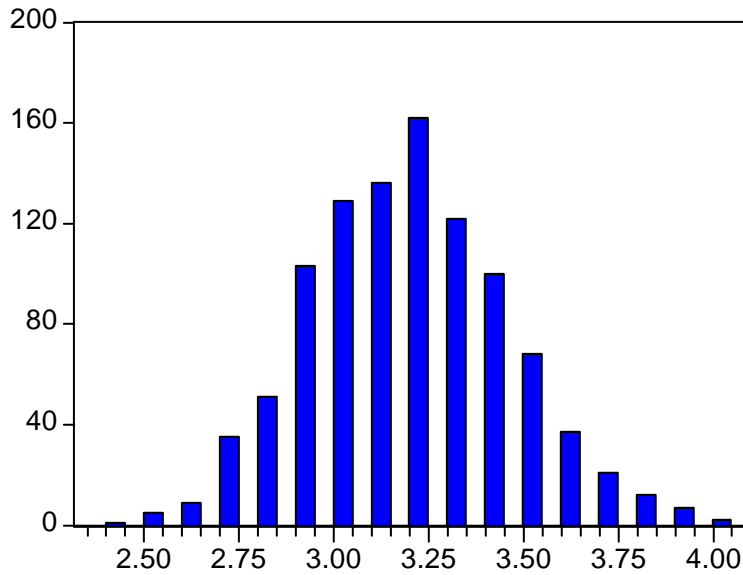
**Fig 11.** Histogram of the ratios between the MSE values associated with the five-minute daily realized variances and the MSE values associated with the optimally sampled daily realized variances for the S&P100 stocks. The optimally sampled realized variances are computed by summing squared continuously compounded returns constructed using midpoint bid-ask quotes sampled every  $D^*$  minutes, where  $D^*$  is the optimal sampling interval from Proposition 3 in the main text. The five-minute realized variances are computed by summing squared continuously-compounded returns constructed by sampling every five minutes. The sample of S&P 100 stocks covers the month of February 2002. For NYSE stocks we use quotes posted on two exchanges, the NYSE and the MIDWEST. The data come from the TAQ database.



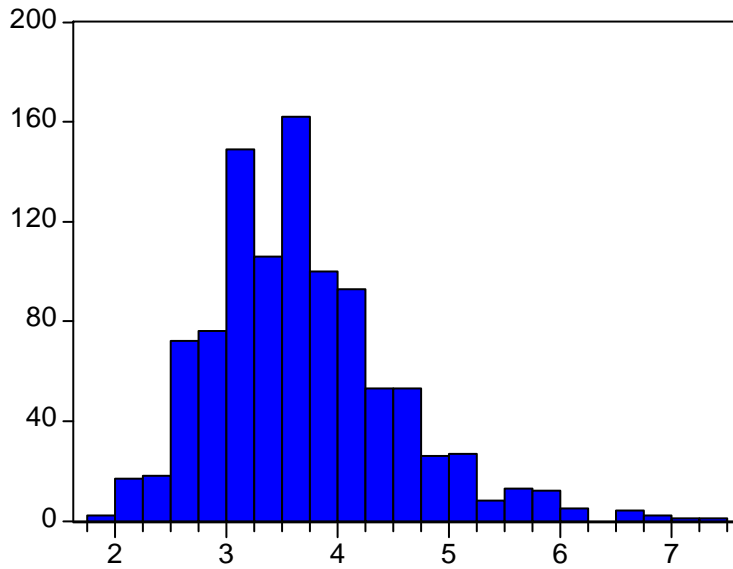
**Fig 12.** Histogram of the ratios between the MSE values associated with the 15-minute realized variances and the MSE values associated with the optimally sampled daily realized variances for the S&P100 stocks. The optimally sampled realized variances are computed by summing squared continuously-compounded returns constructed using midpoint bid-ask quotes sampled every  $D^*$  minutes, where  $D^*$  is the optimal sampling interval from Proposition 3 in the main text. The 15-minute realized variances are computed by summing squared continuously compounded returns constructed by sampling every 15 minutes. The sample of S&P100 stocks covers the month of February 2002. For NYSE stocks we use quotes posted on two exchanges, the NYSE and the MIDWEST. The data come from the TAQ database.



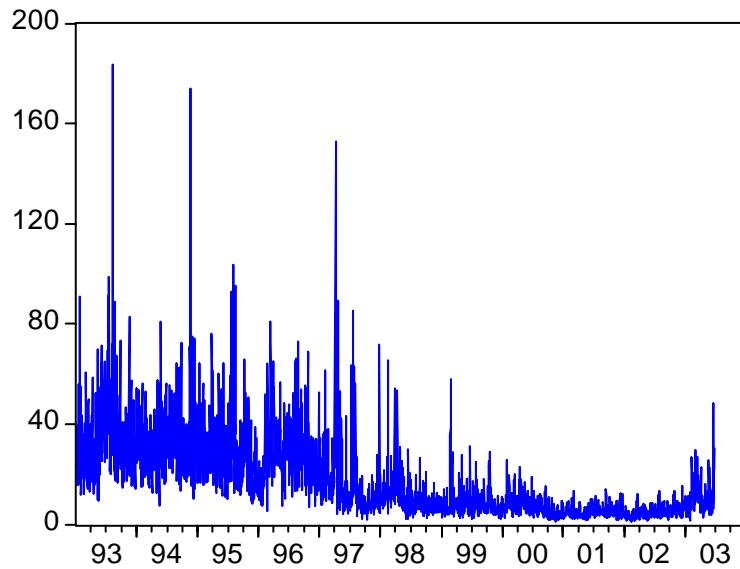
**Fig 13.** Histogram of the average (optimally sampled) daily realized variances for the S&P100 stocks. The optimally sampled realized variances are computed by summing squared continuously compounded returns constructed using midpoint bid-ask quotes sampled every  $D^*$  minutes, where  $D^*$  is the optimal sampling interval from Proposition 3 in the main text. The sample of S&P100 stocks covers the month of February 2002. For NYSE stocks we use quotes posted on two exchanges, the NYSE and the MIDWEST. The data come from the TAQ database.



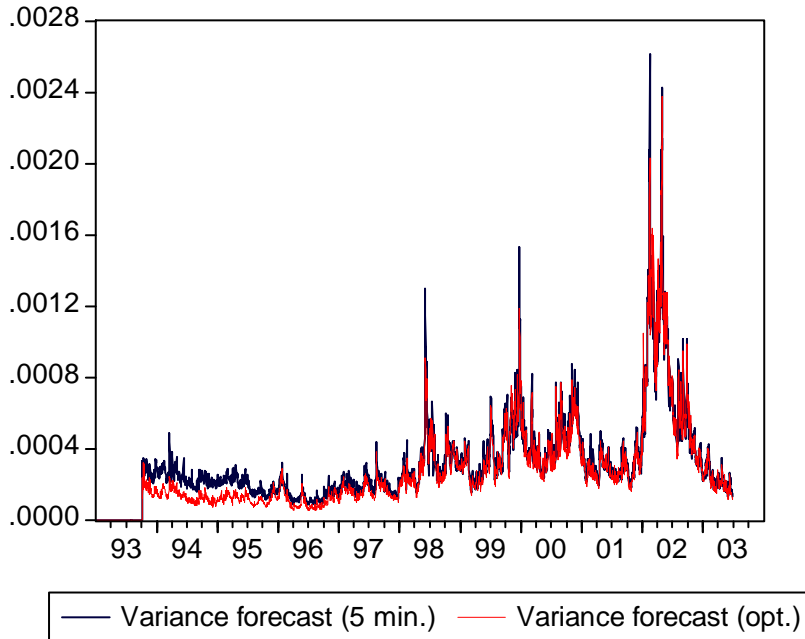
**Fig 14.** Histogram of the simulated distribution of the optimal sampling intervals of the realized variance estimator. As we detail in Section 7, we simulate an observed price process affected by microstructure noise contaminations on the basis of a ratio between microstructure noise variance and average quadratic variation of the underlying efficient price process which mimics the empirical features of Goldman Sachs (GS). The price process is simulated 1,000 times. For every simulation we compute the optimal sampling interval of the realized variance estimator as described in Proposition 3 in the main text. The quarticity used as an input to compute the optimal sampling interval is obtained by sampling continuously compounded returns at the quarticity optimal frequency (i.e., 2.13 minutes in our case). The true optimal sampling interval of the realized variance estimator is 2.8 minutes in our case.



**Fig 15.** Histogram of the simulated distribution of the optimal sampling intervals of the realized variance estimator. As we detail in Section 7, we simulate an observed price process affected by microstructure noise contaminations on the basis of a ratio between microstructure noise variance and average quadratic variation of the underlying efficient price process which mimics the empirical features of Goldman Sachs (GS). The price process is simulated 1,000 times. For every simulation we compute the optimal sampling interval of the realized variance estimator as described in Proposition 3 in the main text. The quarticity used as an input to compute the optimal sampling frequency is obtained by sampling continuously compounded returns every 15 minutes. The true optimal sampling interval of the realized variance estimator is 2.8 minutes in our case.



**Fig 16.** Time series plot of the daily SBC optimal sampling intervals (in minutes) between January 1993 and December 2003. We plot the time series of daily estimates of the optimal intervals to be used to sample continuously compounded midquote returns for the purpose of realized variance computation in the case of SBC communications (SBC). The optimal intervals are obtained as detailed in Proposition 3 in the main text. The SBC data consists of quotes posted on two exchanges, the NYSE and the MIDWEST, between January 1993 and December 2003. The data come from the TAQ data set.



**Fig 17.** Time series plot of the SBC one-day-ahead variance forecasts based on optimally sampled daily realized variances and five-minute daily realized variances. We use quote data for SBC communications (SBC) over the period between January 1993 and December 2003. The data come from the TAQ database. The optimally sampled daily realized variances are computed by summing squared continuously compounded returns constructed using midpoint bid-ask quotes sampled every  $D^*$  minutes, where  $D^*$  is the optimal sampling interval from Proposition 3 in the main text. The five-minute daily realized variances are computed by summing squared continuously compounded returns constructed by sampling every five minutes. The one-day-ahead forecasts are obtained by virtue of an ARFIMA(2, $d$ ,2) model. We use GPH estimates of the  $d$  parameter. The estimated  $d$  values are equal to 0.45 in the case of the optimally sampled realized variances and 0.44 in the case of the five-minute realized variances. We employ 200 observations to construct the first forecast. The total number of one-day-ahead forecasts is 1,875.