Editor's Note: The following article was the *JBES* Invited Address presented at the Joint Statistical Meetings, Minneapolis, Minnesota, August 7–11, 2005.

Realized Variance and Market Microstructure Noise

Peter R. HANSEN
Department of Economics, Stanford University, 579 Serra Mall, Stanford, CA 94305-6072
(*peter.hansen@stanford.edu*)

Asger LUNDE
Department of Marketing and Statistics, Aarhus School of Business, Fuglesangs Alle 4, 8210 Aarhus V, Denmark

We study market microstructure noise in high-frequency data and analyze its implications for the realized variance (RV) under a general specification for the noise. We show that kernel-based estimators can unearth important characteristics of market microstructure noise and that a simple kernel-based estimator dominates the RV for the estimation of integrated variance (IV). An empirical analysis of the Dow Jones Industrial Average stocks reveals that market microstructure noise is time-dependent and correlated with increments in the efficient price. This has important implications for volatility estimation based on high-frequency data. Finally, we apply cointegration techniques to decompose transaction prices and bid–ask quotes into an estimate of the efficient price and noise. This framework enables us to study the dynamic effects on transaction prices and quotes caused by changes in the efficient price.

KEY WORDS: Bias correction; High-frequency data; Integrated variance; Market microstructure noise; Realized variance; Realized volatility; Sampling schemes.

The great tragedy of Science—the slaying of a beautiful hypothesis by an ugly fact (Thomas H. Huxley, 1825–1895).

1. INTRODUCTION

The presence of market microstructure noise in high-frequency financial data complicates the estimation of financial volatility and makes standard estimators, such as the realized variance (RV), unreliable. Thus, from the perspective of volatility estimation, market microstructure noise is an "ugly fact" that challenges the validity of theoretical results that rely on the absence of noise. Volatility estimation in the presence of market microstructure noise is currently a very active area of research. Interestingly, this literature was initiated by an article by Zhou (1996) that was published in this journal a decade ago and was in many ways 10 years ahead of its time.

The best remedy for market microstructure noise depends on the properties of the noise, and the main purpose of this article is to unearth the empirical properties of market microstructure noise. We use a number of kernel-based estimators that are well suited for this problem, and our empirical analysis of high-frequency stock returns reveals the following ugly facts about market microstructure noise:

1. The noise is correlated with the efficient price.
2. The noise is time-dependent.
3. The noise is quite small in the Dow Jones Industrial Average (DJIA) stocks.
4. The properties of the noise have changed substantially over time.

These four empirical "facts" are related to one another and have important implications for volatility estimation. The time dependence in the noise and the correlation between noise and efficient price arise naturally in some models on market microstructure effects, including (a generalized version of) the bid–ask model by Roll (1984) (see Hasbrouck 2004 for a discussion) and models where agents have asymmetric information, such as those by Glosten and Milgrom (1985) and Easley and O'Hara (1987, 1992). Market microstructure noise has many sources, including the discreteness of the data (see Harris 1990, 1991) and properties of the trading mechanism (see, e.g., Black 1976; Amihud and Mendelson 1987). (For additional references to this literature, see, e.g., O'Hara 1995; Hasbrouck 2004.)

The main contributions of this article are as follows: First, we characterize how the RV is affected by market microstructure noise under a general specification for the noise that allows for various forms of stochastic dependencies. Second, we show that market microstructure noise is time-dependent and correlated with efficient returns. Third, we consider some existing theoretical results based on assumptions about the noise that are too simplistic, and discuss when such results provide reasonable approximations. For example, our empirical analysis of the 30 DJIA stocks shows that the noise may be ignored when intraday returns are sampled at relatively low frequencies, such as 20-minute sampling. Assuming the noise is of an independent type seems to be reasonable when intraday returns are sampled every 15 ticks or so. Fourth, we apply cointegration methods to decompose transaction prices and bid/ask quotations into estimates of the efficient price and market microstructure noise. The correlations between these estimated series are consistent with the volatility signature plots. The cointegration analysis enables us to study how a change in the efficient price dynamically affects bid, ask, and transaction prices.

© 2006 American Statistical Association
Journal of Business & Economic Statistics
April 2006, Vol. 24, No. 2
DOI 10.1198/073500106000000071

The interest for empirical quantities based on high-frequency data has surged in recent years (see Barndorff-Nielsen and Shephard 2007 for a recent survey). The RV is a well-known quantity that goes back to Merton (1980). Other empirical quantities include bipower variation and multipower variation, which are particularly useful for detecting jumps (see Barndorff-Nielsen and Shephard 2003, 2004, 2006a,b; Andersen, Bollerslev, and Diebold 2003; Bollerslev, Kretschmer, Pigorsch, and Tauchen 2005; Huang and Tauchen 2005; Tauchen and Zhou 2004), and intraday range-based estimators (see Christensen and Podolskij 2005). High-frequency–based quantities have proven useful for a number of problems. For example, several authors have applied filtering and smoothing techniques to time series of the RV to obtain time series for daily volatility (see, e.g., Maheu and McCurdy 2002; Barndorff-Nielsen, Nielsen, Shephard, and Ysusi 1996; Engle and Sun 2005; Frijns and Lehnert 2004; Koopman, Jungbacker, and Hol 2005; Hansen and Lunde 2005b; Owens and Steigerwald 2005). High-frequency–based quantities are also useful in the context of forecasting (see Andersen, Bollerslev, and Meddahi 2004; Ghysels, Santa-Clara, and Valkanov 2006) and the evaluation and comparison of volatility models (see Andersen and Bollerslev 1998; Hansen, Lunde, and Nason 2003; Hansen and Lunde 2005a, 2006; Patton 2005).

The RV, which is a sum of squared intraday returns, yields a perfect estimate of volatility in the ideal situation where prices are observed continuously and without measurement error (see, e.g., Merton 1980). This result suggests that the RV should be based on intraday returns sampled at the highest possible frequency (tick-by-tick data). Unfortunately, the RV suffers from a well-known bias problem that tends to get worse as the sampling frequency of intraday returns increases (see, e.g., Fang 1996; Andreou and Ghysels 2002; Oomen 2002; Bai, Russell, and Tiao 2004). The source of this bias problem is known as market microstructure noise, and the bias is particularly evident in *volatility signature plots* (see Andersen, Bollerslev, Diebold, and Labys 2000b). Thus there is a trade-off between bias and variance when choosing the sampling frequency, as discussed by Bandi and Russell (2005) and Zhang, Mykland, and Aït-Sahalia (2005). This trade-off is the reason that the RV is often computed from intraday returns sampled at a moderate frequency, such as 5-minute or 20-minute sampling.

A key insight into the problem of estimating the volatility from high-frequency data comes from its similarity to the problem of estimating the long-run variance of a stationary time series. In this literature it is well known that autocorrelation necessitates modifications of the usual sum-of-squared estimator. Those modifications of Newey and West (1987) and Andrews (1991) provided such estimators that are robust to autocorrelation. Market microstructure noise induces autocorrelation in the intraday returns, and this autocorrelation is the source of the RV's bias problem. Given this connection to long-run variance estimation, it is not surprising that "prewhitening" of intraday returns and kernel-based estimators (including the closely related subsample-based estimators) are found to be useful in the present context. Zhou (1996) introduced the use of kernel-based estimators and the subsampling idea to deal with market microstructure noise in high-frequency data. Filtering techniques have been used by Ebens (1999), Andersen, Bollerslev, Diebold, and Ebens (2001), and Maheu and McCurdy (2002) (moving average filter) and Bollen and Inder (2002) (autoregressive filter). Kernel-based estimators were explored by Zhou (1996), Hansen and Lunde (2003), and Barndorff-Nielsen, Hansen, Lunde, and Shephard (2004) and the closely related subsample-based estimators were used in an unpublished paper by Müller (1993) and also by Zhou (1996), Zhang et al. (2005), and Zhang (2004).

The rest of the article is organized as follows. In Section 2 we describe our theoretical framework and discuss sampling schemes in calendar time and tick time. We also characterize the bias of the RV under a general specification for the noise. In Section 3 we consider the case with independent market microstructure noise, which has been used by various authors, including Corsi, Zumbach, Müller, and Dacorogna (2001), Curci and Corsi (2004), Bandi and Russell (2005), and Zhang et al. (2005). We consider a simple kernel-based estimator of Zhou (1996) that we denote by RV_{AC_1} because it uses the first-order autocorrelation to bias-correct the RV. We benchmark RV_{AC_1} to the standard measure of RV and find that the former is superior to the latter in terms of the mean squared error (MSE). We also evaluate the implications for some theoretical results based on assumptions in which market microstructure noise is absent. Interestingly, we find that the root mean squared error (RMSE) of the RV in the presence of noise is quite similar to those that ignore the noise at low sampling frequencies, such as 20-minute sampling. This finding is important because many existing empirical studies have drawn conclusions from 20-minute and 30-minute intraday returns, using the results of Barndorff-Nielsen and Shephard (2002). However, at 5-minute sampling we find that the "true" confidence interval about the RV can be as much as 100% larger than those based on an "absence of noise assumption." In Section 4 we present a robust estimator that is unbiased for a general type of noise and discuss noise that is time-dependent in both calendar time and tick time. We also discuss the subsampling version of Zhou's estimator, which is robust to some forms of time-dependence in tick time. In Section 5 we describe our data and present most of our empirical results. The key result is the overwhelming evidence against the independent noise assumption. This finding is quite robust to the choice of sampling method (calendar time or tick time) and the type of price data (transaction prices or quotation prices). This dependence structure has important implications for many quantities based on ultra-high–frequency data. These features of the noise have important implications for some of the bias corrections that have been used in the literature. Although the independent noise assumption may be fairly reasonable when the tick size is $1/16$, it is clearly not consistent with the recent data. In fact, much of the noise has "evaporated" after the tick size is reduced to 1 cent. In Section 6 we present a cointegration analysis of the vector of bid, ask, and transaction prices. The Granger representation makes it possible to decompose each of the price series into noise and a common efficient price. Further, based on this decomposition we estimate impulse response functions that reveal the dynamic effects on bid, ask, and transaction prices as a response to a change in the efficient price. In Section 7 we provide a summary, and we conclude the article with three appendixes that provide proofs and details about our estimation methods.

2. THE THEORETICAL FRAMEWORK

We let $\{p^*(t)\}$ denote a latent log-price process in continuous time and use $\{p(t)\}$ to denote the observable log-price process. Thus the noise process is given by

$$u(t) \equiv p(t) - p^*(t).$$

The noise process, u, may be due to market microstructure effects, such as bid–ask bounces, but the discrepancy between p and p^* can also be induced by the technique used to construct $p(t)$. For example, p is often constructed artificially from observed transactions or quotes using the *previous tick* method or the *linear interpolation* method, which we define and discuss later in this section.

We work under the following specification for the efficient price process, p^*.

Assumption 1. The efficient price process satisfies $dp^*(t) = \sigma(t)\,dw(t)$, where $w(t)$ is a standard Brownian motion, σ is a random function that is independent of w, and $\sigma^2(t)$ is Lipschitz (almost surely).

In our analysis we condition on the volatility path, $\{\sigma^2(t)\}$, because our analysis focuses on estimators of the integrated variance (IV),

$$IV \equiv \int_a^b \sigma^2(t)\,dt.$$

Thus we can treat $\{\sigma^2(t)\}$ as deterministic even though we view the volatility path as random. The Lipschitz condition is a smoothness condition that requires $|\sigma^2(t) - \sigma^2(t+\delta)| < \epsilon\delta$ for some ϵ and all t and δ (with probability 1). The assumption that w and σ are independent is not essential. The connection between kernel-based and subsample-based estimators (see Barndorff-Nielsen et al. 2004), shows that weaker assumptions, used by Zhang et al. (2005) and Zhang (2004), are sufficient in this framework.

We partition the interval $[a, b]$ into m subintervals, and m plays a central role in our analysis. For example, we derive asymptotic distributions of quantities as $m \to \infty$. This type of *infill asymptotics* is commonly used in spatial data analysis and goes back to Stein (1987). Related to the present context is the use of infill asymptotics for estimation of diffusions (see Bandi and Phillips 2004). For a fixed m, the ith subinterval is given by $[t_{i-1,m}, t_{i,m}]$, where $a = t_{0,m} < t_{1,m} < \cdots < t_{m,m} = b$. The length of the ith subinterval is given by $\delta_{i,m} \equiv t_{i,m} - t_{i-1,m}$, and we assume that $\sup_{i=1,\ldots,m} \delta_{i,m} = O(\frac{1}{m})$, such that the length of each subinterval shrinks to 0 as m increases. The *intraday returns* are now defined by

$$y^*_{i,m} \equiv p^*(t_{i,m}) - p^*(t_{i-1,m}), \qquad i = 1, \ldots, m,$$

and the increments in p and u are defined similarly and denoted by

$$y_{i,m} \equiv p(t_{i,m}) - p(t_{i-1,m}), \qquad i = 1, \ldots, m,$$

and

$$e_{i,m} \equiv u(t_{i,m}) - u(t_{i-1,m}), \qquad i = 1, \ldots, m.$$

Note that the *observed intraday returns* decompose into $y_{i,m} = y^*_{i,m} + e_{i,m}$. The IV over each of the subintervals is defined by

$$\sigma^2_{i,m} \equiv \int_{t_{i-1,m}}^{t_{i,m}} \sigma^2(s)\,ds, \qquad i = 1, \ldots, m,$$

and we note that $\operatorname{var}(y^*_{i,m}) = E(y^{*2}_{i,m}) = \sigma^2_{i,m}$ under Assumption 1.

The RV of p^* is defined by

$$RV^{(m)}_* \equiv \sum_{i=1}^m y^{*2}_{i,m},$$

and $RV^{(m)}_*$ is consistent for the IV as $m \to \infty$ (see, e.g., Protter 2005). A feasible asymptotic distribution theory of RV (in relation to IV) was established by Barndorff-Nielsen and Shephard (2002) (see also Meddahi 2002; Mykland and Zhang 2006; Gonçalves and Meddahi 2005). Whereas $RV^{(m)}_*$ is an ideal estimator, it is not a feasible estimator because p^* is latent. The realized variance of p, given by

$$RV^{(m)} \equiv \sum_{i=1}^m y^2_{i,m},$$

is observable but suffers from a well-known bias problem and is generally inconsistent for the IV (see, e.g., Bandi and Russell 2005; Zhang et al. 2005).

2.1 Sampling Schemes

Intraday returns can be constructed using different types of sampling schemes. The special case where $t_{i,m}$, $i = 1, \ldots, m$, are equidistant in calendar time [i.e., $\delta_{i,m} = (b-a)/m$ for all i] is referred to as *calendar time sampling* (CTS). The widely used exchange rates data from Olsen and associates (see Müller et al. 1990) are equidistant in time, and 5-minute sampling ($\delta_{i,m} =$ 5 min) is often used in practice.

CTS requires the construction of artificial prices from the raw (irregularly spaced) price data (transaction prices or quotations). Given observed prices at the times $t_0 < \cdots < t_N$, one can construct a price at time $\tau \in [t_j, t_{j+1})$, using

$$p(\tau) \equiv p_{t_j}$$

or

$$\tilde{p}(\tau) \equiv p_{t_j} + \frac{\tau - t_j}{t_{j+1} - t_j}\big(p_{t_{j+1}} - p_{t_j}\big).$$

The former is known as the *previous tick* method (Wasserfallen and Zimmermann 1985), and the latter is the *linear interpolation* method (see Andersen and Bollerslev 1997). Both methods have been discussed by Dacorogna, Gencay, Müller, Olsen, and Pictet (2001, sec. 3.2.1). When sampling at ultra-high frequencies, the linear interpolation method has the following unfortunate property, where "$\xrightarrow{p}$" denotes convergence in probability.

Lemma 1. Let N be fixed and consider the RV based on the linear interpolation method. It holds that $RV^{(m)} \xrightarrow{p} 0$ as $m \to \infty$.

The result of Lemma 1 essentially boils down to the fact that the quadratic variation of a straight line is zero. Although this is a limit result (as $m \to \infty$), the lemma does suggest that the linear interpolation method is not suitable for the construction of intraday returns at high frequencies, where sampling may occur multiple times between two neighboring price observations. That the result of Lemma 1 is more than a theoretical artifact is evident from the volatility signature plots of Hansen and Lunde (2003). Given the result of Lemma 1, we avoid the use of the linear interpolation and use the previous tick method to construct CTS intraday returns.

The case where $t_{i,m}$ denotes the time of a transaction/quotation is referred to as *tick time sampling* (TTS). An example of TTS is when $t_{i,m}$, $i = 1, \ldots, m$, are chosen to be the time of every fifth transaction, say.

The case where the sampling times, $t_{0,m}, \ldots, t_{m,m}$, are such that $\sigma^2_{i,m} = IV/m$ for all $i = 1, \ldots, m$ is known as *business time sampling* (BTS) (see Oomen 2006). Zhou (1998) referred to BTS intraday returns as de-volatized returns and discussed distributional advantages of BTS returns. Whereas $t_{i,m}$, $i = 0, \ldots, m$, are observable under CTS and TTS, they are latent under BTS, because the sampling times are defined from the unobserved volatility path. Empirical results of Andersen and Bollerslev (1997) and Curci and Corsi (2004) suggest that BTS can be approximated by TTS. This feature is nicely captured in the framework of Oomen (2006), where the (random) tick times are generated with an intensity directly related to a quantity corresponding to $\sigma^2(t)$ in the present context. Under CTS, we sometimes write $RV^{(x \text{ sec})}$, where x seconds is the period in time spanned by each of the intraday returns (i.e., $\delta_{i,m} = x$ seconds). Similarly, we write $RV^{(y \text{ tick})}$ under TTS when each intraday return spans y ticks (transactions or quotations).

2.2 Characterizing the Bias of the Realized Variance Under General Noise

Initially, we make the following assumptions about the noise process, u.

Assumption 2. The noise process, u, is covariance stationary with mean 0, such that its autocovariance function is defined by $\pi(s) \equiv E[u(t)u(t+s)]$.

The covariance function, π, plays a key role because the bias of $RV^{(m)}$ is tied to the properties of $\pi(s)$ in the neighborhood of 0. Simple examples of noise processes that satisfy Assumption 2 include the independent noise process, which has $\pi(s) = 0$ for all $s \neq 0$, and the Ornstein–Uhlenbeck process. The latter was used by Aït-Sahalia, Mykland, and Zhang (2005a) to study estimation in a parametric diffusion model that is robust to market microstructure noise.

An important aspect of our analysis is that our assumptions allow for a dependence between u and p^*. This is a generalization of the assumptions made in the existing literature, and our empirical analysis shows that this generalization is needed, in particularly when prices are sampled from quotations.

Next, we characterize the RV bias under these general assumptions for the market microstructure noise, u.

Theorem 1. Given Assumptions 1 and 2, the bias of the realized variance under CTS is given by

$$E\big[RV^{(m)} - IV\big] = 2\rho_m + 2m\left[\pi(0) - \pi\left(\frac{b-a}{m}\right)\right], \quad (1)$$

where $\rho_m \equiv E(\sum_{i=1}^m y^*_{i,m} e_{i,m})$.

The result of Theorem 1 is based on the following decomposition of the observed RV:

$$RV^{(m)} = \sum_{i=1}^m y^{*2}_{i,m} + 2\sum_{i=1}^m e_{i,m} y^*_{i,m} + \sum_{i=1}^m e^2_{i,m},$$

where $\sum_{i=1}^m e^2_{i,m}$ is the "realized variance" of the noise process u responsible for the last bias term in (1). The dependence between u and p^* that is relevant for our analysis is given in the form of the correlation between the efficient intraday returns, $y^*_{i,m}$, and the return noise, $e_{i,m}$. By the Cauchy–Schwarz inequality, $\pi(0) \geq \pi(s)$ for all s, such that the bias is always positive when the return noise process, $e_{i,m}$, is uncorrelated with the efficient intraday returns $y^*_{i,m}$ (because this implies that $\rho_m = 0$). Interestingly, the total bias can be negative. This occurs when $\rho_m < -m[\pi(0) - \pi(\Delta_m)]$, which is the case where the downward bias (caused by a negative correlation between $e_{i,m}$ and $y^*_{i,m}$) exceeds the upward bias caused by the "realized variance" of u. This appears to be the case for the RVs that are based on quoted prices, as shown in Figure 1.

The last term of the bias expression in Theorem 1 shows that the bias is tied to the properties of $\pi(s)$ in the neighborhood of 0, and, as $m \to \infty$ (hence $\delta_m \to 0$), we obtain the following result.

Corollary 1. Suppose that the assumptions of Theorem 1 hold and that $\pi(s)$ is differentiable at 0. Then the asymptotic bias is given by

$$\lim_{m\to\infty} E\big[RV^{(m)} - IV\big] = 2\rho - 2(b-a)\pi'(0),$$

provided that $\rho \equiv \lim_{m\to\infty} E(\sum_{i=1}^m y^*_{i,m} e_{i,m})$ is well defined.

Under the independent noise assumption, we can define $\pi'(0) = -\infty$, which is the situation that we analyze in detail in Section 3. A related asymptotic result is obtained whenever the quadratic variation of the bivariate process, $(p^*, u)'$, is well defined, such that $[p, p] = [p^*, p^*] + 2[p^*, u] + [u, u]$, where $[X, Y]$ denotes the quadratic covariation. In this setting we have $IV = [p^*, p^*]$ such that

$$RV^{(m)} - IV \xrightarrow{p} 2[p^*, u] + [u, u] \quad (\text{as } m \to \infty),$$

where $\rho = [p^*, u]$ and $-2(b-a)\pi'(0) = [u, u]$ (almost surely under additional assumptions).

A volatility signature plot provides an easy way to visually inspect the potential bias problems of RV-type estimators. Such plots first appeared in an unpublished thesis by Fang (1996) and were named and made popular by Andersen et al. (2000b). Let $RV^{(m)}_t$ denote the RV based on m intraday returns on day t. A volatility signature plot displays the sample average,

$$\overline{RV}^{(m)} \equiv n^{-1} \sum_{t=1}^n RV^{(m)}_t,$$

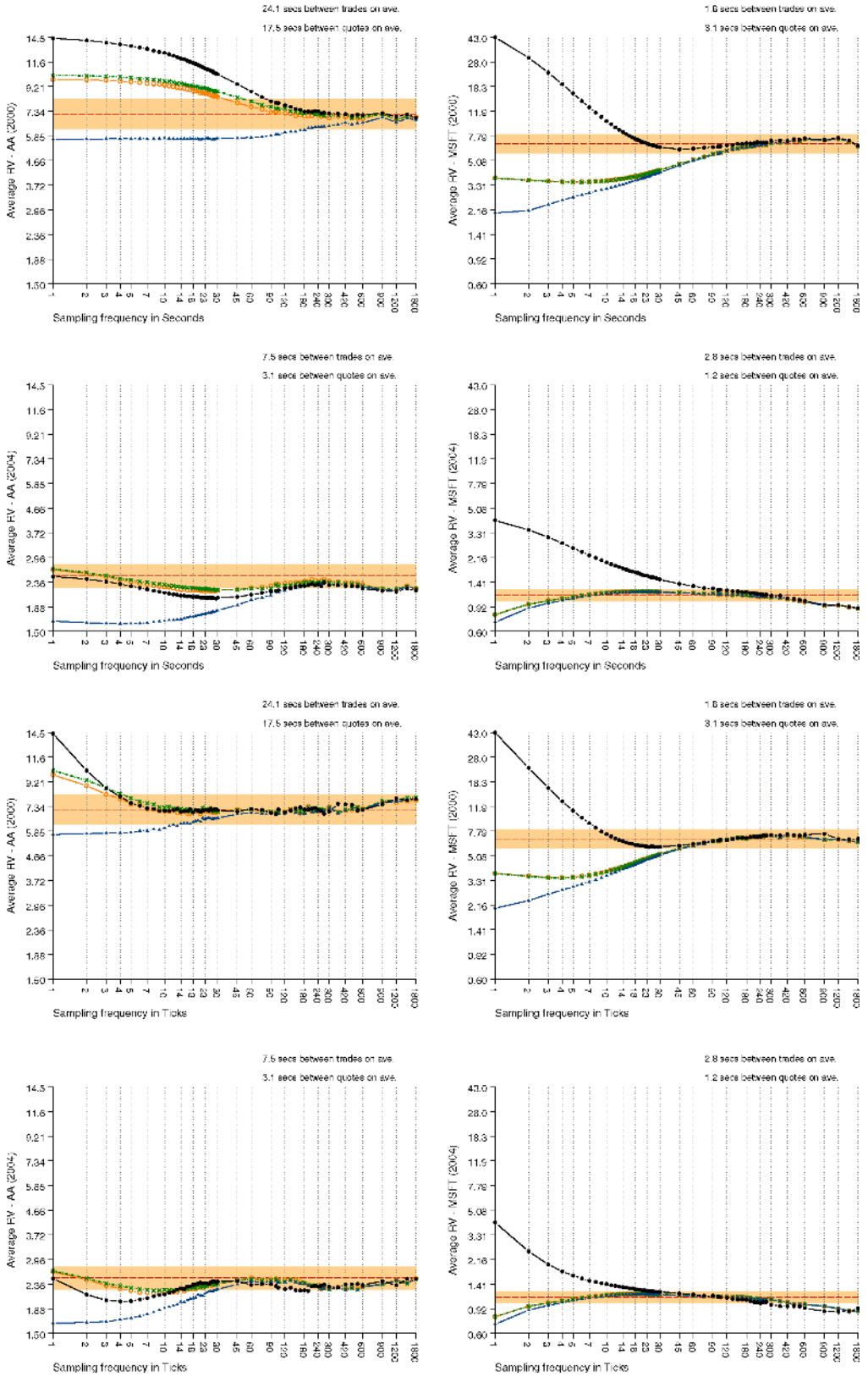

Figure 1. Volatility Signature Plots for RV_t Based on Ask Quotes (▫—▫—▫), Bid Quotes (×···×···×), Mid-Quotes (▲—▲—▲), and Transaction Prices (▪—●—▪). The left column is for AA and the right column is for MSFT. The two top rows are based on calendar time sampling, in contrast to the bottom rows that are based on tick time sampling. The results for 2000 are the panels in rows 1 and 3, and those for 2004 are in rows 2 and 4. The horizontal line represents an estimate of the average IV, $\bar{\sigma}^2 \equiv \overline{RV}^{(1\ tick)}_{ACNW_{30}}$, that is defined in Section 4.2. The shaded area about $\bar{\sigma}^2$ represents an approximate 95% confidence interval for the average volatility.

as a function of the sampling frequencies m, where the average is taken over multiple periods (typically trading days).

Figure 1 presents volatility signature plots for AA (left) and MSFT (right) using both CTS (rows 1 and 2) and TTS (rows 3 and 4) and based on both transaction data and quotation data. The signature plots are based on daily RVs from the years 2000 (rows 1 and 3) and 2004 (rows 2 and 4), where $RV_t^{(m)}$ is calculated from intraday returns spanning the period 9:30 AM to 16:00 PM (the hours that the exchanges are open). The horizontal line represents an estimate of the average IV, $\bar{\sigma}^2 \equiv \overline{RV}^{(1\text{ tick})}_{\text{ACNW}_{30}}$, defined in Section 4.2. The shaded area about $\bar{\sigma}^2$ represents an approximate 95% confidence interval for the average volatility. These confidence intervals are computed using a method described in Appendix B.

From Figure 1, we see that the RVs based on low and moderate frequencies appear to be approximately unbiased. However, at higher frequencies, the RV becomes unreliable, and the market microstructure effects are pronounced at the ultra-high frequencies, particularly for transaction prices. For example, $\overline{RV}^{(1\text{ sec})}$ is about 47 for MSFT in 2000, whereas $\overline{RV}^{(1\text{ min})}$ is much smaller (about 6.0).

A very important result of Figure 1 is that the volatility signature plots for mid-quotes drop (rather than increases) as the sampling frequency increases (as $\delta_{i,m} \to 0$). This holds for both CTS and TTS. Thus these volatility signature plots provide the first piece of evidence for the ugly facts about market microstructure noise.

Fact I. The noise is negatively correlated with the efficient returns.

Our theoretical results show that ρ_m must be responsible for the negative bias of $RV^{(m)}$. The other bias term, $2m[\pi(0) - \pi(\frac{b-a}{m})]$, is always nonnegative, such that time dependence in the noise process cannot (by itself) explain the negative bias seen in the volatility signature plots for mid-quotes. So Figure 1 strongly suggests that the innovations in the noise process, $e_{i,m}$, are negatively correlated with the efficient returns, $y^*_{i,m}$. Although this phenomenon is most evident for mid-quotes, it is quite plausible that the efficient return is also correlated with each of the noise processes embedded in the three other price series: bid, ask, and transaction prices. At this point it is worth recalling Colin Sautar's words: "Just because you're not paranoid doesn't mean they're not out to get you." Similarly, just because we cannot see a negative bias does not mean that ρ_m is 0. In fact, if $\rho_m > 0$, then it would not be exposed in a simple manner in a volatility signature plot. From

$$\operatorname{cov}(y^*_{i,m}, e^{\text{mid}}_{i,m}) = \frac{1}{2}\operatorname{cov}(y^*_{i,m}, e^{\text{ask}}_{i,m}) + \frac{1}{2}\operatorname{cov}(y^*_{i,m}, e^{\text{bid}}_{i,m}),$$

we see that the noise in bid and/or ask quotes must be correlated with the efficient prices if the noise in mid-quotes is found to be correlated with the efficient price. In Section 6 we present additional evidence of this correlation, which is also found for transaction data.

Nonsynchronous revisions of bid and ask quotes when the efficient price changes is a possible explanation for the negative correlation between noise and efficient returns. An upward movement in prices often causes the ask price to increase before the bid does, whereby the bid–ask spread is temporary widened. A similar widening of the spread occurs when prices go down. This has implications for the quadratic variation of mid-quotes, because a one-tick price increment is divided into two half-tick increments, resulting in quadratic terms that add up to only half that of the bid or ask price $[(\frac{1}{2})^2 + (\frac{1}{2})^2$ versus $1^2]$. Such discrete revisions of the observed price toward the effective price has been used in a very interesting framework by Large (2005), who showed that this may result in a negative bias.

Figure 2 presents typical trading scenarios for AA during three 20-minute periods on April 24, 2004. The prevailing bid and ask prices are given by the edges of the shaded area, and the dots represents actual transaction prices. That the spread tends to get wider when prices move up or down is seen in many places, such as the minutes after 10:00 AM and around 12:15 PM.

3. THE CASE WITH INDEPENDENT NOISE

In this section we analyze the special case where the noise process is assumed to be of an independent type. Our assumptions, which we make precise in Assumption 3, essentially amount to assuming that $\pi(s) = 0$ for all $s \neq 0$ and $p^* \perp\!\!\!\perp u$, where we use "$\perp\!\!\!\perp$" to denote stochastic independence. Most of the existing literature has established results assuming this kind of noise, and in this section we shall draw on several important results from Zhou (1996), Bandi and Russell (2005), and Zhang et al. (2005). Although we have already dismissed this form of noise as an accurate description of the noise in our data, there are several good arguments for analyzing the properties of the RV and related quantities under this assumption. The independent noise assumption makes the analysis tractable and provides valuable insight into the issues related to market microstructure noise. Furthermore, although the independent noise assumption is inaccurate at ultra-high sampling frequencies, the implications of this assumptions may be valid at lower sampling frequencies. For example, it may be reasonable to assume that the noise is independent when prices are sampled every minute. On the other hand, for some purposes the independent noise assumption can be quite misleading, as we discuss in Section 5.

We focus on a kernel estimator originally proposed by Zhou (1996) that incorporates the first-order autocovariance. A similar estimator was applied to daily return series by French, Schwert, and Stambaugh (1987). Our use of this estimator has three purposes. First, we compare this simple bias-corrected version of the realized variance to the standard measure of the realized variance, and find that these results are generally quite favorable to the bias-corrected estimator. Second, our analysis makes it possible to quantify the accuracy of results based on no-noise assumptions, such as the asymptotic results by Jacod (1994), Jacod and Protter (1998), Barndorff-Nielsen and Shephard (2002), and Mykland and Zhang (2006) and to evaluate whether the bias-corrected estimator is less sensitive to market microstructure noise. Finally, we use the bias-corrected estimator to analyze the validity of the independent noise assumption.

Assumption 3. The noise process satisfies the following:

(a) $p^* \perp\!\!\!\perp u$, $u(s) \perp\!\!\!\perp u(t)$ for all $s \neq t$, and $E[u(t)] = 0$ for all t

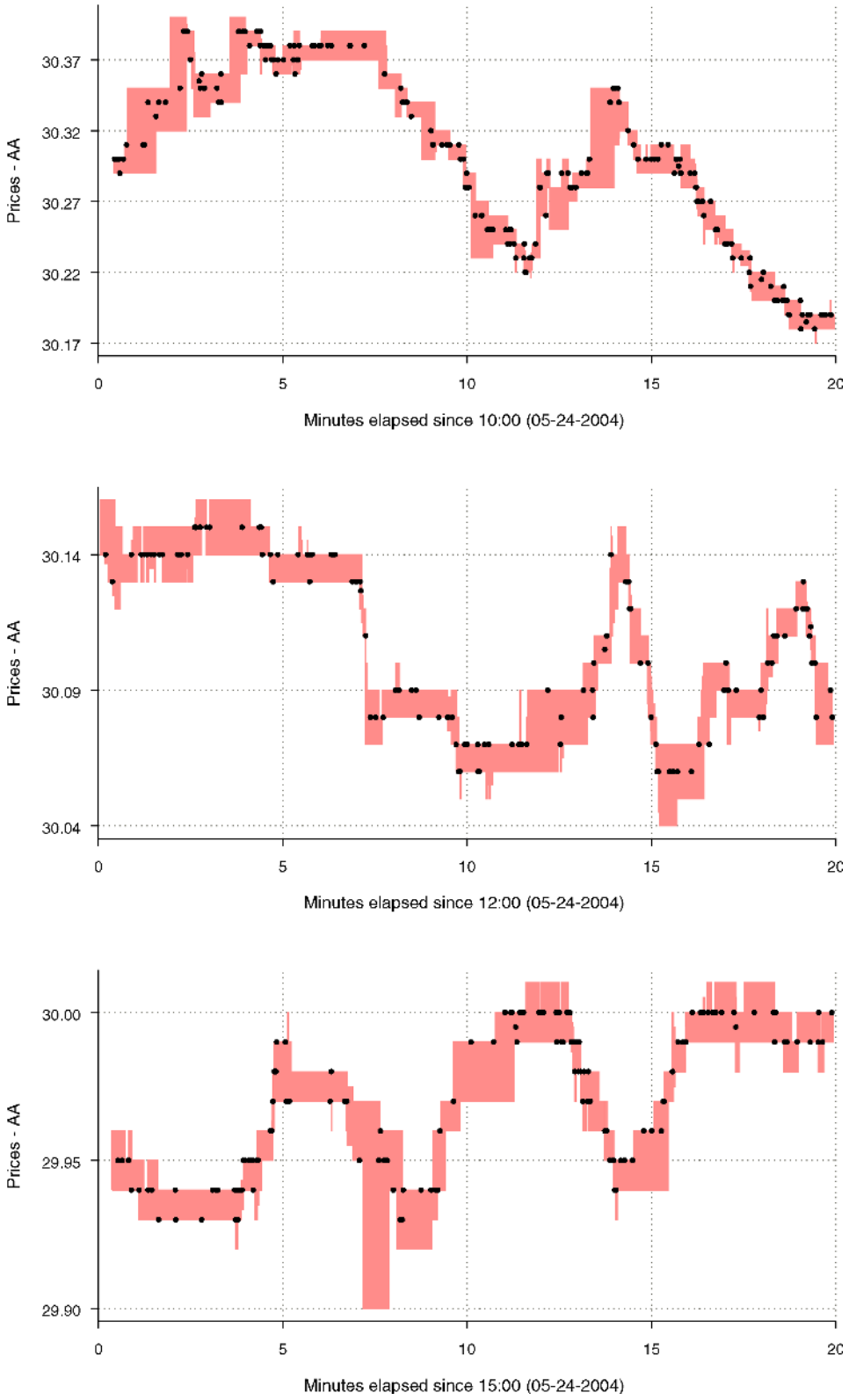

Figure 2. Bid and Ask Quotes (defined by the shaded area) and Actual Transaction Prices (●) Over Three 20-Minute Subperiods on April 24, 2004 for AA.

(b) $\omega^2 \equiv E|u(t)|^2 < \infty$ for all t
(c) $\mu_4 \equiv E|u(t)|^4 < \infty$ for all t.

The independent noise, u, induces an MA(1) structure on the return noise, $e_{i,m}$, which is why this type of noise is sometimes referred to as MA(1) noise. However, $e_{i,m}$ has a very particular MA(1) structure, because it has a unit root. Thus the MA(1) label does not fully characterize the properties of the noise. This is why we prefer to call this type of noise *independent noise*.

Some of the results that we formulate in this section only rely on Assumption 3(a), so we require only (b) and (c) to hold when necessary. Note that ω^2, which is defined in (b), corresponds to $\pi(0)$ in our previous notation. To simplify some of our subsequent expressions, we define the "excess kurtosis ratio", $\kappa \equiv \mu_4/(3\omega^4)$, and note that Assumption 3 is satisfied if u is a Gaussian "white noise" process, $u(t) \sim \mathrm{N}(0, \omega^2)$, in which case $\kappa = 1$.

The existence of a noise process, u, that satisfies Assumption 3, follows directly from Kolmogorov's existence theorem (see Billingsley 1995, chap. 7). It is worthwhile to note that "white noise processes in continuous time" are very erratic processes. In fact, the quadratic variation of a white noise process is unbounded (as is the *r-tic variation* for any other integer). Thus the "realized variance" of a white noise process diverges to infinity in probability as the sampling frequency, m, is increased. This is in stark contrast to the situation for Brownian-type processes that have finite r-tic variation for $r \geq 2$ (see Barndorff-Nielsen and Shephard 2003).

Lemma 2. Given Assumptions 1 and 3(a) and (b), we have that $E(RV^{(m)}) = IV + 2m\omega^2$; if Assumption 3(c) also holds, then

$$\mathrm{var}\big(RV^{(m)}\big) = \kappa 12\omega^4 m + 8\omega^2 \sum_{i=1}^{m} \sigma_{i,m}^2 - (6\kappa - 2)\omega^4 + 2\sum_{i=1}^{m} \sigma_{i,m}^4 \quad (2)$$

and

$$\frac{RV^{(m)} - 2m\omega^2}{\sqrt{\kappa 12\omega^4 m}} = \sqrt{\frac{m}{3\kappa}}\left(\frac{RV^{(m)}}{2m\omega^2} - 1\right) \xrightarrow{d} \mathrm{N}(0, 1), \quad \text{as } m \to \infty.$$

Here we "$\xrightarrow{d}$" to denote convergence in distribution. Thus unlike the situation in Corollary 1, where the noise is time-dependent and the asymptotic bias is finite [whenever $\pi'(0)$ is finite], this situation with independent market microstructure noise leads to a bias that diverges to infinity. This result was first derived in an unpublished thesis by Fang (1996). The expression for the variance [see (2)] is due to Bandi and Russell (2005) and Zhang et al. (2005); the former expressed (2) in terms of the moments of the return noise, $e_{i,m}$.

In the absence of market microstructure noise and under CTS [$\omega^2 = 0$ and $\delta_{i,m} = (b-a)/m$], we recognize a result of Barndorff-Nielsen and Shephard (2002) that

$$\mathrm{var}\big(RV^{(m)}\big) = 2\sum_{i=1}^{m} \sigma_{i,m}^4 = 2\frac{b-a}{m}\int_a^b \sigma^4(s)\,ds + o\left(\frac{1}{m}\right),$$

where $\int_a^b \sigma^4(s)\,ds$ is known as the *integrated quarticity*, introduced by Barndorff-Nielsen and Shephard (2002).

Next, we consider the estimator of Zhou (1996) given by

$$RV_{\mathrm{AC}_1}^{(m)} \equiv \sum_{i=1}^{m} y_{i,m}^2 + \sum_{i=1}^{m} y_{i,m}y_{i-1,m} + \sum_{i=1}^{m} y_{i,m}y_{i+1,m}. \quad (3)$$

This estimator incorporates the empirical first-order autocovariance, which amounts to a bias correction that "works" in much the same way that robust covariance estimators, such as that of Newey and West (1987), achieve their consistency. Note that (3) involves $y_{0,m}$ and $y_{m+1,m}$, which are intraday returns outside the interval $[a, b]$. If these two intraday returns are unavailable, then one could simply use the estimator $\sum_{i=2}^{m-1} y_{i,m}^2 + \sum_{i=2}^{m} y_{i,m}y_{i-1,m} + \sum_{i=1}^{m-1} y_{i,m}y_{i+1,m}$ that estimates $\int_{a+\delta_{1,m}}^{b-\delta_{m,m}} \sigma^2(s)\,ds = IV + O(\frac{1}{m})$. Here we follow Zhou (1996) and use the formulation in (3) because it simplifies the analysis and several expressions. Our empirical implementation is based on a version that does not rely on intraday returns outside the $[a, b]$ interval. We describe the exact implementation in Section 5.

Next, we formulate results for $RV_{\mathrm{AC}_1}^{(m)}$ that are similar to those for $RV^{(m)}$ in Lemma 2.

Lemma 3. Given Assumptions 1 and 3(a), we have that $E(RV_{\mathrm{AC}_1}^{(m)}) = IV$. If Assumption 3(b) also holds, then

$$\mathrm{var}\big(RV_{\mathrm{AC}_1}^{(m)}\big) = 8\omega^4 m + 8\omega^2 \sum_{i=1}^{m} \sigma_{i,m}^2 - 6\omega^4 + 6\sum_{i=1}^{m} \sigma_{i,m}^4 + O(m^{-2})$$

under CTS and BTS, and

$$\frac{RV_{\mathrm{AC}_1}^{(m)} - IV}{\sqrt{8\omega^4 m}} \xrightarrow{d} \mathrm{N}(0, 1), \quad \text{as } m \to \infty.$$

An important result of Lemma 3 is that $RV_{\mathrm{AC}_1}^{(m)}$ is unbiased for the IV at any sampling frequency, m. Also note that Lemma 3 requires slightly weaker assumptions than those needed for $RV^{(m)}$ in Lemma 2. The first result relies on only Assumption 3(a); (c) is not needed for the variance expression. This is achieved because the expression for $RV_{\mathrm{AC}_1}^{(m)}$ can be rewritten in a way that does not involve squared noise terms, $u_{i,m}^2$, $i = 1, \ldots, m$, as does the expression for $RV^{(m)}$, where $u_{i,m} \equiv u(t_{i,m})$. A somewhat remarkable result of Lemma 3 is that the bias-corrected estimator, $RV_{\mathrm{AC}_1}^{(m)}$, has a smaller asymptotic variance (as $m \to \infty$) than the unadjusted estimator, $RV^{(m)}$ ($8m\omega^4$ vs. $12\kappa m\omega^4$). Usually, bias correction is accompanied by a larger asymptotic variance. Also note that the asymptotic results of Lemma 3 are somewhat more useful than those of Lemma 2 (in terms of estimating IV), because the results of Lemma 2 do not involve the object of interest, IV, but shed light only on aspects of the noise process. This property was used by Bandi and Russell (2005) and Zhang et al. (2005) to estimate ω^2; we discuss this aspect in more detail in our empirical analysis in Section 5. It is important to note that the asymptotic results of Lemma 3 do not suggest that $RV_{\mathrm{AC}_1}^{(m)}$ should be based on intraday returns sampled at the highest possible frequency, because the asymptotic variance is increasing in m! Thus we could drop IV from the quantity that converges in distribution to N(0, 1) and simply write $RV_{\mathrm{AC}_1}^{(m)}/\sqrt{8\omega^4 m} \xrightarrow{d} \mathrm{N}(0, 1)$. In other words, whereas $RV_{\mathrm{AC}_1}^{(m)}$ is "centered" about the object of interest, IV, it is unlikely to be close to IV as $m \to \infty$.

In the absence of market microstructure noise ($\omega^2 = 0$), we note that

$$\operatorname{var}\big[RV^{(m)}_{\mathrm{AC}_1}\big] \approx 6\sum_{i=1}^{m} \sigma^4_{i,m},$$

which shows that the variance of $RV^{(m)}_{\mathrm{AC}_1}$ is about three times larger than that of $RV^{(m)}$ when $\omega^2 = 0$. Thus, in the absence of noise, we see an increase in the asymptotic variance as a result of the bias correction. Interestingly, this increase in the variance is identical to that of the maximum likelihood estimator in a Gaussian specification, where $\sigma^2(s)$ is constant and $\omega^2 = 0$ (see Aït-Sahalia et al. 2005a).

It is easy to show that $\tau_i^* = c/m$, $i = 1, \ldots, m$, solves the following constrained minimization problem:

$$\min_{\tau_1,\ldots,\tau_m} \sum_{i=1}^{m} \tau_i^2 \quad \text{subject to} \quad \sum_{i=1}^{m} \tau_i = c.$$

Thus, if we set $\tau_i = \sigma^2_{i,m}$ and $c = IV$, then we see that $\sum_{i=1}^m \sigma^4_{i,m}$ (for fixed m) is minimized under BTS. This highlights one of the advantages of BTS over CTS. This result was shown to hold in a related (pure jump) framework by Oomen (2005). In the present context, we have that, under BTS $\sum_{i=1}^m \sigma^4_{i,m} = IV^2/m$, and specifically it holds that

$$\frac{IV^2}{m} \le \int_a^b \sigma^4(s)\,ds\,\frac{b-a}{m}.$$

The variance expression under CTS [$\delta_{i,m} = (b-a)/m$] is approximately given by

$$\operatorname{var}\big[RV^{(m)}_{\mathrm{AC}_1}\big] \approx 8\omega^4 m + 8\omega^2 \int_a^b \sigma^2(s)\,ds - 6\omega^4 + 6\frac{b-a}{m}\int_a^b \sigma^4(s)\,ds.$$

Next, we compare $RV^{(m)}_{\mathrm{AC}_1}$ and $RV^{(m)}$ in terms of their MSEs and their respective optimal sampling frequencies for a special case that reveals key features of the two estimators.

Corollary 2. Define $\lambda \equiv \omega^2/IV$, suppose that $\kappa = 1$, and let $t_{0,m}, \ldots, t_{m,m}$ be such that $\sigma^2_{i,m} = IV/m$ (BTS). The MSEs are given by

$$\operatorname{MSE}\big(RV^{(m)}\big) = IV^2\left[4\lambda^2 m^2 + 12\lambda^2 m + 8\lambda - 4\lambda^2 + 2\frac{1}{m}\right] \quad (4)$$

and

$$\operatorname{MSE}\big(RV^{(m)}_{\mathrm{AC}_1}\big) = IV^2\left[8\lambda^2 m + 8\lambda - 6\lambda^2 + 6\frac{1}{m} - 2\frac{1}{m^2}\right]. \quad (5)$$

The optimal sampling frequencies for $RV^{(m)}$ and $RV^{(m)}_{\mathrm{AC}_1}$ are given implicitly as the real (positive) solutions to $4\lambda^2 m^3 + 6\lambda^2 m^2 - 1 = 0$ and $4\lambda^2 m^3 - 3m + 2 = 0$.

We denote the optimal sampling frequencies for $RV^{(m)}$ and $RV^{(m)}_{\mathrm{AC}_1}$ by m_0^* and m_1^*. These are approximately given by

$$m_0^* \approx (2\lambda)^{-2/3} \qquad \text{and} \qquad m_1^* \approx \sqrt{3}(2\lambda)^{-1}.$$

The expression for m_0^* was derived in Bandi and Russell (2005) and Zhang et al. (2005) under more general conditions than those used in Corollary 2, whereas the expression for m_1^* was derived earlier by Zhou (1996).

In our empirical analysis, we often find that $\lambda \le 10^{-3}$, such that

$$m_1^*/m_0^* \approx 3^{1/2}2^{-1/3}(\lambda^{-1})^{1/3} \ge 10,$$

which shows that m_1^* is several times larger than m_0^* when the noise-to-signal is as small as we find it to be in practice. In other words, $RV^{(m)}_{\mathrm{AC}_1}$ permits more frequent sampling than does the "optimal" RV. This is quite intuitive, because $RV^{(m)}_{\mathrm{AC}_1}$ can use more information in the data without being affected by a severe bias. Naturally, when TTS is used, the number of intraday returns, m, cannot exceed the total number of transactions/quotations, so in practice it might not be possible to sample as frequently as prescribed by m_1^*. Furthermore, these results rely on the independent noise assumption, which may not hold at the highest sampling frequencies.

Corollary 2 captures the salient features of this problem and characterizes the MSE properties of $RV^{(m)}$ and $RV^{(m)}_{\mathrm{AC}_1}$ in terms of a single parameter, λ (noise-to-signal). Thus the simplifying assumptions of Corollary 2 yield an attractive framework for comparing $RV^{(m)}$ and $RV^{(m)}_{\mathrm{AC}_1}$ and for analyzing their (lack of) robustness to market microstructure noise.

From Corollary 2, we note that the RMSEs of $RV^{(m)}$ and $RV^{(m)}_{\mathrm{AC}_1}$ are proportional to the IV and given by $r_0(\lambda, m)IV$ and $r_1(\lambda, m)IV$, where

$$r_0(\lambda, m) \equiv \sqrt{4\lambda^2 m^2 + 12\lambda^2 m + 8\lambda - 4\lambda^2 + \frac{2}{m}}$$

and

$$r_1(\lambda, m) \equiv \sqrt{8\lambda^2 m + 8\lambda - 6\lambda^2 + \frac{6}{m} - \frac{2}{m^2}}.$$

Figure 3 plots $r_0(\lambda, m)$ and $r_1(\lambda, m)$ using empirical estimates of λ. The estimates are based on high-frequency stock returns of Alcoa (left panels) and Microsoft (right panels) in year the 2000. The details about the estimation of λ are deferred to Section 5. The upper panels present $r_0(\hat{\lambda}, m)$ and $r_1(\hat{\lambda}, m)$, where the x-axis is $\delta_{i,m} = (b-a)/m$ in units of seconds. For both equities, we note that the $RV^{(m)}_{\mathrm{AC}_1}$ dominates the $RV^{(m)}$ except at the very lowest frequencies. The minimums of $r_0(\hat{\lambda}, m)$ and $r_1(\hat{\lambda}, m)$ identify their respective optimal sampling frequencies, m_0^* and m_1^*. For the AA returns, we find that the optimal sampling frequencies are $m^*_{0,\mathrm{AA}} = 44$ and $m^*_{1,\mathrm{AA}} = 511$ (corresponding to intraday returns spanning 9 minutes and 46 seconds) and that the theoretical reduction of the RMSE is 33.1%. The curvatures of $r_0(\hat{\lambda}, m)$ and $r_1(\hat{\lambda}, m)$ in the neighborhood of m_0^* and m_1^* show that $RV^{(m)}_{\mathrm{AC}_1}$ is less sensitive than $RV^{(m)}$ to the choice of m.

The middle panels of Figure 3 display the relative RMSE of $RV^{(m)}_{\mathrm{AC}_1}$ to that of (the optimal) $RV^{(m_0^*)}$ and the relative RMSE of $RV^{(m)}$ to that of (the optimal) $RV^{(m_1^*)}_{\mathrm{AC}_1}$. These panels show that the $RV^{(m)}_{\mathrm{AC}_1}$ continues to dominate the "optimal" $RV^{(m_0^*)}$ for a wide ranges of frequencies, not just in a small neighborhood of the optimal value, m_1^*. This robustness of RV_{AC_1} is quite useful in practice, where λ and (hence) m_1^* are not known with certainty. The result shows that a reasonably precise estimate

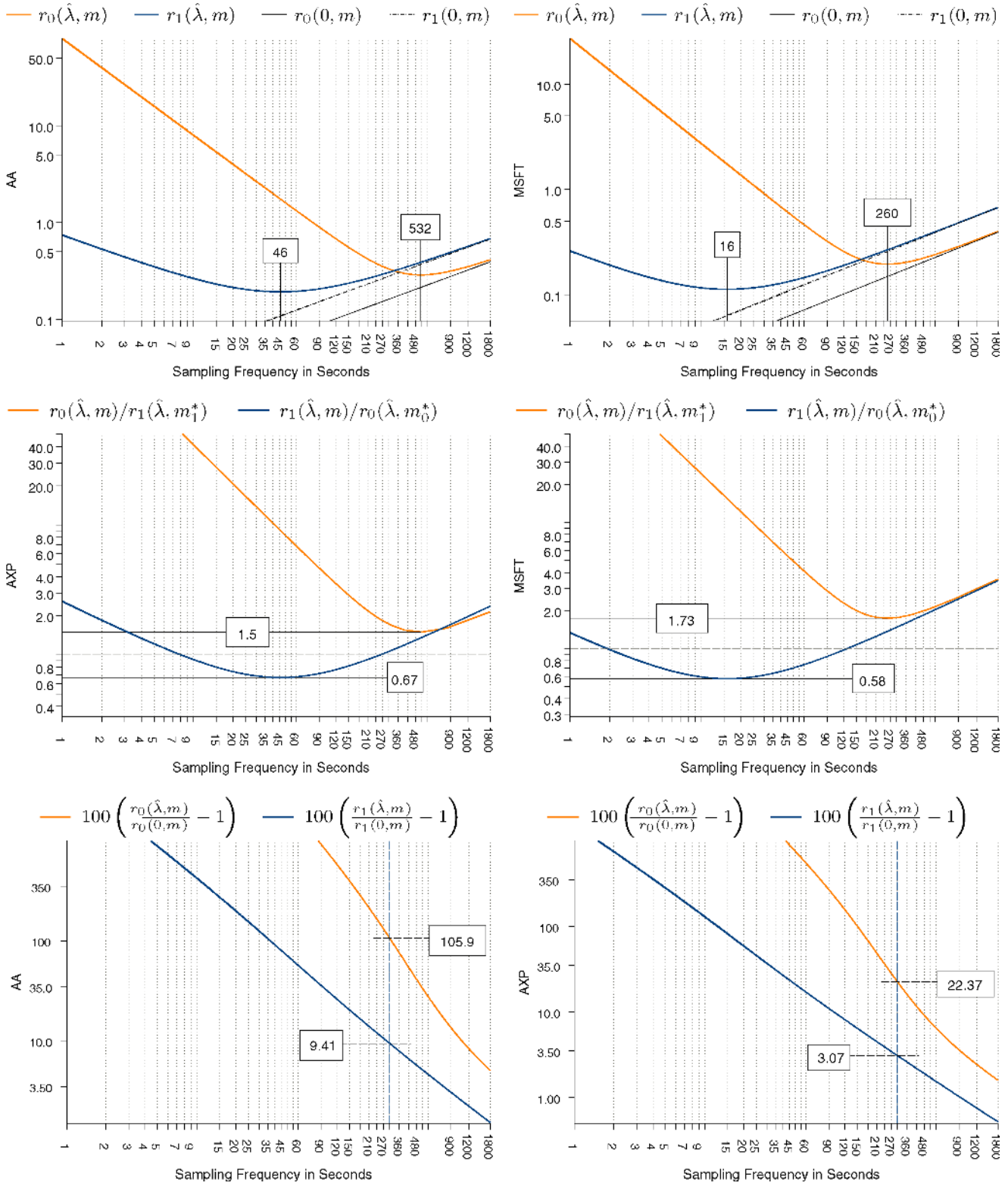

Figure 3. RMSE Properties of RV and RV_{AC_1} Under Independent Market Microstructure Noise Using Empirical Estimates of λ for 2000. The upper panels display the RMSEs for RV and RV_{AC_1} using estimates of λ, $r_0(\hat{\lambda}, m)$ and $r_1(\hat{\lambda}, m)$, and the corresponding RMSEs in the absence of noise, $r_0(0, m)$ and $r_1(0, m)$. The middle panels are the relative RMSEs of $RV^{(m)}$ and $RV_{AC_1}^{(m)}$ to $RV^{(m_0^*)}$ and $RV_{AC_1}^{(m_1^*)}$, as defined by $r_0(\hat{\lambda}, m)/r_1(\hat{\lambda}, m_1^*)$ and $r_1(\hat{\lambda}, m)/r_0(\hat{\lambda}, m_0^*)$. The lower panels show the percentage increase in the RMSE for different sampling frequencies caused by market microstructure noise. The x-axis gives the sampling frequency of intraday returns as defined by $\delta_{i,m} = (b - a)/m$ in units of seconds, where $b - a = 6.5$ hours (a trading day).

of λ (and hence m_1^*) will lead to a RV_{AC_1} that dominates RV. This result is not surprising, because recent developments in this literature have shown that it is possible to construct kernel-based estimators that are even more accurate than RV_{AC_1} (see Barndorff-Nielsen et al. 2004; Zhang 2004).

A second, very interesting aspect that can be analyzed based on the results of Corollary 2 is the accuracy of theoretical results derived under the assumption that $\lambda = 0$ (no market microstructure noise). For example, the accuracy of a confidence interval for IV, which is based on asymptotic results that ignore the presence of noise, will depend on λ and m. The expressions of Corollary 2 provide a simple way to quantify the theoretical accuracy of such confidence intervals, including those of Barndorff-Nielsen and Shephard (2002). Figure 3 provides valuable information on this question. The upper panels of Figure 3 present the RMSEs of $RV^{(m)}$ and $RV^{(m)}_{\mathrm{AC}_1}$, using both $\hat{\lambda} > 0$ (the case with noise) and $\lambda = 0$ (the case without noise). For small values of m, we see that $r_0(\hat{\lambda}, m) \approx r_0(0, m)$ and $r_1(\hat{\lambda}, m) \approx r_1(0, m)$, whereas the effects of market microstructure noise are pronounced at the higher sampling frequencies. The lower panels of Figure 3 quantify the discrepancy between the two "types" of RMSEs as a function of the sampling frequency. These plots present $100[r_0(\hat{\lambda}, m) - r_0(0, m)]/r_0(0, m)$ and $100[r_1(\hat{\lambda}, m) - r_1(0, m)]/r_1(0, m)$ as a function of m. Thus the former reveals the percentage increase of the RV's RMSE due to market microstructure noise, and the second line similarly shows the increase of the RV_{AC_1}'s RMSE due to noise. The increase in the RMSE may be translated into a widening of a confidence intervals for IV (about $RV^{(m)}$ or $RV^{(m)}_{\mathrm{AC}_1}$). The vertical lines in the right panels mark the sampling frequency corresponding to 5-minute sampling under CTS and show that the "actual" confidence interval (based on $RV^{(m)}$) is 105.94% larger than the "no-noise" confidence interval for AA, whereas the enlargement is 22.37% for MSFT. At 20-minute sampling, the discrepancy is less than a couple of percent, so in this case the size distortion from being oblivious to market microstructure noise is quite small. The corresponding increases in the RMSE of $RV^{(m)}_{\mathrm{AC}_1}$ are 9.41% and 3.07%. Thus a "no-noise" confidence interval about $RV^{(m)}_{\mathrm{AC}_1}$ is more reliable than that about $RV^{(m)}$ at moderate sampling frequencies. Here we have used an estimator of λ based on data from the year 2000, before the tick size was reduced to 1 cent. In our empirical analysis we find the noise to be much smaller in recent years, such that "no-noise approximations" are likely to be more accurate after decimalization of the tick size.

Figure 4 presents the volatility signature plots for $RV^{(m)}_{\mathrm{AC}_1}$, where we have used the same scale as in Figure 1. When sampling in calender time (the four upper panels), we see a pronounced bias in $RV^{(m)}_{\mathrm{AC}_1}$ when intraday returns are sampled more frequently than every 30 seconds. The main explanation for this is that CTS will sample the same price multiple times when m is large, which induces (artificial) autocorrelation in intraday returns. Thus, when intraday returns are based on CTS, it is necessary to incorporate higher-order autocovariances of $y_{i,m}$ when m becomes large. The plots in rows 3 and 4 are signature plots when intraday returns are sampled in tick time. These also reveal a bias in $RV^{(x \text{ tick})}_{\mathrm{AC}_1}$ at the highest frequencies, which shows that the noise is time dependent in tick time. For example, the MSFT 2000 plot suggests that the time dependence lasts for 30 ticks, perhaps longer.

Fact II. The noise is autocorrelated.

We provide additional evidence of this fact, based on other empirical quantities, in the following sections.

4. THE CASE WITH DEPENDENT NOISE

In this section we consider the case where the noise is time-dependent and possibly correlated with the efficient returns, $y^*_{i,m}$. Following earlier versions of the present article, issues related to time dependence and noise–price correlation have been addressed by others, including Aït-Sahalia et al. (2005b), Frijns and Lehnert (2004), and Zhang (2004). The time scale of the dependence in the noise plays a role in the asymptotic analysis. Although the "clock" at which the memory in the noise decays can follow any time scale, it seems reasonable for it to be tied to calendar time, tick time, or a combination of the two. We first consider a situation where the time dependence is specific to calendar time, then consider the case with time dependence in tick time.

4.1 Dependence in Calender Time

To bias-correct the RV under the general time-dependent type of noise, we make the following assumption about the time dependence in the noise process.

Assumption 4. The noise process has finite dependence in the sense that $\pi(s) = 0$ for all $s > \theta_0$ for some finite $\theta_0 \geq 0$, and $E[u(t)|p^*(s)] = 0$ for all $|t - s| > \theta_0$.

The assumption is trivially satisfied under the independent noise assumption used in Section 3. A more interesting class of noise processes with finite dependence are those of the moving average type, $u(t) = \int_{t-\theta_0}^{t} \psi(t - s)\, dB(s)$, where $B(s)$ represents a standard Brownian motion and $\psi(s)$ is a bounded (nonrandom) function on $[0, \theta_0]$. The autocorrelation function for a process of this kind is given by $\pi(s) = \int_s^{\theta_0} \psi(t)\psi(t - s)\, dt$, for $s \in [0, \theta_0]$.

Theorem 2. Suppose that Assumptions 1, 2, and 4 hold and let q_m be such that $q_m/m > \theta_0$. Then (under CTS),

$$E\big(RV^{(m)}_{\mathrm{AC}_{q_m}} - IV\big) = 0,$$

where

$$RV^{(m)}_{\mathrm{AC}_{q_m}} \equiv \sum_{i=1}^{m} y^2_{i,m} + \sum_{h=1}^{q_m} \sum_{i=1}^{m} (y_{i-h,m} y_{i,m} + y_{i,m} y_{i+h,m}).$$

A drawback of $RV^{(m)}_{\mathrm{AC}_{q_m}}$ is that it may produce a negative estimate of volatility, because the covariances are not scaled downward in a way that would guarantee positivity. This is particularly relevant in the situation where intraday returns have a "sharp negative autocorrelation" (see West 1997), which has been observed in high-frequency intraday returns constructed from transaction prices. To rule out the possibility of a negative estimate, one could use a different kernel, such as the Bartlett kernel. Although a different kernel may not be entirely unbi-

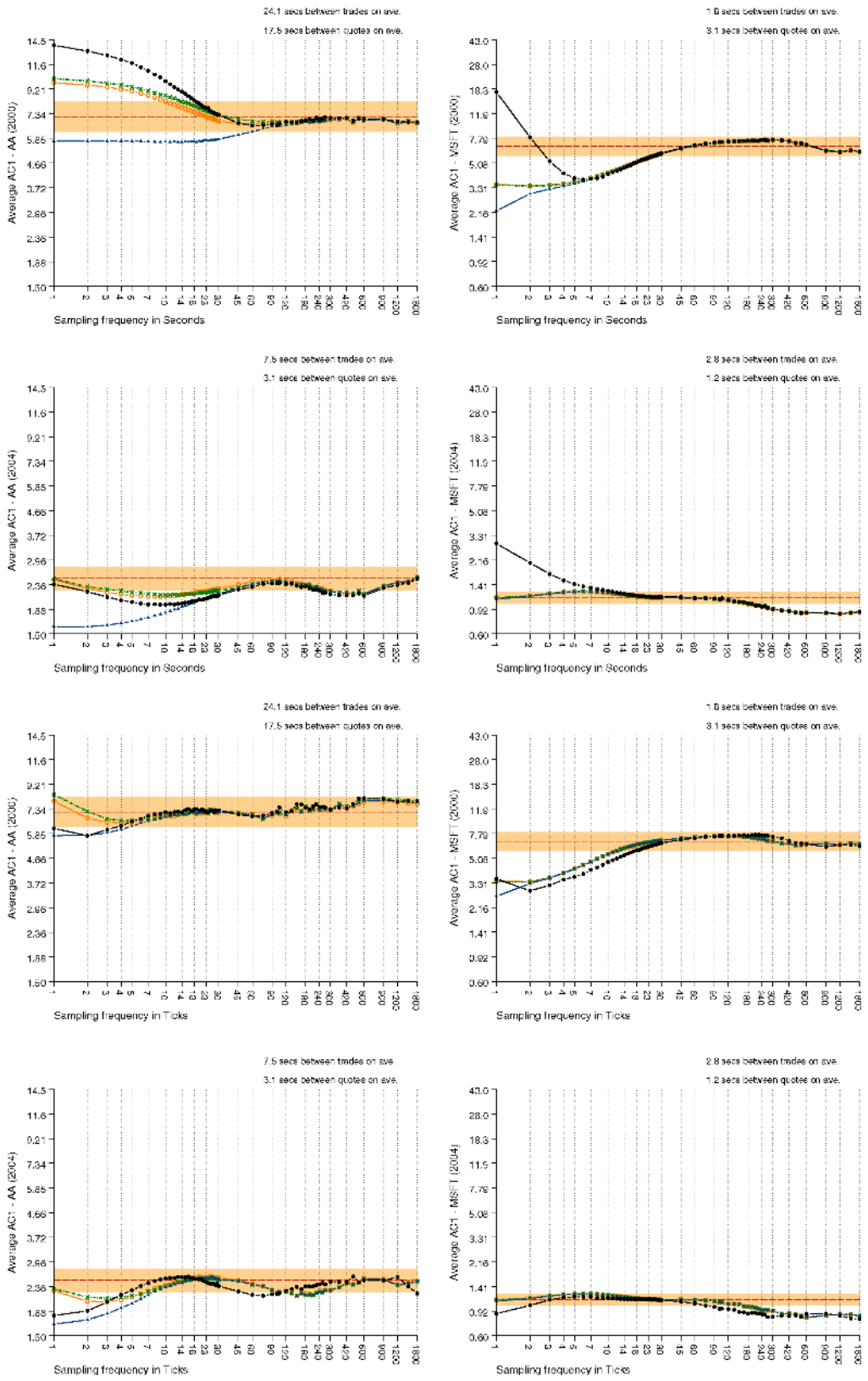

Figure 4. Volatility Signature Plots for RV_{AC_1} Based on Ask Quotes (▫–▫–▫), Bid Quotes (×–×–×), Mid-Quotes (▲–▲–▲), and Transaction Prices (▪–●–▪). The left column is for AA and the right column is for MSFT. The two top rows are based on calendar time sampling; the bottom rows are based on tick time sampling. The results for 2000 are the panels in rows 1 and 3, and those for 2004 are in rows 2 and 4. The horizontal line represents an estimate of the average IV, $\bar{\sigma}^2 \equiv \overline{RV}^{(1\ tick)}_{ACNW_{30}}$, that is defined in Section 4.2. The shaded area about $\bar{\sigma}^2$ represents an approximate 95% confidence interval for the average volatility.

ased, it may result is a smaller MSE than that of RV_{AC}. Interestingly, Barndorff-Nielsen et al. (2004) have shown that the subsample estimator of Zhang et al. (2005) is almost identical to the Bartlett kernel estimator.

In the time series literature, the lag length, q_m, is typically chosen such that $q_m/m \to 0$ as $m \to \infty$, for example, $q_m = \lceil 4(m/100)^{2/9} \rceil$, where $\lceil x \rceil$ denotes the smallest integer that is greater than or equal to x. But if the noise were dependent in calendar time, then this would be inappropriate, because it would lead to $q_m = 3$ when a typical trading day (390 minutes) were divided into 78 intraday returns (5-minute returns) and to $q_m = 6$ if the day were divided into 780 intraday returns (30-second returns). So the former q would cover 15 minutes, whereas the latter would cover 3 minutes (6×30 seconds), and in fact the period would shrink to 0 as $m \to \infty$. Under Assumptions 2 and 4, the autocorrelation in intraday returns is specific to a period in calendar time, which does not depend on m; thus it is more appropriate to keep the width of the "autocorrelation window," q_m/m, constant. This also makes $RV^{(m)}_{\mathrm{AC}}$ more comparable across different frequencies, m. Thus we set $q_m = \lceil \frac{w}{(b-a)/m} \rceil$, where w is the desired width of the lag window and $b - a$ is the length of the sampling period (both in units of time), such that $(b - a)/m$ is the period covered by each intraday return. In this case we write $RV^{(m)}_{\mathrm{AC}_w}$ in place of $RV^{(m)}_{\mathrm{AC}_{qm}}$. Therefore, if we were to sample in calendar time and set $w = 15$ min and $b - a = 390$ min, then we would include $q_m = \lceil m/26 \rceil$ autocovariance terms.

When q_m is such that $q_m/m > \theta_0 \geq 0$, this implies that $RV^{(m)}_{\mathrm{AC}_{qm}}$ cannot be consistent for IV. This property is common for estimators of the long-run variance in the time series literature whenever q_m/m does not converge to 0 sufficiently fast (see, e.g., Kiefer, Vogelsang, and Bunzel 2000; and Jansson 2004). The lack of consistency in the present context can be understood without consideration of market microstructure noise. In the absence of noise, we have that $\mathrm{var}(y^2_{i,m}) = 2\sigma^4_{i,m}$ and $\mathrm{var}(y_{i,m}y_{i+h,m}) = \sigma^2_{i,m}\sigma^2_{i+h,m} \approx \sigma^4_{i,m}$, such that

$$\mathrm{var}\big[RV^{(m)}_{\mathrm{AC}_{q_m}}\big] \approx 2\sum_{i=1}^{m}\sigma^4_{i,m} + \sum_{h=1}^{q_m}(2)^2\sum_{i=1}^{m}\sigma^4_{i,m} = 2(1+2q_m)\sum_{i=1}^{m}\sigma^4_{i,m},$$

which approximately equals

$$2(1+2q_m)\frac{b-a}{m}\int_0^1 \sigma^4(s)\,ds$$

under CTS. This shows that the variance does not vanish when q_m is such that $q_m/m > \theta_0 > 0$.

The upper four panels of Figure 5 represent a new type of signature plots for $RV^{(1\text{ sec})}_{\mathrm{AC}_q}$. Here we sample intraday returns every second using the previous-tick method and now plot q along the x-axis. Thus these signature plots provide information on time dependence in the noise process. The fact that the $RV^{(1\text{ sec})}_{\mathrm{AC}_q}$ of the four price series differ and have not leveled off is evidence of time dependence. Thus in the upper four panels, where we sample in calendar time, it appears that the dependence lasts for as long as 2 minutes (AA, year 2000) or as short as 15 seconds (MSFT, year 2004). We comment on the lower four panels in the next section, where we discuss intraday returns sampled in tick time.

4.2 Time Dependence in Tick Time

When sampling at ultra-high frequencies, we find it more natural to sample in tick time, such that the same observation is not sampled multiple times. Furthermore, the time dependence in the noise process may be in tick time rather than calender time. Several results of Bandi and Russell (2005) allow for time dependence in tick time (while the price–noise correlation is assumed away).

The following example gives a situation with market microstructure noise that is time-dependent in tick time and correlated with efficient returns.

Example 1. Let $t_0 < t_1 < \cdots < t_m$ be the times at which prices are observed, and consider the case where we sample intraday returns at the highest possible frequency in tick time. We suppress the subscript m to simplify the notation. Suppose that the noise is given by $u_i = \alpha y^*_i + \varepsilon_i$, where ε_i is a sequence of iid random variables with mean 0 and variance $\mathrm{var}(\varepsilon_i) = \omega^2$. Thus $\alpha = 0$ corresponds to the case with independent noise assumption, and $\alpha = \omega^2 = 0$ corresponds to the case without noise. It now follows that

$$e_i = \alpha(y^*_i - y^*_{i-1}) + \varepsilon_i - \varepsilon_{i-1},$$

such that

$$E(e^2_i) = \alpha^2(\sigma^2_i + \sigma^2_{i-1}) + 2\omega^2$$

and

$$E(e_i y^*_i) = \alpha\sigma^2_i,$$

where $\sum_{i=1}^m \sigma^2_i = IV$. Thus

$$E\big[RV^{(1\text{ tick})}\big] = IV + 2\alpha(1+\alpha)IV + 2m\omega^2,$$

with a bias given by $2\alpha(1+\alpha)IV + 2m\omega^2$. This bias may be negative if $\alpha < 0$ (the case where u_i and y^*_i are negatively correlated). Now, we also have

$$E(e_i e_{i-1}) = -\alpha^2\sigma^2_{i-1} - \omega^2$$

and

$$E(e_i y^*_{i-1}) = -\alpha\sigma^2_{i-1},$$

such that

$$\sum_{i=1}^{m} E[y_i y_{i+1}] = -\alpha^2 IV - 2m\omega^2 - \alpha IV,$$

which shows that $RV^{(1\text{ tick})}_{\mathrm{AC}_1}$ is almost unbiased for the IV.

In this simple example, u_i is only contemporaneously correlated with y^*_i. In practice, it is plausible that u_i could also be correlated with lagged values of y^*_i, which would yield a more complicated time dependence in tick time. In this situation we could use $RV^{(1\text{ tick})}_{\mathrm{AC}_q}$, with a q sufficiently large to capture the time dependence.

Assumption 4 and Theorem 2 are formulated for the case with CTS, but a similar estimator can be defined under dependence in tick time. The lower four panels of Figure 5 are the

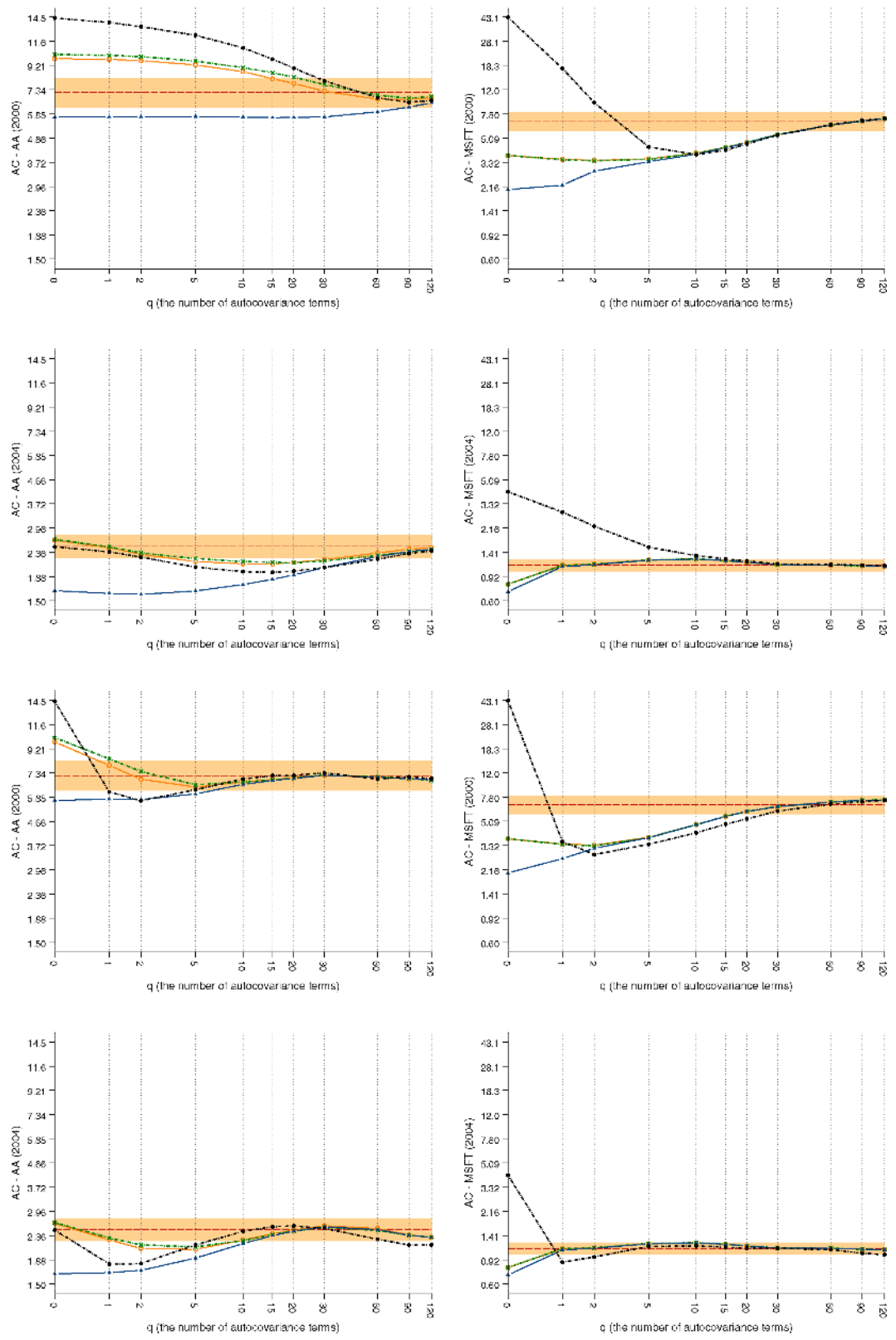

Figure 5. Volatility Signature Plots of $RV^{(1\ sec)}_{AC_q}$ (four upper panels) and $RV^{(1\ tick)}_{AC_q}$ (four lower panels) for Each of the Price Series: Ask Quotes (▫–▫–▫), Bid Quotes (×–×–×), Mid-Quotes (▲–▲–▲), and Transaction Prices (■–●–■). The x-axis is the number of autocovariance terms, q, included in $RV^{(1\ sec)}_{AC_q}$. The left column is for AA, and the right column is for MSFT. The results for 2000 are the panels in rows 1 and 3, and those for 2004 are in rows 2 and 4. The horizontal line represents an estimate of the average IV, $\bar{\sigma}^2 \equiv \overline{RV}^{(1\ tick)}_{ACNW_{30}}$, that is defined in Section 4.2. The shaded area about $\bar{\sigma}^2$ represents an approximate 95% confidence interval for the average volatility.

signature plots for $RV^{(1\text{ tick})}_{\text{AC}_q}$, where q is the number of autocovariances used to bias correct the standard RV. From these plots, we see that a correction for the first couple of autocovariances has a substantial impact on the estimator, but higher-order autocovariances are also important, because the volatility signature plots do not stabilize until $q \geq 30$ in some cases (e.g., MSFT in 2000). This time dependence was longer than we had anticipated; thus we examined whether a few "unusual" days were responsible for this result. However, the upward-sloping volatility signature (until q is about 30) is actually found in most daily plots of $RV^{(1\text{ tick})}_{\text{AC}_q}$ against q for MSFT in the year 2000.

In Section 3 we analyzed the simple kernel estimator that incorporates only the first-order autocovariance of intraday returns, which we now generalize by including higher-order autocovariances. We did this to make the estimator, $RV^{(1\text{ tick})}_{\text{AC}_q}$, robust to both time dependence in the noise and correlation between noise and efficient returns. Interestingly, Zhou (1996) also proposed a subsample version of this estimator, although he did not refer to it as a subsample estimator. As is the case for $RV^{(1\text{ tick})}_{\text{AC}_q}$, this estimator is robust to time dependence that is finite in tick time. Next, we describe the subsample-based version of Zhou's estimator.

Let $t_0 < t_1 < \cdots < t_N$ be the times at which prices are observed in the interval $[a, b]$, where $a = t_0$ and $b = t_N$. Here m need not be equal to N (unlike the situation in the previous example), because we use intraday returns that span several price observations. We use the following notation for such (skip-k) intraday returns:

$$y_{t_i,t_{i+k}} \equiv p_{t_{i+k}} - p_{t_i}.$$

Thus $y_{t_i,t_{i+k}}$ is the intraday return over the time interval $[t_i, t_{i+k}]$. This leads to the identity

$$RV^{(k\text{ tick})}_{\text{AC}_1} = \sum_{i\in\{0,k,2k,\ldots,N-k\}} \left(y^2_{t_i,t_{i+k}} + y_{t_{i-k},t_i}y_{t_i,t_{i+k}} + y_{t_i,t_{i+k}}y_{t_{i+k},t_{i+2k}}\right)$$

(assuming that N/k is an integer), which is a sum involving $m = N/k$ terms. The subsample version $RV^{(m)}_{\text{AC}_1}$, proposed by Zhou (1996), can be expressed as

$$\frac{1}{k}\sum_{i=0}^{N-1}\left(y^2_{t_i,t_{i+k}} + y_{t_{i-k},t_i}y_{t_i,t_{i+k}} + y_{t_i,t_{i+k}}y_{t_{i+k},t_{i+2k}}\right). \tag{6}$$

Thus, for $k = 2$, we have

$$\frac{1}{2}\sum_{i=1}^{N}(y_i + y_{i+1})^2 + (y_{i-1} + y_{i-2})(y_i + y_{i+1}) + (y_i + y_{i+1})(y_{i+2} + y_{i+3}),$$

where $y_i \equiv y_{t_{i-1},t_i}$. Rearranging the terms, we see that this sum is approximately given by

$$\frac{1}{2}\sum_{i=1}^{N} 2y_i^2 + 4y_iy_{i+1} + 4y_iy_{i+2} + 2y_iy_{i+3} = \hat\gamma_0 + (\hat\gamma_{-1} + \hat\gamma_1) + (\hat\gamma_{-2} + \hat\gamma_2) + \frac{1}{2}(\hat\gamma_{-3} + \hat\gamma_3),$$

where $\hat\gamma_j \equiv \sum_{i=1}^{N} y_iy_{i+j}$. For the general case, $k \geq 2$, we can show that this subsample estimator is approximately given by

$$RV^{(1\text{ tick})}_{\text{ACNW}_k} \equiv \hat\gamma_0 + \sum_{j=1}^{k}(\hat\gamma_{-j} + \hat\gamma_j) + \sum_{j=1}^{k}\frac{k-j}{k}(\hat\gamma_{-j-k} + \hat\gamma_{j+k}).$$

Thus this estimator equals the $RV^{(1\text{ tick})}_{\text{AC}_k}$ plus an additional term, a Bartlett-type weighted sum of higher-order covariances. Interestingly, Zhou (1996) showed that the subsample version of his estimator [see (6)] has a variance that is (at most) of order $O(\frac{k}{N}) + O(\frac{1}{k}) + O(\frac{N}{k^2})$ (assuming constant volatility). This term is of order $O(N^{-1/3})$ if k is chosen to be proportional to $N^{2/3}$, as was done by Zhang et al. (2005). It appears that Zhou may have considered k as fixed in his asymptotic analysis, because he referred to this estimator as being inconsistent (see Zhou 1998, p. 114). Therefore, the great virtues of subsample-based estimators in this context were first recognized by Zhang et al. (2005).

5. EMPIRICAL ANALYSIS

We now analyze stock returns for the 30 equities of the DJIA. The sample period spans 5 years, from January 3, 2000 to December 31, 2004. We report results for each of the years individually, but give some of the more detailed results only for the years 2000 and 2004 to conserve space. The tick size was reduced from 1/16 of 1 dollar to 1 cent on January 29, 2001, and to avoid mixing mix days with different tick sizes, we drop most of the days during January 2001 from our sample. The data are transaction prices and quotations from NYSE and NASDAQ, and all data were extracted from the Trade and Quote (TAQ) database.

We filtered the raw data for outliers, and discarded transactions outside the period 9:30 AM–4:00 PM and removed days with less than 5 hours of trading from the sample. This reduced the sample to the number of days reported in the last column of Table 1. The filtering procedure removed obvious data errors, such as zero prices. We also removed transaction prices that were more than one spread away from the bid and ask quotes. (Details of the filtering procedure are described in a technical appendix available at our website.) The average number of transactions/quotations per day are given for each year in our sample; these reveal a steady increase in the number of transactions and quotations over the 5-year period. The numbers in parentheses are the percentages of transaction prices that differ from the proceeding transaction price and similarly for the quoted prices. The same price is often observed in several consecutive transactions/quotations, because a large trade may be divided into smaller transactions, and a "new" quote may simply reflect a revision of the "depth" while the bid and ask prices remain unchanged. We use all price observations in our analysis. Censoring all of the zero intraday returns does not affect the RV, but has an impact on the autocorrelation of intraday returns.

Our analysis of quotation data is based on bid and ask prices and the average of these (mid-quotes). The RVs are calculated for the hours that the market is open, approximately 390 minutes per day (6.5 hours for most days). Our tables present results for all 30 equities, whereas our figures present

Table 1. Equity Data: Summary Statistics

Symbol	Name	Exchange	Transactions/day					Quotes/day					No. of days
			2000	2001	2002	2003	2004	2000	2001	2002	2003	2004	
AA	Alcoa	NYSE	997(41)	1,479(50)	1,898(50)	2,153(45)	3,150(43)	1,384(33)	1,875(44)	2,775(38)	5,309(32)	7,560(34)	1,223
AXP	American Express	NYSE	1,584(45)	2,449(54)	2,791(53)	3,371(52)	2,752(44)	2,004(42)	3,123(46)	3,745(41)	7,293(43)	7,464(34)	1,221
BA	Boeing Company	NYSE	1,052(39)	1,730(49)	2,306(52)	2,961(48)	3,037(47)	1,587(30)	2,492(46)	3,552(43)	6,475(42)	8,022(39)	1,221
C	Citigroup	NYSE	2,597(44)	3,129(51)	3,997(49)	4,424(46)	4,803(47)	3,076(27)	3,994(42)	5,039(40)	7,993(37)	9,316(37)	1,221
CAT	Caterpillar Inc.	NYSE	747(40)	1,260(50)	1,673(53)	2,414(54)	3,154(56)	1,039(33)	2,002(43)	2,879(40)	5,852(46)	8,185(51)	1,221
DD	Du Pont De Nemours	NYSE	1,257(48)	1,853(54)	2,100(53)	2,963(51)	3,077(49)	1,858(30)	2,915(43)	3,418(39)	6,663(41)	8,069(38)	1,220
DIS	Walt Disney	NYSE	1,304(39)	2,013(53)	2,882(54)	3,448(48)	3,564(46)	1,525(25)	2,893(43)	4,750(36)	7,768(33)	8,480(34)	1,221
EK	Eastman Kodak	NYSE	775(42)	1,128(50)	1,392(45)	2,008(45)	1,941(44)	1,181(35)	1,941(44)	2,322(37)	4,910(35)	5,715(31)	1,220
GE	General Electric	NYSE	3,191(41)	3,157(52)	4,712(54)	4,828(48)	4,771(44)	2,888(34)	3,646(48)	5,559(42)	8,369(31)	10,051(24)	1,221
GM	General Motors	NYSE	988(37)	1,357(50)	2,448(49)	2,874(49)	2,855(47)	1,572(35)	2,009(44)	4,246(37)	6,262(35)	6,889(34)	1,221
HD	Home Depot Inc.	NYSE	1,961(42)	2,648(51)	3,533(52)	3,843(49)	3,642(46)	2,161(31)	2,946(51)	4,181(42)	7,246(36)	8,005(31)	1,221
HON	Honeywell	NYSE	1,122(38)	1,453(52)	1,872(47)	2,482(47)	2,668(47)	1,378(33)	2,364(47)	3,120(40)	5,385(40)	6,542(39)	1,221
HPQ	Hewlett–Packard	NYSE	1,900(46)	2,321(51)	2,543(46)	3,263(43)	3,708(43)	2,394(45)	3,170(43)	3,226(36)	6,536(31)	8,700(27)	1,218
IBM	Int. Business Machines	NYSE	2,322(55)	3,637(63)	3,492(61)	4,111(60)	4,553(55)	3,319(53)	4,729(55)	6,688(37)	7,790(49)	9,447(51)	1,220
INTC	Intel. Corp.	NASD	14,982(73)	15,637(83)	15,194(80)	10,918(71)	9,078(67)	10,599(17)	12,512(32)	17,622(45)	19,554(39)	19,828(42)	1,230
IP	International Paper	NYSE	1,002(40)	1,509(50)	1,971(50)	2,569(47)	2,283(44)	1,368(31)	2,346(41)	3,134(37)	5,997(40)	6,417(34)	1,221
JNJ	Johnson & Johnson	NYSE	1,492(49)	2,059(57)	2,541(60)	3,272(53)	3,853(48)	1,739(36)	2,322(47)	3,091(42)	6,529(39)	8,687(36)	1,221
JPM	J. P. Morgan	NYSE	1,317(52)	2,555(51)	2,973(52)	3,836(49)	3,939(46)	1,839(52)	3,384(46)	4,079(39)	7,387(36)	8,808(30)	1,221
KO	Coca-Cola	NYSE	1,376(45)	1,662(51)	2,349(52)	2,854(50)	3,320(48)	1,635(32)	2,063(48)	3,221(42)	6,240(39)	7,498(35)	1,221
MCD	McDonalds	NYSE	1,109(39)	1,779(50)	2,226(49)	2,638(45)	2,790(43)	1,578(21)	2,334(42)	3,065(37)	5,763(33)	7,331(31)	1,221
MMM	Minnesota Mng. Mfg.	NYSE	868(44)	1,563(56)	2,054(57)	3,092(58)	3,370(48)	1,157(45)	2,462(50)	3,253(48)	7,100(52)	8,228(46)	1,220
MO	Philip Morris	NYSE	1,283(38)	1,942(53)	3,047(54)	3,473(50)	3,404(48)	2,365(13)	3,266(34)	4,174(40)	6,776(38)	7,721(38)	1,220
MRK	Merck	NYSE	1,832(41)	1,943(51)	2,574(52)	3,639(50)	3,663(45)	2,021(35)	2,381(48)	3,262(41)	6,927(43)	7,658(34)	1,221
MSFT	Microsoft	NASD	13,900(68)	14,479(86)	15,257(85)	12,013(73)	8,643(66)	8,275(14)	13,341(40)	18,234(66)	19,833(45)	19,669(41)	1,229
PG	Procter & Gamble	NYSE	1,518(44)	1,881(54)	2,631(52)	3,314(52)	3,668(50)	2,584(27)	3,137(43)	4,555(42)	7,535(47)	8,397(45)	1,221
SBC	Sbc Communications	NYSE	1,603(37)	2,267(52)	3,038(52)	3,060(46)	3,364(43)	2,042(24)	2,791(48)	3,909(39)	6,478(31)	8,346(26)	1,220
T	AT&T Corp.	NYSE	2,233(29)	1,851(40)	2,224(43)	2,353(44)	2,412(39)	1,657(26)	2,238(40)	2,931(32)	5,307(30)	7,091(20)	1,221
UTX	United Technologies	NYSE	767(41)	1,354(51)	1,986(53)	2,758(55)	3,107(55)	1,111(46)	2,183(46)	3,075(45)	6,657(49)	7,996(53)	1,220
WMT	Wal-Mart Stores	NYSE	1,946(43)	2,368(54)	2,847(58)	2,860(51)	4,394(50)	2,609(27)	2,675(47)	3,971(44)	5,610(39)	8,744(40)	1,221
XOM	Exxon Mobil	NYSE	1,720(41)	2,227(53)	3,467(54)	4,198(48)	4,489(47)	1,979(33)	2,712(43)	4,677(37)	8,340(32)	9,950(29)	1,219

NOTE: Symbols and names of the equities used in our empirical analysis. The third column is the exchange that we extracted data from and the following columns give the average number of transactions/quotations per day for each of the five years in our sample. The percentage of observations for which the transaction price or one of the quoted prices was different from the previous one is given in parentheses. The last column is the total number of data in our sample.

results for two equities, Alcoa (AA) and Microsoft (MSFT), which represent DJIA equities with low and high trading activities. The corresponding figures for the other 28 DJIA equities are available on request.

5.1 Empirical Implementation of Estimators

In practice, we do not rely on the intraday returns outside the $[a, b]$ interval. Thus in our empirical analysis we substitute (for $h > 0$)

$$\tilde{\gamma}_h \equiv \frac{m}{m-h} \sum_{i=1}^{m-h} y_{i,m} y_{i+h,m}$$

for the theoretical quantity

$$\hat{\gamma}_h \equiv \sum_{i=1}^{m} y_{i,m} y_{i+h,m},$$

because the latter relies on $y_{m+1,m}, \ldots, y_{m+h,m}$. (For $h < 0$, we define $\tilde{\gamma}_h \equiv \tilde{\gamma}_{|h|}$.) In the expression for $\tilde{\gamma}_h$ we use an upward scaling, $m/(m-h)$, of the "autocovariances" to compensate for the "missing" autocovariance terms. Thus our empirical implementation of the simplest kernel estimator is given by $RV^{(m)}_{\mathrm{AC}_1} = \sum_{i=1}^{m} y_{i,m}^2 + 2\frac{m}{m-1} \sum_{i=1}^{m-1} y_{i,m} y_{i+1,m}$.

5.2 Estimation of Market Microstructure Noise Parameters

Under the independent noise assumption used in Section 3, we have, from Lemma 2, that

$$\tilde{\omega}^2 \equiv \frac{RV^{(m)}}{2m} \xrightarrow{p} \omega^2 \quad \text{as } m \to \infty.$$

This estimator was proposed by Bandi and Russell (2005) and Zhang et al. (2005) for different purposes. Bandi and Russell (2005) used $\tilde{\omega}_t^2$ to determine the optimal sampling frequency m_0^*, whereas Zhang et al. (2005) used it to select the number of subsamples and to serve as a bias-correction device of their "second-best" one-scale subsample estimator.

Under the independent noise assumption, we have, from Lemma 2, that $E(RV^{(m)}) = IV + 2m\omega^2$. Whereas $\tilde{\omega}^2$ is asymptotically justified, our empirical results reveal that ω^2 is very small in practice, so small that $2m\omega^2$ is small relative to the IV, with the exception of the most liquid assets, such as INTC and MSFT. Whenever the $IV/2m$ is nonnegligible, $\tilde{\omega}^2$ will overestimate ω^2. Thus, better estimators are given by

$$\check{\omega}^2 \equiv \frac{RV^{(m)} - RV^{(13)}}{2(m-13)}$$

and

$$\hat{\omega}^2 \equiv \frac{RV^{(m)} - \widehat{IV}}{2m},$$

where $RV^{(13)}$ is based on intraday returns that span about 30 minutes each and $\widehat{IV}$ is some unbiased estimator of IV. From Lemmas 2 and 3, it follows that $\tilde{\omega}^2$, $\check{\omega}^2$, and $\hat{\omega}^2$ are asymptotically equivalent in the sense that they have the same probability limit as $m \to \infty$. But $\check{\omega}^2$ and $\hat{\omega}^2$ are unbiased for ω^2 for any finite m, and we show that $\tilde{\omega}^2$ is quite biased in many cases. Another problem is that the independent noise assumption need not hold at ultra-high frequencies, in which case the asymptotic bias is not given by $2m\omega^2$. Clearly, this is problematic for all three estimators.

Table 2 presents annual sample averages of $\tilde{\omega}^2$, $\check{\omega}^2$, and $\hat{\omega}^2$ for the 5 years of our sample. Here we use $RV^{(1\text{ tick})}_{\mathrm{AC}_1}$ as our choice for $\widehat{IV}$, which is unbiased under the independent noise assumption. The first estimator, $\tilde{\omega}^2$, assumes that $IV/2m$ is negligible, in which case all three estimators should be similar. Because $\check{\omega}^2$ and $\hat{\omega}^2$ generally agree, whereas $\tilde{\omega}^2$ is typically much larger, it is evident that $\tilde{\omega}^2$ overestimates ω^2. A related observation was made by Engle and Sun (2005).

Fact III. The noise is smaller than one might think.

Here, by small we mean that the bias of the RV due to noise is small. This is particularly so in the more recent years (see, e.g., the 2004 volatility signature plot for AA in Fig. 1). By small, we do not mean that the noise is unimportant, but rather that it has a less dramatic impact than suggested by the independent noise assumption.

The fact that the noise in each of the intraday returns is relatively small (even when sampling occurs at the highest possible frequency) will likely affect the properties of the suggested implementation of the two-scale estimator by Zhang et al. (2005). In fact, the one-scale estimator proposed by Zhang et al. (2005) may be more accurate when the noise is small, as Table 2 suggests. Similarly, when determining the optimal sampling frequency for RV (as in Bandi and Russell 2005), one should be careful not to overestimate ω^2, which would lead to a lower-than-optimal sampling frequency. The important message is that one should incorporate RV_{AC_1}, or some other unbiased estimator of IV, when estimating ω^2 from RV.

Another observation from Table 2 is that the noise has changed. For example, our 2001 estimates of ω^2 (using $\hat{\omega}^2$) are on average <20% of those of 2000, and are even smaller in subsequent years. A large portion of this reduction is due to decimalization. We discuss the changed empirical properties of the noise in more detail in our discussion of the results in Figures 6 and 7.

Now, if we were to believe the independent noise assumption (which is less of a stretch in 2000 than in subsequent years), then we could construct the following estimate of λ:

$$\hat{\lambda} \equiv \bar{\omega}^2 / \overline{IV},$$

where $\bar{\omega}^2 \equiv n^{-1} \sum_{t=1}^{n} \hat{\omega}_t^2$ and $\overline{IV} \equiv n^{-1} \sum_{t=1}^{n} RV^{(1\text{ tick})}_{\mathrm{AC}_1,t}$. The latter is an estimate of the average daily IV over the sample $t = 1, \ldots, n$, because RV_{AC_1} is unbiased for IV under the independent noise assumption. Similarly, $\bar{\omega}^2$ is an estimate of the average daily noise. If both ω^2 and IV are constant across days, such that λ is the same for all days, then $\hat{\lambda}$ is consistent for λ whenever a law of large numbers applies to $\hat{\omega}_t^2$ and $RV^{(1\text{ tick})}_{\mathrm{AC}_1,t}$, $t = 1, \ldots, n$. In practice, both ω^2 and IV are likely to vary across days, so $\hat{\lambda}$ should be viewed only as a proxy for the noise-to-signal ratio.

Table 3 contains empirical results for all 30 equities using both transaction and quotation data from 2000. Several interesting observations can be made based on these results. For the transaction data, we note that $\hat{\lambda}$ is typically found to be <.1%, and the theoretical reduction of the RMSE, $100[r_0(\hat{\lambda}, m_0^*) -$

Table 2. Estimates of the Noise Variance for Transaction Data (annual averages)

	2000			2001			2002			2003			2004		
Asset	$\tilde{\omega}^2$	$\check{\omega}^2$	$\hat{\omega}^2$	$\tilde{\omega}^2$	$\check{\omega}^2$	$\hat{\omega}^2$	$\tilde{\omega}^2$	$\check{\omega}^2$	$\hat{\omega}^2$	$\tilde{\omega}^2$	$\check{\omega}^2$	$\hat{\omega}^2$	$\tilde{\omega}^2$	$\check{\omega}^2$	$\hat{\omega}^2$
AA	1.786	.995	1.040	.447	.098	.117	.409	.150	.100	.201	.040	.055	.090	.011	.024
AXP	.618	.189	.261	.271	.054	.056	.257	.075	.060	.084	.023	.023	.033	.006	.010
BA	1.174	.610	.681	.292	.026	.045	.324	.118	.069	.162	.070	.044	.060	.010	.014
C	.633	.407	.438	.222	.085	.028	.256	.035	.030	.062	.016	.014	.034	.016	.013
CAT	1.672	.724	.834	.263	*−.032*	.028	.207	*−.025*	.029	.080	*−.004*	.008	.041	.001	.003
DD	1.130	.662	.676	.231	.050	.058	.228	.068	.048	.087	.035	.024	.053	.020	.016
DIS	1.638	1.179	1.205	.563	.273	.194	.539	.344	.258	.231	.151	.113	.119	.080	.062
EK	.912	.265	.413	.382	*−.084*	.020	.361	*−.031*	.037	.164	.010	.038	.124	*−.003*	.028
GE	.541	.390	.361	.196	.062	.031	.186	.084	.050	.089	.052	.049	.051	.031	.033
GM	.639	.018	.225	.222	*−.014*	.036	.178	*−.001*	.043	.092	.013	.025	.052	.001	.014
HD	.971	.592	.613	.256	.058	.054	.245	.091	.083	.114	.050	.053	.058	.017	.024
HON	1.374	.458	.599	.537	.089	.090	.451	.079	.080	.217	.088	.058	.098	.030	.024
HPQ	.821	.232	.246	.716	.320	.193	.711	.350	.254	.248	.108	.101	.142	.089	.078
IBM	.342	.134	.149	.112	.031	.021	.118	.029	.020	.040	.013	.009	.024	.011	.006
INTC	.232	.186	.208	.209	.171	.186	.077	.042	.054	.055	.034	.039	.046	.030	.032
IP	2.015	1.049	1.156	.431	.167	.092	.234	.068	.051	.117	.040	.026	.067	.007	.016
JNJ	.373	.196	.212	.132	.048	.049	.161	.066	.065	.067	.018	.021	.029	.008	.011
JPM	.426	*−.004*	.021	.368	.187	.097	.523	.194	.128	.125	.056	.052	.044	.016	.019
KO	.776	.435	.498	.188	.035	.035	.146	.046	.033	.073	.026	.019	.044	.020	.016
MCD	1.966	1.517	1.466	.336	.172	.117	.351	.174	.126	.246	.120	.115	.099	.044	.046
MMM	.492	*−.033*	.071	.190	*−.007*	*−.002*	.119	.014	.010	.032	.004	.003	.030	.001	.002
MO	2.891	2.376	2.487	.167	.028	.044	.130	.060	.038	.097	.028	.025	.043	.002	.011
MRK	.499	.242	.288	.159	.019	.015	.157	.029	.023	.056	.009	.011	.050	.013	.019
MSFT	.225	.194	.206	.073	.052	.060	.025	.007	.013	.036	.025	.027	.037	.029	.029
PG	.630	.328	.315	.185	.072	.045	.085	.015	.012	.031	.009	.004	.027	.007	.006
SBC	.992	.596	.662	.199	.031	.049	.361	.131	.087	.176	.064	.061	.094	.052	.052
T	2.135	1.704	1.756	.467	.157	.138	.544	.198	.222	.227	.058	.086	.203	.103	.111
UTX	.977	*−.049*	.120	.246	*−.044*	*−.006*	.189	*−.003*	.017	.064	.005	.006	.038	.008	.005
WMT	.892	.462	.545	.184	.029	.030	.131	.025	.025	.054	.002	.009	.032	.008	.012
XOM	.348	.148	.205	.143	.046	.031	.152	.084	.037	.063	.037	.025	.031	.012	.015

NOTE: Annual averages of estimates of ω^2 for transaction data using the three different estimators: $\tilde{\omega}^2 \equiv RV^{(m)}/(2m)$, $\check{\omega}^2 \equiv (RV^{(m)} - RV^{(13)})/(2(m-13))$, and $\hat{\omega}^2 \equiv (RV^{(m)} - \widehat{IV})/(2m)$. All numbers have been multiplied by 100.

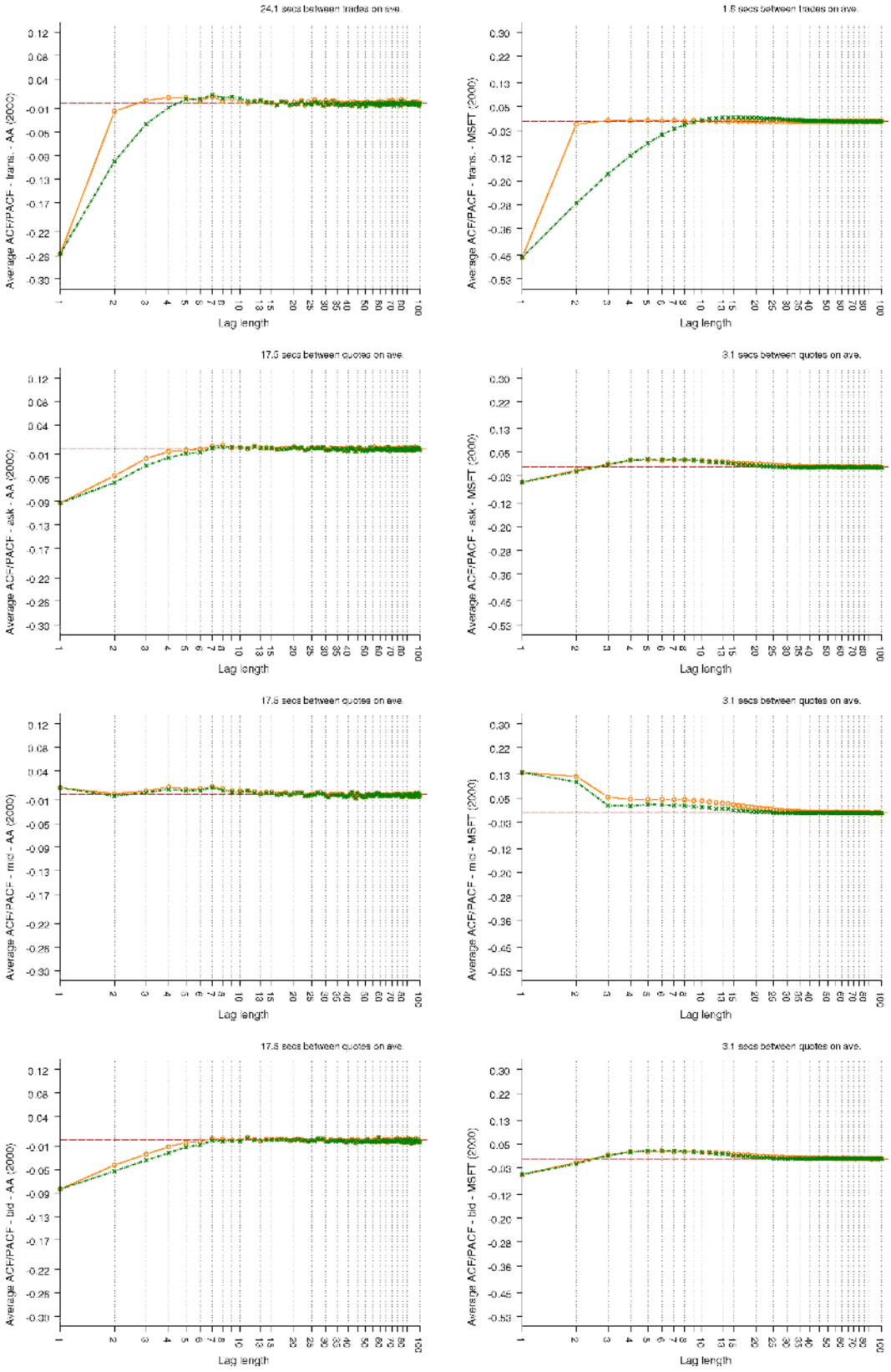

Figure 6. The ACF (▫—▫—▫) and PACF (×--×--×) for Four Different Tick-by-Tick Price Series: Transaction Prices, Bid Quotes, Ask Quotes, and Mid-Quotes. The left column is for AA, and the right column is for MSFT. The plotted series are the annual averages over the days in the year 2000.

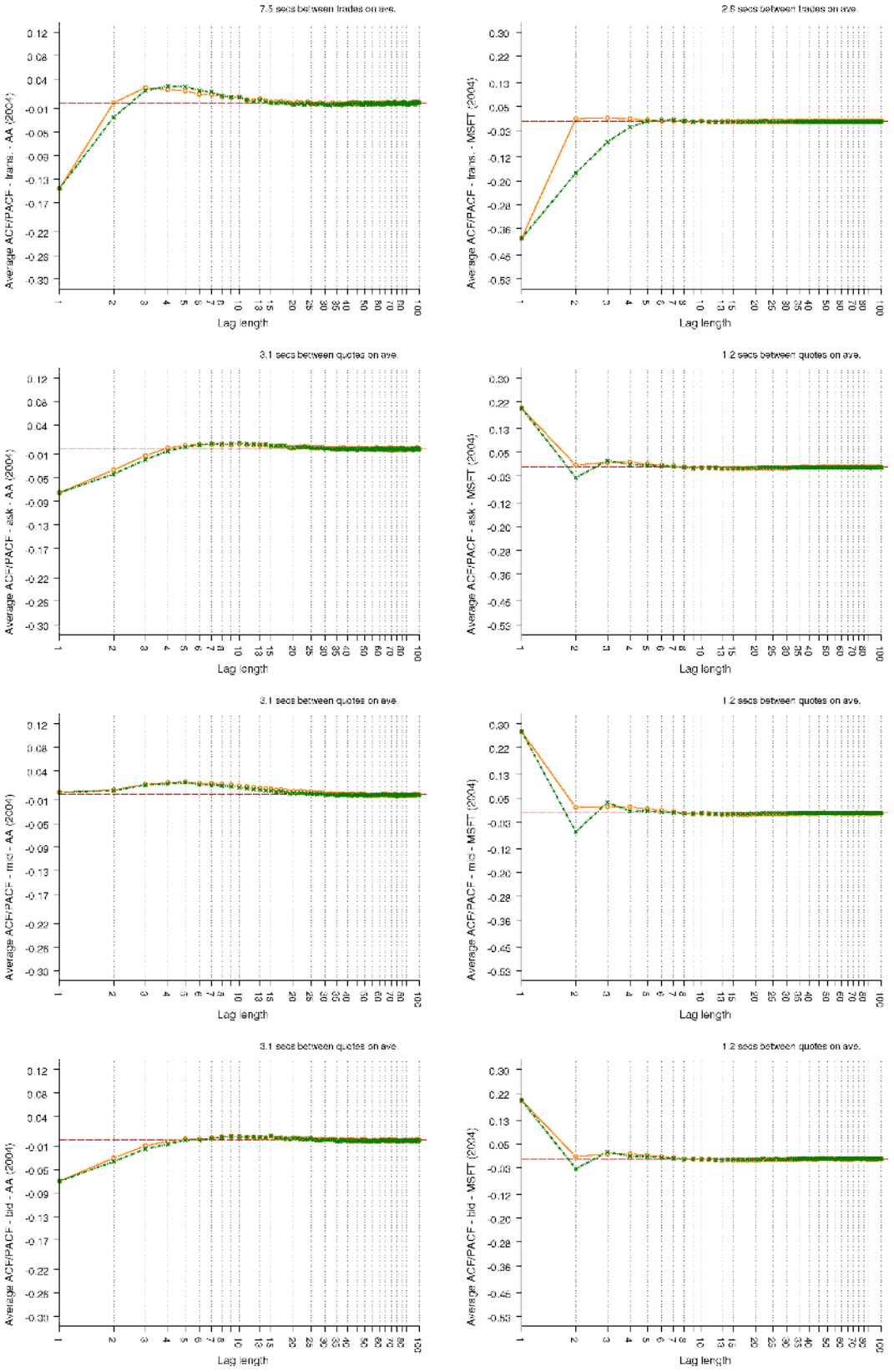

Figure 7. The ACF (□—□—□) and PACF (×- -×- -×) for Four Different Tick-by-Tick Price Series: Transaction Prices, Bid Quotes, Ask Quotes, and Mid-Quotes. The left column is for AA, and the right column is for MSFT. The plotted series are the annual averages over the days in the year 2004.

Table 3. Estimated Noise-to-Signal Ratios and "Optimal" Sampling Frequencies for the Year 2000

	Trades					Quotes	
Asset	$\hat{\omega}^2 \cdot 100$	$\overline{IV}$	$\hat{\lambda} \cdot 100$	m_0^*	m_1^*	$\Delta RMSE$	$\hat{\omega}^2 \cdot 100$
AA	1.0400	6.141	.1693	44	511	33.1%	.0026
AXP	.2605	5.244	.0497	100	1,743	43.6%	−.0351
BA	.6807	4.181	.1628	45	531	33.5%	−.0207
C	.4383	4.610	.0951	65	910	38.2%	−.0367
CAT	.8344	5.239	.1593	46	543	33.7%	−.0192
DD	.6756	5.769	.1171	56	739	36.4%	−.0274
DIS	1.2050	4.322	.2789	31	310	28.7%	−.0292
EK	.4133	3.495	.1183	56	732	36.3%	−.0153
GE	.3609	4.740	.0762	75	1,137	40.1%	−.0221
GM	.2249	3.242	.0694	80	1,248	40.9%	−.0338
HD	.6125	5.883	.1041	61	831	37.4%	−.0513
HON	.5991	6.669	.0898	67	964	38.7%	−.0671
HPQ	.2457	10.31	.0238	163	3,632	49.4%	−.0643
IBM	.1493	5.120	.0292	143	2,969	47.8%	−.0042
INTC	.2077	5.884	.0353	126	2,453	46.3%	−.0446
IP	1.1560	7.520	.1538	47	563	34.0%	−.0286
JNJ	.2116	2.445	.0866	69	1,000	39.0%	−.0254
JPM	.0212	5.726	.0037	566	23,350	62.0%	−.0387
KO	.4978	3.656	.1361	51	636	35.1%	−.0346
MCD	1.4660	4.557	.3218	28	269	27.4%	.0098
MMM	.0707	3.377	.0209	178	4,134	50.3%	−.0549
MO	2.4870	4.092	.6078	18	142	21.6%	.0276
MRK	.2883	3.287	.0877	68	987	38.9%	−.0143
MSFT	.2063	3.557	.0580	90	1,493	42.3%	−.0256
PG	.3145	4.719	.0667	82	1,299	41.2%	−.0104
SBC	.6617	3.912	.1691	44	512	33.1%	−.0415
T	1.7560	4.748	.3698	26	234	26.1%	−.0504
UTX	.1204	5.691	.0212	177	4,094	50.3%	−.0743
WMT	.5454	5.857	.0931	66	929	38.4%	−.0535
XOM	.2046	2.160	.0947	65	914	38.2%	−.0259

NOTE: Estimates of ω^2, IV, and the noise-to-signal ratio, λ (annual averages for 2000). Columns five and six are the optimal sampling frequencies for $RV^{(m)}$ and $RV^{(m)}_{AC_1}$ based on the listed value of $\hat{\lambda}$. The corresponding reduction in the RMSE, $100[r_0(\hat{\lambda}, m_0^*) - r_1(\hat{\lambda}, m_1^*)]/r_0(\hat{\lambda}, m_0^*)$ is listed in column seven. The last column contains estimates of $\hat{\omega}^2$ (annual average) using mid-quotes, where negative values are indicative of negative correlation between noise and efficient returns.

$r_1(\hat{\lambda}, m_1^*)]/r_0(\hat{\lambda}, m_0^*)$, is typically in the 25–50% range. For example, for Alcoa, that $\hat{\omega}^2_{AA} = 1.04\%$ and $\hat{\lambda}_{AA} = 1.04\%/6.14 = .1693\%$, which leads to the optimal sampling frequencies $m_0^* = 44$ and $m_1^* = 511$. For a typical trading day spanning 6.5 hours, this corresponds to intraday returns that (on average) span 9 minutes and 46 seconds. Bandi and Russell (2005) and Oomen (2005, 2006) reported "optimal" sampling frequencies for $RV^{(m)}$ that are similar to our estimates of m_0^*. By plugging these numbers into the formulas of Corollary 2, we find the reduction of the RMSE (from using $RV^{(m_1^*)}_{AC_1}$ rather than $RV^{(m_0^*)}$) to be 33%.

The noise-to-signal ratio, λ, is likely to differ across days, in which case the optimal sampling frequencies, m_0^* and m_1^*, will also differ across days. Thus $\hat{\lambda}$ should be viewed as a proxy for λ on a typical trading day, and our estimates should be viewed as approximations for "daily average values," in the sense that $m_0 = 90$ and $m_1 = 1{,}493$ appear to be sensible sampling frequencies to use with the MSFT transaction data. Fortunately, Figure 1 shows that $RV^{(m)}_{AC_1}$ is relatively insensitive to small deviations from m_1^*, such that a reasonable estimate for λ does produce a more accurate estimator based on RV_{AC_1} than one based on RV.

For the quotation data, almost all of our estimates of ω^2 are negative, which occurs whenever the sample average of $RV^{(1\text{ tick})}$ is smaller than that of $RV^{(1\text{ tick})}_{AC_1}$. This is obviously in conflict with the results of Lemmas 2 and 3, which dictate that the population difference, $E[RV^{(1\text{ tick})} - RV^{(1\text{ tick})}_{AC_1}]$, be positive. The expected difference is $2\omega^2$ times a number that is proportional to the average number of transactions/quotations per day. One explanation for the observation that $\hat{\omega}^2 < 0$ is that $\omega^2 \simeq 0$, such that the "wrong" sign occurs simply by chance. But this is highly improbable, because all but two of the estimates of ω^2 (not just about half of them) are found to be negative for the quotation data. Thus these negative estimates provide additional evidence against the independent noise assumption.

The autocorrelation function (ACF) and the partial autocorrelation function (PACF) for intraday returns provide a simple eyeball test of the independent noise assumption. Figures 6 and 7 present annual averages of the ACF and PACF estimated for each day using one-tick sampling. The results for 2000 are given in Figure 6; those for 2004, in Figure 7. The upper panels are those for transaction prices, and the subsequent panels correspond to ask, mid, and bid quotes.

The results for AA in 2000 suggest that time dependence in the noise process may be specific to tick time and that the memory in transaction prices lasts only a few ticks, slightly longer for quoted prices. In 2004 the time dependence in transaction prices appears to be longer—at least 10 transactions—and because we are looking at an annual average, it may be even longer for some days. The duration between transactions

in 2004 was about a third of what it was in 2000. Thus when the time dependence in tick time is converted into calendar time, the time dependence in 2000 and 2004 is about the same. The results for MSFT are in some respects very different from those for AA. The average ACF and PACF for the transaction data may suggest that the independent noise assumption is appropriate for this price series; however, many of the higher-order autocovariances are nontrivial, which explains why $RV^{(1\text{ tick})}_{\text{AC}_1}$ is often very different from $RV^{(1\text{ tick})}_{\text{AC}_{30}}$. For the quoted price series, the time dependence is slightly more involved, particularly for mid-quotes.

Comparing the results in Figures 6 and 7 shows that the noise properties have changed after decimalization of the tick size. For transaction data, the change in the noise properties is most evident for AA, whereas in quoted prices the change is most pronounced for MSFT. For example, the first-order autocovariances for bid and ask quotes have opposite signs in 2000 and 2004. Thus, based on the results in Table 2 and Figures 6 and 7, we are led to the following fact.

Fact IV. The properties of the noise have changed over time.

Because the ACFs and PACFs plotted in Figures 6 and 7 are averages over daily estimates, there may be a great deal of variation across days, although we did not find this to be the case in the daily ACFs and PACFs. (For formal hypothesis tests that address properties of market microstructure noise [day-by-day], see Awartani, Corradi, and Distaso 2004.)

6. PRICE DECOMPOSITION BY COINTEGRATION METHODS

Empirical studies on volatility estimation from contaminated high-frequency returns have typically used univariate price series, such as transaction prices, ask quotes, bid quotes, or mid-quotes. All of these series are proxies for the same efficient price, and thus it is natural to incorporate the information from all series when estimating the volatility of the latent efficient price.

One way to combine the information from multiple series is to compute an RV-type measure for each series and take the average of these, as was done by Hansen and Lunde (2005b), who took the average of estimators based on bid and ask quotes. Here we take a different approach and use cointegration methods to extract the efficient price (and noise) from a vector consisting of bid and ask quotes and transaction prices. This approach can be extended to include additional price series, for example, limit order books can be used to define additional bid and ask price series that depend on the volume offered at various prices, and prices from multiple exchanges could also be included in the analysis.

The purpose of this analysis is threefold. First, the method makes it possible to decompose the different prices into a common stochastic trend (efficient price) and transitory components (one noise process for each price series). This makes it possible to compare the properties of these series with those that we observe in our volatility signature plots. Second, the decomposition reveals how the efficient price is tied to innovations in the different price series. This shows which price series are most informative about the efficient price. Third, we can study the dynamic impact on quotes and transaction prices as a response to a change in the efficient price. This can be done with standard impulse response analysis. (For alternative ways of decomposing the observed price series, see, e.g., Engle and Sun 2005, who used a ACD–GARCH specification where the noise process has a two-component ARMA structure; also see Frijns and Lehnert 2004.)

We use the vector autoregressive framework that was used to analyze price discovery (from multiple markets) by Harris, McInish, Shoesmith, and Wood (1995), Hasbrouck (1995), and Harris, McInish, and Wood (2002), among others. Our analysis differs from this literature because we apply cointegration techniques to quotes and transactions in conjunction, not to transaction prices from different exchanges. Because it is possible to obtain the prevailing bid and ask prices at the time at which a transaction occurs, we avoid issues related to nonsynchronous trading. Our impulse response analysis, which we believe to be novel, shows how bid, ask, and transaction prices dynamically respond to a change in the efficient price.

We let t_i, $i = 0, 1, \ldots, m$, denote the times when transactions occur during some trading day and drop the subscript m to simplify the notation. We define the vector of "prices" by

$$\mathbf{p}_{t_i} = \begin{pmatrix} \text{transaction price at time } t_i \\ \text{prevailing ask price at time } t_i \\ \text{prevailing bid price at time } t_i \end{pmatrix},$$

where we use log-prices as in the previous sections.

Suppose that the dynamics of the vector of prices, $\mathbf{p}_{t_i}$, can be approximated by the vector autoregressive error correction model,

$$\Delta\mathbf{p}_{t_i} = \boldsymbol{\alpha}\boldsymbol{\beta}'\mathbf{p}_{t_{i-1}} + \sum_{j=1}^{k-1} \boldsymbol{\Gamma}_j \Delta\mathbf{p}_{t_{i-j}} + \boldsymbol{\mu} + \boldsymbol{\varepsilon}_{t_i}, \tag{7}$$

where $\boldsymbol{\varepsilon}_{t_i}$, $i = 1, \ldots, m$, is a sequence of uncorrelated error terms. It is reasonable to assume that each of the three observed price series shares the same stochastic trend such that any pair of prices cointegrate, because the spread is presumed to be stationary. This implies that $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are 3×2 matrices with full column rank and that we may impose the two natural cointegration vectors,

$$\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = \begin{pmatrix} 1 & 0 \\ -\frac{1}{2} & 1 \\ -\frac{1}{2} & -1 \end{pmatrix}. \tag{8}$$

The space spanned by the two columns of $\boldsymbol{\beta}$ defines the set of cointegration relations. Thus any two linearly independent vectors in this subspace can be used to define $\boldsymbol{\beta}$. We choose these two vectors because they are simple to interpret; $\boldsymbol{\beta}_1'\mathbf{p}_{t_i}$ is the difference between the transaction price and the mid-quote, and $\boldsymbol{\beta}_2'\mathbf{p}_{t_i}$ is the bid–ask spread.

Key to this analysis are the 3×1 vectors, $\boldsymbol{\alpha}_\perp$ and $\boldsymbol{\beta}_\perp$, which are orthogonal to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. [Thus $\boldsymbol{\alpha}_\perp'\boldsymbol{\alpha} = \boldsymbol{\beta}_\perp'\boldsymbol{\beta} = (0, 0)$.] As is the case for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, these vectors are not fully identified, so we impose the normalizations

$$\boldsymbol{\beta}_\perp = \boldsymbol{\iota} \qquad \text{and} \qquad \boldsymbol{\alpha}_\perp'\boldsymbol{\iota} = 1,$$

where $\boldsymbol{\iota} \equiv (1, 1, 1)'$.

It follows from the Granger representation theorem that

$$\mathbf{p}_t = \boldsymbol{\beta}_\perp(\boldsymbol{\alpha}_\perp'\boldsymbol{\Gamma}\boldsymbol{\beta}_\perp)^{-1}\boldsymbol{\alpha}_\perp'\sum_{s=1}^{t}\boldsymbol{\varepsilon}_s + \mathbf{C}(L)(\boldsymbol{\varepsilon}_t + \boldsymbol{\mu}) + \mathbf{A}_0,$$

where $\boldsymbol{\Gamma} \equiv \mathbf{I} - \boldsymbol{\Gamma}_1 - \cdots - \boldsymbol{\Gamma}_{k-1}$ and $\mathbf{A}_0$ is a that depending on initial values of $\mathbf{p}_t$. The Granger representation theorem is due to Johansen (1988) and Hansen (2005), who obtained a recursive formula for the coefficients of $\mathbf{C}(L)$ (and details about $\mathbf{A}_0$), which we use in our analysis.

6.1 Common Stochastic Trend

There are a number of ways to define the *common stochastic trend* from the Granger representation (see Johansen 1996). The most natural definition in this framework is

$$p_{t_i}^* \equiv (\boldsymbol{\alpha}_\perp'\boldsymbol{\Gamma}\boldsymbol{\beta}_\perp)^{-1}\sum_{j=1}^{i}\boldsymbol{\alpha}_\perp'\boldsymbol{\varepsilon}_{t_j} + \text{initial value}.$$

This definition has the desired martingale property, because the corresponding intraday returns,

$$y_{t_i}^* \equiv \frac{\boldsymbol{\alpha}_\perp'\boldsymbol{\varepsilon}_{t_i}}{\boldsymbol{\alpha}_\perp'\boldsymbol{\Gamma}\boldsymbol{\beta}_\perp},$$

are uncorrelated. Thus the Granger representation can be expressed as

$$\mathbf{p}_{t_i} = \begin{pmatrix} p_{t_i}^* \\ p_{t_i}^* \\ p_{t_i}^* \end{pmatrix} + \mathbf{C}(L)\boldsymbol{\varepsilon}_{t_i} + \mathbf{A}_0.$$

This definition of the common stochastic trend, $p_{t_i}^*$, corresponds to that of Hasbrouck (1995, 2002). Johansen (1996) also discussed alternative definitions of the common stochastic trend, such as $\boldsymbol{\beta}_\perp'\mathbf{p}_{t_i}$ and $\boldsymbol{\alpha}_\perp'\boldsymbol{\Gamma}\mathbf{p}_{t_i}$, which are linear combinations of the observed price vector; these are known as Granger–Gonzalo stochastic trends in the literature (see Gonzalo and Granger 1995). The connection between $p_{t_i}^*$ and a linear combination of $\mathbf{p}_{t_i}$ follows directly from the Granger representation, because the latter involves a component that is proportional to $p_{t_i}^*$ and a second component that relates to the stationary part of the Granger representation. This relation was discussed by Johansen (1996) (see also Hansen and Johansen 1998, pp. 27–28), and similar arguments were given by de Jong (2002) and Baillie, Booth, Tse, and Zabotina (2002) (see also Harris et al. 2002; Lehmann 2002).

It is the martingale property of $p_{t_i}^*$ that makes $p_{t_i}^*$ the most natural definition of the efficient price, and the RV of $p_{t_i}^*$ is given by

$$RV_{p^*} \equiv \sum_{i=1}^{m}(y_{t_i}^*)^2 = \frac{\sum_{i=1}^{m}(\boldsymbol{\alpha}_\perp'\boldsymbol{\varepsilon}_{t_i})^2}{(\boldsymbol{\alpha}_\perp'\boldsymbol{\Gamma}\boldsymbol{\beta}_\perp)^2}.$$

Naturally, the parameters need to be estimated in practice, so we rely on

$$\hat{p}_{t_i}^* = (\hat{\boldsymbol{\alpha}}_\perp'\hat{\boldsymbol{\Gamma}}\boldsymbol{\beta}_\perp)^{-1}\sum_{j=1}^{i}\hat{\boldsymbol{\alpha}}_\perp'\hat{\boldsymbol{\varepsilon}}_{t_j},$$

where $\hat{\varepsilon}_{t_i}$ are the residuals from (7). The estimation is described in Appendix C. Given $\hat{p}_{t_i}^*$, we can now define

$$RV_{\hat{p}^*} \equiv \sum_{i=1}^{m}\big(\hat{y}_{t_j}^*\big)^2 = \frac{\sum_{i=1}^{m}(\hat{\boldsymbol{\alpha}}_\perp'\hat{\boldsymbol{\varepsilon}}_{t_i})^2}{(\hat{\boldsymbol{\alpha}}_\perp'\hat{\boldsymbol{\Gamma}}\boldsymbol{\beta}_\perp)^2}.$$

6.2 Empirical Results

We estimate the parameters in (7) for each of the days individually, with the number of lags, k, chosen to be the smallest number ($\leq$10) that made Ljung–Box tests at lags 5, 10, 15, and 30 insignificant at the 5% significance level. Some additional details about the estimation are given in Appendix C.

Assuming that the ultra-high–frequency price observations are well described by (7) is problematic for several reasons. The duration between observations is an important aspect of the data ignored by model (7), and the discreteness of the data has implications for the properties of the "errors" $\boldsymbol{\varepsilon}_{t_i}$, $i = 1, \ldots, m$, and the interdependence with the vector of prices. The reader should be aware that we have ignored all such issues, and thus the results that we draw from this analysis should be viewed as a rough approximation.

A key parameter in our analysis is $\boldsymbol{\alpha}_\perp$ (3×1 vector) that shows how the efficient price is related to "innovations" in the different price series. In our empirical analysis the elements of $\boldsymbol{\alpha}_\perp$ correspond to transaction prices, ask quotes, and bid quotes, and so we use the following notation:

$$\boldsymbol{\alpha}_\perp' = (\alpha_{\perp,\text{tr}}, \alpha_{\perp,\text{ask}}, \alpha_{\perp,\text{bid}}).$$

Table 4 reports a summary of our empirical estimates, which are averages of the daily estimates, $\hat{\boldsymbol{\alpha}}_\perp$, and $\overline{RV}_{\hat{p}^*}$. For comparison, we also report the average RVquantities $\overline{RV}^{(1\text{ tick})}_{\text{ACNW}_{15}}$, $\overline{RV}^{(1\text{ tick})}_{\text{ACNW}_{30}}$, and $\overline{RV}^{(13)}$. To get a sense of the dispersion of the daily estimates, we also report the 5% and 95% quantiles in brackets below each of the averages.

It is striking how similar the average $\hat{\boldsymbol{\alpha}}_\perp$ is across equities listed on the NYSE; the average point estimates are generally quite close to $\boldsymbol{\alpha}_\perp = (\frac{1}{2}, \frac{1}{4}, \frac{1}{4})'$. In contrast, the two NASDAQ stocks, INTC and MSFT, produce average estimates that stand out by being closer to $\boldsymbol{\alpha}_\perp = (\frac{1}{4}, \frac{3}{8}, \frac{3}{8})'$. Thus the instantaneous correlation between innovations in the transaction price and the efficient price is larger for NYSE stocks, and, consequently, the quoted prices are more closely related to the efficient price for NASDAQ stocks than to that for NYSE stocks. We find this to be a very interesting empirical observation that is likely tied to the differences by which the two exchanges operate. Specifically, quotes on the NASDAQ are competitive and binding, such that movements in these are more likely to be tied to movements in the efficient price than are movements in the specialist quotes on the NYSE.

The estimated model allows us to compute the daily estimates,

$$\sum_{i=1}^{m}\hat{e}_{t_i}^2 \qquad \text{and} \qquad \sum_{i=1}^{m}\hat{y}_{t_i}^*\hat{e}_{t_i},$$

where $\hat{e}_{t_i}$ is an element of $\hat{\mathbf{e}}_{t_i} \equiv \hat{\mathbf{u}}_{t_i} - \hat{\mathbf{u}}_{t_{i-1}}$ and $\hat{\mathbf{u}}_{t_i} = \mathbf{p}_{t_i} - p_{t_i}^*$. (De-meaning the elements of $\hat{\mathbf{u}}_{t_i}$ is not required because it does

Table 4. Cointegration Results for the Year 2004

	Lags	$\hat{\alpha}_{\perp,\,tr}$	$\hat{\alpha}_{\perp,\,ask}$	$\hat{\alpha}_{\perp,\,bid}$	$\overline{RV}_{\hat{p}^*}$	$\overline{RV}^{(1\ tick)}_{ACNW_{15}}$	$\overline{RV}^{(1\ tick)}_{ACNW_{30}}$	$\overline{RV}^{(13)}$
AA	5.59	.51	.26	.22	2.30	2.52	2.50	2.21
	[2.00; 1.0]	[.35; .68]	[.09; .41]	[.04; .37]	[1.00; 4.61]	[1.03; 4.70]	[.90; 4.89]	[.50; 5.30]
AXP	4.09	.46	.29	.26	.67	.71	.70	.69
	[1.00; 10.0]	[.33; .63]	[.13; .45]	[.08; .40]	[.29; 1.33]	[.28; 1.46]	[.24; 1.53]	[.13; 1.91]
BA	5.06	.46	.29	.24	1.46	1.52	1.53	1.49
	[2.00; 10.0]	[.33; .61]	[.17; .44]	[.07; .39]	[.63; 2.84]	[.66; 3.01]	[.58; 3.35]	[.34; 3.60]
C	4.18	.48	.27	.25	1.01	.99	.91	.83
	[1.00; 10.0]	[.33; .63]	[.16; .38]	[.14; .35]	[.43; 2.06]	[.40; 2.05]	[.35; 2.10]	[.19; 1.80]
CAT	5.00	.45	.29	.26	1.61	1.64	1.59	1.49
	[2.00; 10.0]	[.33; .57]	[.16; .42]	[.13; .42]	[.72; 3.08]	[.73; 3.29]	[.67; 3.27]	[.42; 3.51]
DD	4.28	.44	.29	.27	1.12	1.11	1.03	1.02
	[1.40; 10.0]	[.32; .56]	[.13; .43]	[.15; .39]	[.52; 2.06]	[.49; 2.02]	[.42; 2.00]	[.22; 2.50]
DIS	4.21	.41	.31	.29	1.68	1.64	1.51	1.36
	[1.00; 10.0]	[.30; .53]	[.18; .44]	[.12; .41]	[.70; 3.07]	[.68; 3.23]	[.55; 3.07]	[.31; 3.31]
EK	5.37	.47	.29	.24	2.30	2.41	2.44	2.46
	[2.00; 10.0]	[.28; .69]	[.07; .48]	[−.04; .43]	[.79; 4.72]	[.84; 4.76]	[.69; 4.73]	[.40; 6.27]
GE	3.90	.46	.28	.26	.87	.96	.97	.87
	[1.00; 8.00]	[.26; .62]	[.15; .43]	[.13; .40]	[.39; 1.80]	[.39; 1.94]	[.36; 2.01]	[.26; 1.93]
GM	5.33	.48	.26	.26	1.28	1.39	1.42	1.45
	[2.00; 10.0]	[.33; .67]	[.10; .41]	[.10; .42]	[.54; 2.36]	[.57; 2.83]	[.50; 3.27]	[.33; 3.53]
HD	4.93	.47	.27	.26	1.37	1.47	1.48	1.37
	[2.00; 10.0]	[.31; .63]	[.11; .40]	[.10; .40]	[.60; 2.67]	[.65; 2.80]	[.61; 2.94]	[.32; 3.06]
HON	5.15	.47	.28	.25	2.03	2.02	1.92	1.82
	[1.00; 10.0]	[.33; .64]	[.12; .44]	[.09; .40]	[.85; 3.94]	[.83; 4.06]	[.77; 4.19]	[.48; 3.55]
HPQ	4.36	.40	.31	.29	2.07	2.08	1.97	1.84
	[1.00; 10.0]	[.28; .54]	[.17; .42]	[.15; .42]	[.80; 4.07]	[.72; 4.60]	[.66; 4.47]	[.37; 4.35]
IBM	4.92	.43	.29	.27	.90	.88	.81	.71
	[2.00; 10.0]	[.33; .56]	[.18; .42]	[.16; .39]	[.36; 1.77]	[.32; 1.69]	[.30; 1.64]	[.18; 1.53]
INTC	4.60	.25	.37	.38	2.36	2.34	2.30	2.01
	[2.00; 9.00]	[.18; .34]	[.21; .50]	[.24; .52]	[1.07; 4.21]	[1.02; 4.03]	[.95; 3.94]	[.59; 4.17]
IP	4.69	.45	.28	.27	1.19	1.27	1.26	1.27
	[1.00; 10.0]	[.30; .62]	[.12; .42]	[.11; .44]	[.46; 2.50]	[.51; 2.58]	[.42; 2.44]	[.32; 3.11]
JNJ	4.45	.45	.29	.26	.81	.82	.80	.79
	[2.00; 10.0]	[.32; .58]	[.14; .43]	[.08; .39]	[.33; 1.66]	[.34; 1.62]	[.29; 1.71]	[.15; 1.96]
JPM	3.87	.46	.28	.26	1.08	1.13	1.11	1.08
	[1.00; 9.60]	[.30; .62]	[.12; .42]	[.13; .38]	[.46; 2.43]	[.47; 2.64]	[.44; 2.93]	[.28; 2.67]
KO	4.19	.41	.29	.30	.95	.97	.92	.81
	[1.00; 10.0]	[.28; .55]	[.16; .40]	[.15; .42]	[.42; 1.85]	[.43; 1.84]	[.36; 1.89]	[.22; 1.95]
MCD	4.55	.42	.31	.27	1.39	1.43	1.45	1.38
	[1.00; 10.0]	[.27; .60]	[.16; .46]	[.09; .41]	[.60; 3.14]	[.52; 3.19]	[.49; 3.26]	[.32; 3.45]
MMM	4.86	.45	.28	.26	1.09	1.11	1.07	.96
	[2.00; 10.0]	[.33; .57]	[.14; .44]	[.13; .40]	[.43; 2.11]	[.44; 2.17]	[.39; 2.37]	[.23; 2.10]
MO	5.04	.48	.28	.25	1.48	1.51	1.51	1.52
	[2.00; 10.0]	[.33; .67]	[.09; .42]	[.09; .39]	[.34; 3.17]	[.29; 3.35]	[.25; 3.31]	[.17; 3.55]
MRK	4.60	.48	.27	.25	1.30	1.38	1.36	1.30
	[2.00; 10.0]	[.33; .63]	[.12; .40]	[.08; .40]	[.42; 3.84]	[.45; 3.79]	[.40; 3.55]	[.28; 3.94]
MSFT	4.43	.24	.40	.36	1.16	1.16	1.12	.89
	[2.00; 9.00]	[.16; .32]	[.21; .59]	[.16; .54]	[.50; 2.19]	[.50; 2.22]	[.48; 2.18]	[.21; 2.18]
PG	5.06	.49	.28	.23	.87	.87	.83	.75
	[2.00; 10.0]	[.36; .64]	[.15; .45]	[.10; .34]	[.34; 1.64]	[.33; 1.67]	[.29; 1.67]	[.18; 1.65]
SBC	4.00	.38	.31	.30	1.36	1.44	1.38	1.28
	[1.00; 10.0]	[.21; .55]	[.15; .47]	[.14; .45]	[.56; 2.64]	[.52; 2.83]	[.44; 2.87]	[.26; 3.12]
T	4.12	.34	.34	.32	1.65	1.77	1.86	2.00
	[1.00; 10.0]	[.21; .49]	[.21; .46]	[.17; .47]	[.62; 3.32]	[.69; 3.87]	[.67; 4.43]	[.48; 4.81]
UTX	5.16	.46	.28	.26	1.19	1.17	1.11	1.06
	[2.00; 10.0]	[.33; .62]	[.15; .42]	[.12; .38]	[.45; 2.47]	[.45; 2.51]	[.38; 2.39]	[.23; 2.70]
WMT	4.98	.55	.23	.21	1.11	1.14	1.10	1.04
	[2.00; 10.0]	[.42; .72]	[.09; .36]	[.08; .34]	[.53; 2.22]	[.57; 2.21]	[.50; 2.05]	[.30; 2.30]
XOM	4.09	.47	.27	.26	.78	.85	.85	.84
	[1.00; 9.65]	[.29; .62]	[.16; .40]	[.13; .39]	[.34; 1.60]	[.34; 1.61]	[.30; 1.68]	[.23; 1.71]

NOTE: Annual averages of the selected lag length, $\hat{\alpha}_{\perp}$, realized variance of $\hat{p}^*_t$, two measures of $RV^{(1\ tick)}_{ACNW_q}$, and the standard RV based on 30-minute sampling, where $\overline{RV}^{(1\ tick)}_{ACNW_q} \equiv \frac{1}{n}\sum_{t=1}^{n}(RV^{(1\ tick),\,tr}_{ACNW_{q,t}} + RV^{(1\ tick),\,bid}_{ACNW_{q,t}} + RV^{(1\ tick),\,ask}_{ACNW_{q,t}})/3$, and similarly for $\overline{RV}^{(13)}$. The numbers in the squared bracket are the 5% and 95% quantiles for the different statistics.

not affect $\hat{\mathbf{e}}_{t_i}$.) Table 5 reports daily averages for the different price series.

Figure 8 plots our estimate of the efficient price, $\hat{p}^*_{t_i}$, for the same window of time plotted in Figure 2. It is comforting to see that our estimate of the efficient price is in line with the quoted bid and ask prices. Rarely is $\hat{p}^*_{t_i}$ outside the bid–ask bounds, so there is no immediate evidence that a profit can be made from the discrepancies between $\hat{p}^*_{t_i}$ and the observed prices.

6.3 Impulse Response Function

Define the two matrices,

$$\mathbf{\Pi} = \boldsymbol{\alpha}\boldsymbol{\beta}' \qquad \text{and} \qquad \mathbf{C} = \boldsymbol{\beta}_{\perp}(\boldsymbol{\alpha}'_{\perp}\mathbf{\Gamma}\boldsymbol{\beta}_{\perp})^{-1}\boldsymbol{\alpha}'_{\perp}.$$

Table 5. Variations and Covariation for Noise and Efficient Price for the Year 2004

	$\overline{RV}_{\hat{p}^*}$	$\overline{\sum_i e_i^2}^{tr}$	$\overline{\sum_i e_i^2}^{ask}$	$\overline{\sum_i e_i^2}^{mid}$	$\overline{\sum_i e_i^2}^{bid}$	$\overline{\sum_i e_i y_i^*}^{tr}$	$\overline{\sum_i e_i y_i^*}^{ask}$	$\overline{\sum_i e_i y_i^*}^{mid}$	$\overline{\sum_i e_i y_i^*}^{bid}$
AA	2.30	.79	1.47	.89	1.73	−.30	−.75	−.81	−.86
	[1.00; 4.61]	[.39; 1.58]	[.66; 2.92]	[.31; 2.10]	[.77; 3.41]	[−1.01; .13]	[−2.13; −.05]	[−2.05; −.09]	[−2.30; −.12]
AXP	.67	.31	.41	.20	.46	−.08	−.16	−.17	−.19
	[.29; 1.33]	[.17; .54]	[.19; .76]	[.07; .41]	[.21; .89]	[−.28; .04]	[−.43; .00]	[−.45; −.01]	[−.50; −.01]
BA	1.46	.63	.92	.46	1.10	−.17	−.35	−.40	−.45
	[.63; 2.84]	[.28; 1.19]	[.42; 1.80]	[.18; .96]	[.51; 2.26]	[−.70; .09]	[−.93; −.03]	[−1.00; −.08]	[−1.20; −.05]
C	1.01	.49	.73	.35	.78	.04	−.20	−.22	−.23
	[.43; 2.06]	[.26; .72]	[.33; 1.33]	[.12; .74]	[.37; 1.57]	[−.19; .17]	[−.57; −.01]	[−.62; −.03]	[−.66; −.02]
CAT	1.61	.63	.94	.45	1.12	−.36	−.37	−.42	−.46
	[.72; 3.08]	[.27; 1.16]	[.35; 1.92]	[.15; .84]	[.45; 2.18]	[−.88; −.07]	[−.86; −.03]	[−.93; −.08]	[−1.04; −.05]
DD	1.12	.57	.81	.34	.87	−.04	−.22	−.22	−.22
	[.52; 2.06]	[.34; .92]	[.37; 1.54]	[.14; .74]	[.42; 1.57]	[−.28; .11]	[−.64; .01]	[−.63; −.02]	[−.61; −.01]
DIS	1.68	1.65	1.57	.62	1.66	.35	−.19	−.23	−.26
	[.70; 3.07]	[.88; 2.70]	[.61; 3.11]	[.23; 1.21]	[.70; 3.13]	[.07; .74]	[−.66; .10]	[−.69; −.01]	[−.82; .04]
EK	2.30	.89	1.28	.74	1.52	−.50	−.67	−.73	−.79
	[.79; 4.72]	[.38; 1.72]	[.51; 2.77]	[.21; 1.80]	[.56; 3.42]	[−1.66; .03]	[−1.96; −.02]	[−2.05; −.04]	[−2.18; −.03]
GE	.87	.87	.80	.40	.83	.22	−.24	−.25	−.26
	[.39; 1.80]	[.52; 1.28]	[.33; 1.55]	[.11; .83]	[.38; 1.52]	[.10; .38]	[−.62; .00]	[−.66; −.03]	[−.68; −.02]
GM	1.28	.45	.75	.39	.81	−.15	−.35	−.37	−.39
	[.54; 2.36]	[.23; .80]	[.32; 1.52]	[.13; .91]	[.38; 1.52]	[−.53; .08]	[−.96; −.04]	[−.94; −.06]	[−1.03; −.07]
HD	1.37	.70	.93	.48	.98	−.04	−.38	−.39	−.40
	[.60; 2.67]	[.36; 1.28]	[.38; 1.78]	[.17; 1.01]	[.43; 2.03]	[−.33; .15]	[−.95; −.03]	[−.95; −.05]	[−1.00; −.02]
HON	2.03	.85	1.41	.67	1.59	−.17	−.45	−.49	−.53
	[.85; 3.94]	[.43; 1.43]	[.61; 2.86]	[.21; 1.47]	[.69; 3.07]	[−.69; .19]	[−1.45; .02]	[−1.42; −.04]	[−1.52; .00]
HPQ	2.07	2.02	1.69	.75	1.79	.27	−.33	−.36	−.40
	[.80; 4.07]	[1.28; 2.96]	[.79; 3.17]	[.27; 1.42]	[.81; 3.10]	[−.03; .59]	[−1.28; .08]	[−1.11; .04]	[−1.26; .07]
IBM	.90	.40	.63	.26	.69	−.03	−.19	−.22	−.25
	[.36; 1.77]	[.17; .67]	[.26; 1.23]	[.09; .54]	[.31; 1.29]	[−.23; .10]	[−.51; −.01]	[−.59; −.03]	[−.69; −.04]
INTC	2.36	3.55	.82	.66	.80	−.13	−.83	−.83	−.82
	[1.07; 4.21]	[2.08; 5.24]	[.34; 1.43]	[.23; 1.25]	[.34; 1.35]	[−.96; .45]	[−1.60; −.30]	[−1.59; −.30]	[−1.60; −.29]
IP	1.19	.51	.79	.35	.85	−.14	−.26	−.28	−.31
	[.46; 2.50]	[.21; 1.07]	[.30; 1.65]	[.11; .70]	[.34; 1.70]	[−.47; .09]	[−.76; .02]	[−.73; −.01]	[−.77; −.01]

Table 5 (continued).

	$\overline{RV}_{\hat{p}^*}$	$\overline{\sum_i e_i^{2\,tr}}$	$\overline{\sum_i e_i^{2\,ask}}$	$\overline{\sum_i e_i^{2\,mid}}$	$\overline{\sum_i e_i^{2\,bid}}$	$\overline{\sum_i e_i y_i^{*\,tr}}$	$\overline{\sum_i e_i y_i^{*\,ask}}$	$\overline{\sum_i e_i y_i^{*\,mid}}$	$\overline{\sum_i e_i y_i^{*\,bid}}$
JNJ	.81	.40	.50	.27	.57	−.06	−.21	−.23	−.24
	[.33; 1.66]	[.25; .63]	[.22; .98]	[.09; .53]	[.26; .99]	[−.24; .05]	[−.55; −.01]	[−.57; −.03]	[−.56; −.02]
JPM	1.08	.60	.74	.36	.82	0	−.23	−.23	−.24
	[.46; 2.43]	[.33; .98]	[.31; 1.64]	[.11; .92]	[.37; 1.71]	[−.24; .13]	[−.79; .01]	[−.77; −.01]	[−.83; .00]
KO	.95	.53	.63	.27	.63	−.02	−.20	−.20	−.20
	[.42; 1.85]	[.31; .87]	[.32; 1.13]	[.09; .53]	[.31; 1.11]	[−.26; .13]	[−.59; .00]	[−.58; −.01]	[−.56; .00]
MCD	1.39	.94	1.05	.46	1.13	.04	−.26	−.29	−.31
	[.60; 3.14]	[.54; 1.49]	[.46; 2.09]	[.15; 1.23]	[.52; 2.20]	[−.30; .26]	[−1.16; .05]	[−1.07; .00]	[−1.07; .06]
MMM	1.09	.32	.56	.27	.59	−.20	−.28	−.30	−.32
	[.43; 2.11]	[.14; .62]	[.23; 1.11]	[.10; .59]	[.26; 1.14]	[−.52; −.01]	[−.71; −.05]	[−.75; −.08]	[−.77; −.07]
MO	1.48	.49	.84	.48	.89	−.23	−.48	−.49	−.49
	[.34; 3.17]	[.20; .81]	[.27; 1.90]	[.10; 1.16]	[.28; 1.85]	[−.73; .07]	[−1.31; −.01]	[−1.24; −.02]	[−1.28; .00]
MRK	1.30	.61	.90	.47	.97	−.06	−.35	−.37	−.40
	[.42; 3.84]	[.21; 1.52]	[.30; 2.50]	[.12; 1.40]	[.31; 2.36]	[−.43; .18]	[−1.36; .00]	[−1.40; −.02]	[−1.42; −.01]
MSFT	1.16	2.79	.41	.33	.41	.08	−.37	−.37	−.37
	[.50; 2.19]	[1.97; 3.95]	[.17; .77]	[.13; .63]	[.17; .79]	[−.22; .26]	[−.81; −.11]	[−.82; −.12]	[−.84; −.13]
PG	.87	.32	.58	.29	.67	−.10	−.22	−.24	−.27
	[.34; 1.64]	[.13; .59]	[.23; 1.10]	[.10; .60]	[.31; 1.27]	[−.30; .04]	[−.52; −.02]	[−.52; −.04]	[−.64; −.04]
SBC	1.36	1.24	.96	.43	1.02	.09	−.26	−.27	−.27
	[.56; 2.64]	[.74; 1.74]	[.36; 1.77]	[.10; .95]	[.41; 1.99]	[−.27; .34]	[−.96; .05]	[−.91; .00]	[−.91; .03]
T	1.65	1.94	1.29	.50	1.42	.10	−.21	−.23	−.25
	[.62; 3.32]	[1.05; 3.14]	[.57; 2.39]	[.15; 1.09]	[.58; 2.74]	[−.26; .38]	[−.84; .15]	[−.84; .08]	[−1.01; .13]
UTX	1.19	.46	.76	.33	.84	−.16	−.24	−.26	−.29
	[.45; 2.47]	[.17; .90]	[.31; 1.56]	[.11; .66]	[.35; 1.58]	[−.52; .04]	[−.68; −.01]	[−.65; −.04]	[−.77; −.02]
WMT	1.11	.37	.73	.43	.81	−.04	−.33	−.36	−.38
	[.53; 2.22]	[.18; .67]	[.38; 1.32]	[.20; .83]	[.41; 1.43]	[−.33; .10]	[−.74; −.08]	[−.81; −.11]	[−.83; −.11]
XOM	.78	.45	.55	.28	.58	.05	−.20	−.20	−.21
	[.34; 1.60]	[.26; .69]	[.22; 1.11]	[.08; .63]	[.23; 1.20]	[−.09; .18]	[−.54; −.01]	[−.60; −.02]	[−.62; −.02]

NOTE: Annual averages of the realized variance of efficient price and noise processes corresponding to different price series. The efficient price and the noise processes were deduced from daily estimates of a cointegration vector autoregression model. The numbers in the squared brackets are the 5% and 95% quantiles for the different statistics.

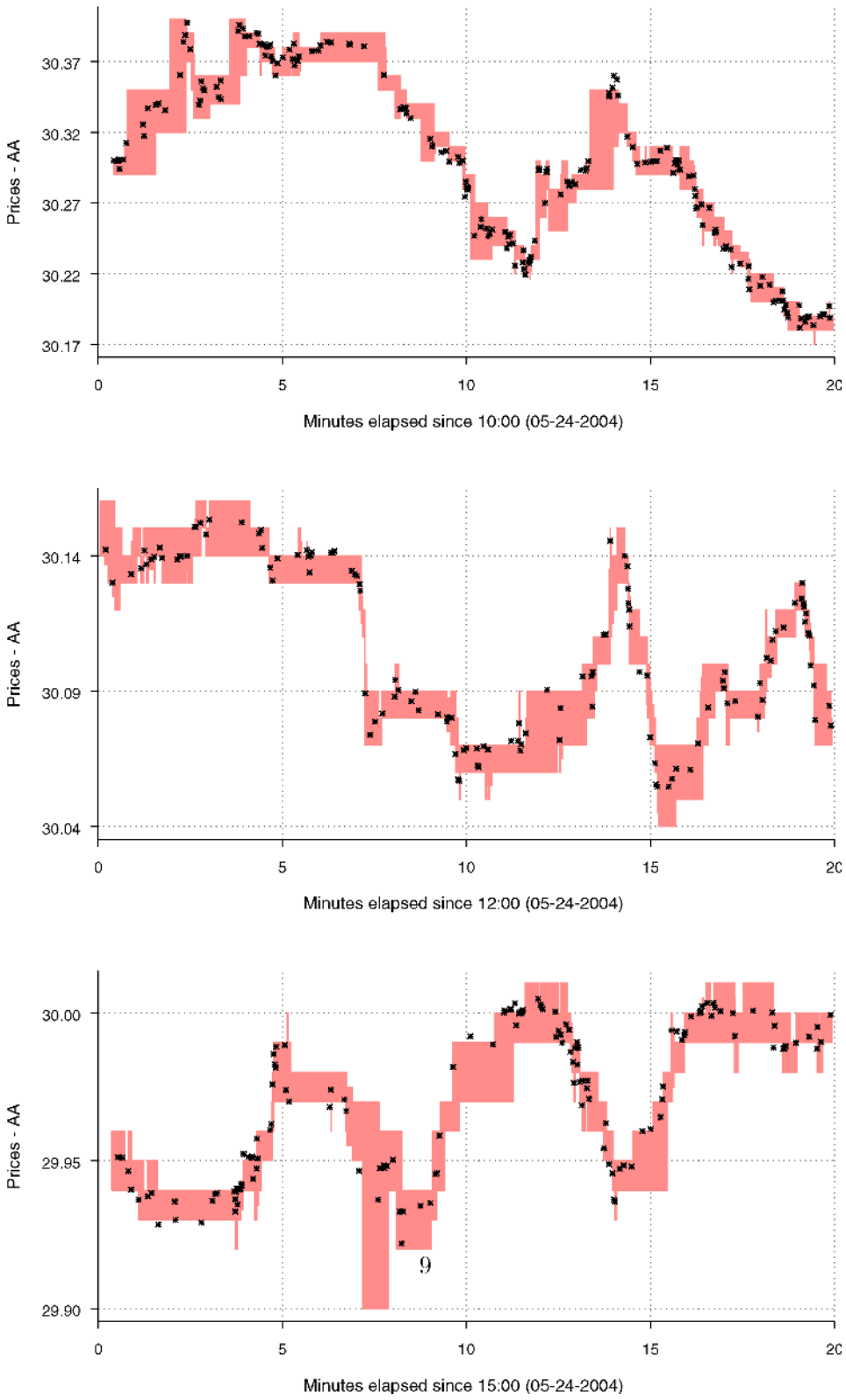

Figure 8. The Estimated Efficient Price Over Three 20-Minute Subperiods on April 24, 2004 for AA. The shaded area shows the best bid and ask quotes, and the star (✱) represent actual transaction prices.

As shown by Hansen (2005), the coefficients of $\mathbf{C}(L) = \mathbf{C}_0 + \mathbf{C}_1 L + \mathbf{C}_2 L^2 + \cdots$ are given recursively by the Yule–Walker type equation

$$\Delta \mathbf{C}_i = \mathbf{\Pi} \mathbf{C}_{i-1} + \mathbf{\Gamma}_1 \Delta \mathbf{C}_{i-1} + \cdots + \mathbf{\Gamma}_{k-1} \Delta \mathbf{C}_{i-k+1} \quad \text{for } i \geq 1, \tag{9}$$

where $\Delta \mathbf{C}_i = \mathbf{C}_i - \mathbf{C}_{i-1}$, $\mathbf{C}_0 = \mathbf{I} - \mathbf{C}$, and $\mathbf{C}_{-1} = \mathbf{C}_{-2} = \cdots = -\mathbf{C}$. An interesting question is how transaction prices and quotes are affected by a change in the efficient price. This question can be addressed with reference to the Granger representation.

Although the Granger representation theorem is an algebraic result that does not rely on distributional assumptions, interpretations drawn from it do rely on additional assumptions (see Hansen 2005). Under the restrictive Gaussian assumptions, with homoscedastic innovations, $\mathbf{\Omega} \equiv \text{var}(\boldsymbol{\varepsilon}_{t_i})$, it follows that

$$\begin{aligned}\frac{d\mathbf{p}_{t_{i+h}}}{dp^*_{t_i}} &= \frac{d\mathbf{p}_{t_{i+h}}}{d\boldsymbol{\varepsilon}_{t_i}} \times \frac{d\boldsymbol{\varepsilon}_{t_i}}{d(\boldsymbol{\alpha}'_\perp \boldsymbol{\varepsilon}_{t_i})} \times \frac{d(\boldsymbol{\alpha}'_\perp \boldsymbol{\varepsilon}_{t_i})}{dp^*_{t_i}} \\ &= (\mathbf{C} + \mathbf{C}_h) \times \mathbf{\Omega}\boldsymbol{\alpha}_\perp \frac{1}{\boldsymbol{\alpha}'_\perp \mathbf{\Omega}\boldsymbol{\alpha}_\perp} \times \boldsymbol{\alpha}'_\perp \mathbf{\Gamma}\boldsymbol{\beta}_\perp \\ &= \delta(\mathbf{C} + \mathbf{C}_h)\mathbf{\Omega}\boldsymbol{\alpha}_\perp = \boldsymbol{\iota} + \delta\mathbf{C}_h\mathbf{\Omega}\boldsymbol{\alpha}_\perp,\end{aligned}$$

where

$$\delta \equiv \frac{\boldsymbol{\alpha}'_\perp \mathbf{\Gamma}\boldsymbol{\beta}_\perp}{\boldsymbol{\alpha}'_\perp \mathbf{\Omega}\boldsymbol{\alpha}_\perp}.$$

Given the lack of results that hold under weaker (and more appropriate) assumptions, we compute estimates of $\boldsymbol{\iota} + \delta\mathbf{C}_h\mathbf{\Omega}\boldsymbol{\alpha}_\perp$, $h = 0, 1, \ldots,$ and interpret these as approximate impulse response functions (IRFs). Thus these "projections" should be viewed as only indicative of the dynamic effect on bid, ask, and transaction prices as a response to a change in the efficient price. The long-run effect is unity for all prices, as it should be, because $d\mathbf{p}_{t_{i+h}}/dp^*_{t_i} \to \boldsymbol{\iota}$ as $h \to \infty$. This follows from $\mathbf{C}_h \to \mathbf{0}$ as $h \to \infty$, which holds under the standard $I(1)$ assumptions (see, e.g., Hansen 2005).

Figure 9 displays the estimated IRFs for transaction prices, bid and ask quotes as a response to a change in the efficient price. These results are based on the daily estimates for 2004. The solid lines correspond to the average IRFs, the dashed lines are the median IRF, and the shaded area is bounded by the 5% and 95% quantiles across the days in 2004. Although most of the price change occurs instantaneously, the estimated IRFs suggest that it takes about 10 transactions before the full effect is absorbed into transaction and quoted prices for AA, and only 3–5 transactions for MSFT. Also note that the instantaneous effect on quotes is larger for MSFT than for AA ($\approx$.9 compared with $\approx$.8), and the total effect on quoted prices is absorbed quicker. This is likely explained by the differences between specialist quotes on the NYSE and the competitive quotes on NASDAQ.

7. SUMMARY AND CONCLUDING REMARKS

We have analyzed the properties of market microstructure noise and its effect on empirical measures of volatility. Our use of kernel-based estimators revealed several important properties about market microstructure noise, and we have shown that kernel-based estimators are very useful in this context.

More importantly, our empirical analysis uncovered several characteristics of market microstructure noise, the most notable of which are that the noise process is time-dependent and that the noise process is correlated with the latent efficient returns. These results were established for both transaction data and quotation data and were found to hold for intraday returns based on both calendar time and tick time sampling.

We also evaluated the accuracy of distributional results based on an assumption that there is no market microstructure noise. We showed that a "no-noise" RMSE based on the results of Barndorff-Nielsen and Shephard (2002) provides a reasonably accurate approximation when intraday returns are sampled at low frequencies, such as 20-minute sampling. At low sampling frequencies, small-sample issues can arise such that coverage probabilities may be inaccurate (see, e.g., Gonçalves and Meddahi 2005), but market microstructure noise is not to be blamed for such issues. When intraday returns are sampled at higher frequencies, the accuracy of "no-noise" approximations is likely to deteriorate. The analogous "no-noise" confidence interval about $RV^{(m)}_{\text{AC}_1}$ yields a more accurate approximation than that of $RV^{(m)}$, but both are very misleading when intraday returns are sampled at high frequencies.

Several results in the literature (including our theoretical results given in Sec. 3) that analyze volatility estimation from high-frequency data contaminated with market microstructure noise have assumed that the noise process is independent of the efficient price and uncorrelated in time. Our empirical results suggest that the implications of these assumptions may hold (at least approximately) when intraday returns are sampled at relatively low frequencies. Thus the conclusions of these articles may hold as long as intraday returns are not sampled more frequently than, say, every 30 ticks. On the other hand, our empirical results have also shown that sampling at ultra-high frequencies, such as every few ticks, necessitates more general assumptions about the dependence structure of market microstructure noise. We established these results in Section 4, where we used a general specification for the noise process that can accommodate both types of dependency. Our cointegration analysis produced results consistent with both forms of dependencies. Volatility signature plots revealed a negative noise–price correlation in quoted price series, and the cointegration analysis showed that the negative correlation is found in all price series, including transaction prices.

Although the main focus of our analysis has been on the properties of market microstructure noise, our analysis also provides some insight into the problem of volatility estimation in the presence of market microstructure noise. Under the independent noise assumption, our comparison of $RV^{(m)}_{\text{AC}_1}$ to the standard measure of realized variance revealed a substantial improvement in the precision, because the theoretical reduction of the RMSE is about 25–50%. These gains were achieved with a simple bias correction that incorporates the first-order autocovariance, and additional improvements are possible with more sophisticated corrections of the realized variance. For example, the kernel estimators of Barndorff-Nielsen et al. (2004) and subsample estimators of Zhang et al. (2005) and Zhang (2004) have better asymptotic properties than $RV^{(m)}_{\text{AC}_1}$. Among the estimators that we have analyzed in this article, $RV^{(1\text{ tick})}_{\text{ACNW}_{30}}$ appears to capture the time dependence found in the noise. However, there are potential gains from allowing for some bias in exchange for a reduction of the variance. This may particularly be the case in recent years, where we find the noise (and hence the bias) to be very small. Thus, correcting for a smaller number of autocovariances may be better in terms of the RMSE, so $RV^{(1\text{ tick})}_{\text{ACNW}_{10}}$ may be a better estimator than $RV^{(1\text{ tick})}_{\text{ACNW}_{30}}$, although the former is more likely to be biased.

Although the literature has made great progress on the problem of estimating the quadratic variation in the presence of

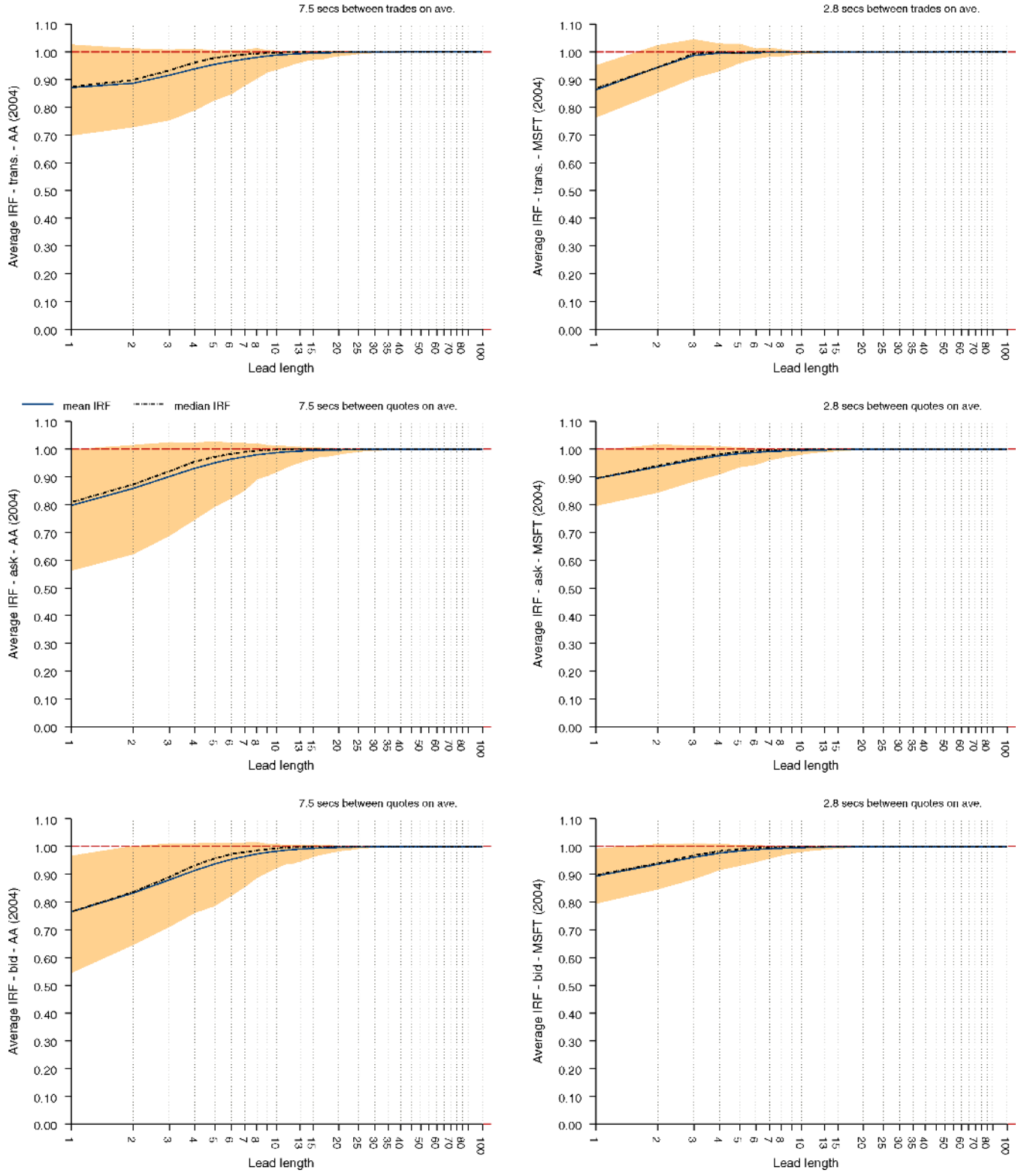

Figure 9. The Estimated IRFs (average over days in the year 2004) for Transaction Prices and Bid and Ask Quotes That Show the Dynamic Effect of an Increase in the Efficient Price. The left panels are for AA, and the right panels are for MSFT. The solid lines correspond to the average IRFs, the dashed lines are the median IRF, and the shaded area is bounded by the 5% and 95% quantiles.

noise, many aspects of this problem are not yet fully understood, mainly because market microstructure noise appears to be more complex than can be accommodated by a simple specification for the noise. Furthermore, the existing estimators of volatility may be improved on by incorporating additional information, such as volume, number of transactions per day, and

the information from limit order books. Exploiting the discreteness of the data (as in Large 2005; Oomen 2006) is another possible avenue for improving existing estimators.

ACKNOWLEDGMENTS

The authors thank seminar participants at University of Copenhagen, University of Aarhus, Nuffield College, Carnegie Mellon University, the Econometric Forecasting and High-Frequency Data Analysis Symposium in Singapore, the conference Innovations in Financial Econometrics in Celebration of the 2003 Nobel, and the 2005 Princeton–Chicago Conference on the Econometrics of High-Frequency Financial Data for many valuable comments. They are particularly grateful to Frederico Bandi, Albert Chun, Eric Ghysels, Joel Hasbrouck, Jeremy Large, Bruce Lehmann, Per Mykland, Neil Shephard, two anonymous referees, and Torben G. Andersen (the editor) for their suggestions that improved this manuscript. All errors remain the responsibility of the authors. Financial support from the Danish Research Agency (grant no. 24-00-0363) is gratefully acknowledged.

APPENDIX A: PROOFS

As stated earlier, we condition on $\{\sigma^2(t)\}$ in our analysis. Thus, without loss of generality, we treat $\sigma^2(t)$ as a deterministic function in our derivations.

Proof of Lemma 1

First, we note that $\tilde{p}(\tau)$ is continuous and piecewise linear on $[a, b]$. Thus $\tilde{p}(\tau)$ satisfies the Lipschitz condition, $\exists\, \delta > 0$, such that $|\tilde{p}(\tau) - \tilde{p}(\tau + \epsilon)| \le \delta\epsilon$ for all $\epsilon > 0$ and all τ. With $\epsilon = (b-a)/m$, we have that

$$\sum_{i=1}^{m} y_{i,m}^2 \le \sum_{i=1}^{m} \delta^2 (b-a)^2 m^{-2} = \frac{\delta^2 (b-a)^2}{m},$$

which demonstrates that $RV^{(m)} \xrightarrow{p} 0$ as $m \to \infty$ and that the QV of $\tilde{p}(\tau)$ is 0 with probability 1.

Proof of Theorem 1

From the identities $RV^{(m)} = \sum_{i=1}^{m} [y_{i,m}^{*2} + 2y_{i,m}^* e_{i,m} + e_{i,m}^2]$, and $E(\sum_{i=1}^{m} y_{i,m}^* e_{i,m}) = \rho_m$ it follows that the bias is given by $\sum_{i=1}^{m} 2E(y_{i,m}^* e_{i,m}) + E(e_{i,m}^2) = 2\rho_m + mE(e_{i,m}^2)$. The result of the theorem now follows from the identity

$$E(e_{i,m}^2) = E\big[u(i\delta_m) - u((i-1)\delta_m)\big]^2 = 2[\pi(0) - \pi(\delta_m)],$$

given Assumption 2.

Proof of Corollary 1

Because $m = (b-a)/\delta_m$, we have that

$$\lim_{m\to\infty} m[\pi(0) - \pi(\delta_m)] = \lim_{m\to\infty} (b-a)\frac{\pi(0) - \pi(\delta_m)}{\delta_m} = -(b-a)\pi'(0),$$

under the assumption that $\pi'(0)$ is well defined.

Proof of Lemma 2

The bias follows directly from the decomposition $y_{i,m}^2 = y_{i,m}^{*2} + e_{i,m}^2 + 2y_{i,m}^* e_{i,m}$, because $E(e_{i,m}^2) = E(u_{i,m} - u_{i-1,m})^2 = E(u_{i,m}^2) + E(u_{i-1,m}^2) - 2E(u_{i,m}u_{i-1,m}) = 2\omega^2$, where we have used Assumption 3(a) and (b). Similarly, we see that

$$\operatorname{var}\big(RV^{(m)}\big) = \operatorname{var}\left(\sum_{i=1}^{m} y_{i,m}^{*2}\right) + \operatorname{var}\left(\sum_{i=1}^{m} e_{i,m}^2\right) + 4\operatorname{var}\left(\sum_{i=1}^{m} y_{i,m}^* e_{i,m}\right),$$

because the three sums are uncorrelated. The first sum involves uncorrelated terms such that $\operatorname{var}(\sum_{i=1}^{m} y_{i,m}^{*2}) = \sum_{i=1}^{m} \operatorname{var}(y_{i,m}^{*2}) = 2\sum_{i=1}^{m} \sigma_{i,m}^4$, where the last equality follows from the Gaussian assumption. For the second sum, we find that

$$\begin{aligned} E(e_{i,m}^4) &= E(u_{i,m} - u_{i-1,m})^4 = E(u_{i,m}^2 + u_{i-1,m}^2 - 2u_{i,m}u_{i-1,m})^2 \\ &= E(u_{i,m}^4 + u_{i-1,m}^4 + 4u_{i,m}^2 u_{i-1,m}^2 + 2u_{i,m}^2 u_{i-1,m}^2) + 0 \\ &= 2\mu_4 + 6\omega^4 \end{aligned}$$

and

$$\begin{aligned} E(e_{i,m}^2 e_{i+1,m}^2) &= E(u_{i,m} - u_{i-1,m})^2 (u_{i+1,m} - u_{i,m})^2 \\ &= E(u_{i,m}^2 + u_{i-1,m}^2 - 2u_{i,m}u_{i-1,m}) \\ &\quad \times (u_{i+1,m}^2 + u_{i,m}^2 - 2u_{i+1,m}u_{i,m}) \\ &= E(u_{i,m}^2 + u_{i-1,m}^2)(u_{i+1,m}^2 + u_{i,m}^2) + 0 \\ &= \mu_4 + 3\omega^4, \end{aligned}$$

where we have used Assumption 3(a)–(c). Thus $\operatorname{var}(e_{i,m}^2) = 2\mu_4 + 6\omega^4 - [E(e_{i,m}^2)]^2 = 2\mu_4 + 2\omega^4$ and $\operatorname{cov}(e_{i,m}^2, e_{i+1,m}^2) = \mu_4 - \omega^4$. Because $\operatorname{cov}(e_{i,m}^2, e_{i+h,m}^2) = 0$ for $|h| \ge 2$, it follows that

$$\begin{aligned} \operatorname{var}\left(\sum_{i=1}^{m} e_{i,m}^2\right) &= \sum_{i=1}^{m} \operatorname{var}(e_{i,m}^2) + \sum_{\substack{i,j=1 \\ i\neq j}}^{m} \operatorname{cov}(e_{i,m}^2, e_{j,m}^2) \\ &= m(2\mu_4 + 2\omega^4) + 2(m-1)(\mu_4 - \omega^4) \\ &= 4m\mu_4 - 2(\mu_4 - \omega^4). \end{aligned}$$

The last sum involves uncorrelated terms such that

$$\operatorname{var}\left(\sum_{i=1}^{m} e_{i,m} y_{i,m}^*\right) = \sum_{i=1}^{m} \operatorname{var}(e_{i,m} y_{i,m}^*) = 2\omega^2 \sum_{i=1}^{m} \sigma_{i,m}^2.$$

By the substitution $3\kappa\omega^4 = \mu_4$, we obtain the expression for the variance. The asymptotic normality has been proven by Zhang et al. (2005) using with an argument similar to one that we use for $RV_{\mathrm{AC}_1}^{(m)}$ in our proof of Lemma 3, and that $2\sum_{i=1}^{m} \sigma_{i,m}^4 + 2\omega^2 \sum_{i=1}^{m} \sigma_{i,m}^2 - 4\omega^4 = O(1)$.

Proof of Lemma 3

First, note that $RV^{(m)}_{\mathrm{AC}_1} = \sum_{i=1}^{m} Y_{i,m} + U_{i,m} + V_{i,m} + W_{i,m}$, where

$$Y_{i,m} \equiv y^*_{i,m}(y^*_{i-1,m} + y^*_{i,m} + y^*_{i+1,m}),$$
$$U_{i,m} \equiv (u_{i,m} - u_{i-1,m})(u_{i+1,m} - u_{i-2,m}),$$
$$V_{i,m} \equiv y^*_{i,m}(u_{i+1,m} - u_{i-2,m}),$$

and

$$W_{i,m} \equiv (u_{i,m} - u_{i-1,m})(y^*_{i-1,m} + y^*_{i,m} + y^*_{i+1,m}),$$

because $y_{i,m}(y_{i-1,m} + y_{i,m} + y_{i+1,m}) = (y^*_{i,m} + u_{i,m} - u_{i-1,m}) \times (y^*_{i-1,m} + y^*_{i,m} + y^*_{i+1,m} + u_{i+1,m} - u_{i-2,m}) = Y_{i,m} + U_{i,m} + V_{i,m} + W_{i,m}$. Thus the properties of $RV^{(m)}_{\mathrm{AC}_1}$ are given from those of $Y_{i,m}$, $U_{i,m}$, $V_{i,m}$, and $W_{i,m}$. Given Assumptions 1 and 3(a), it follows directly that $E(Y_{i,m}) = \sigma^2_{i,m}$ and $E(U_{i,m}) = E(V_{i,m}) = E(W_{i,m}) = 0$, demonstrating $E[RV^{(m)}_{\mathrm{AC}_1}] = \sum_{i=1}^{m} \sigma^2_{i,m}$. Note that $E(U_{i,m})$ consists of terms $E(u_{i,m}u_{j,m})$, where $i \neq j$, so Assumption 3(a) suffices to establish that the expected value is 0. Given Assumptions 1 and 3(a) and (b), the variance of $RV^{(m)}_{\mathrm{AC}_1}$ is given by

$$\begin{aligned}\operatorname{var}\big[RV^{(m)}_{\mathrm{AC}_1}\big] &= \operatorname{var}\left[\sum_{i=1}^{m} Y_{i,m} + U_{i,m} + V_{i,m} + W_{i,m}\right] \\ &= (1) + (2) + (3) + (4) + (5).\end{aligned}$$

Because all other sums are uncorrelated, the five parts are given by $(1) = \operatorname{var}(\sum_{i=1}^{m} Y_{i,m})$, $(2) = \operatorname{var}(\sum_{i=1}^{m} U_{i,m})$, $(3) = \operatorname{var}(\sum_{i=1}^{m} V_{i,m})$, $(4) = \operatorname{var}(\sum_{i=1}^{m} W_{i,m})$, and $(5) = 2 \times \operatorname{cov}(\sum_{i=1}^{m} V_{i,m}, \sum_{i=1}^{m} W_{i,m})$. We derive the expressions of these five terms as follow:

1. $Y_{i,m} = y^*_{i,m}(y^*_{i-1,m} + y^*_{i,m} + y^*_{i+1,m})$, and, given Assumption 1, it follows that $E[y^{*2}_{i,m}y^{*2}_{j,m}] = \sigma^2_{i,m}\sigma^2_{j,m}$ for $i \neq j$ and $E[y^{*2}_{i,m}y^{*2}_{j,m}] = E[y^{*4}_{i,m}] = 3\sigma^4_{i,m}$ for $i = j$, such that

$$\begin{aligned}\operatorname{var}(Y_{i,m}) &= 3\sigma^4_{i,m} + \sigma^2_{i,m}\sigma^2_{i-1,m} + \sigma^2_{i,m}\sigma^2_{i+1,m} - [\sigma^2_{i,m}]^2 \\ &= 2\sigma^4_{i,m} + \sigma^2_{i,m}\sigma^2_{i-1,m} + \sigma^2_{i,m}\sigma^2_{i+1,m}.\end{aligned}$$

The first-order autocorrelation of $Y_{i,m}$ is

$$\begin{aligned}&E[Y_{i,m}Y_{i+1,m}] \\ &\quad = E\big[y^*_{i,m}(y^*_{i-1,m} + y^*_{i,m} + y^*_{i+1,m}) \\ &\qquad \times y^*_{i+1,m}(y^*_{i,m} + y^*_{i+1,m} + y^*_{i+2,m})\big] \\ &\quad = E[y^*_{i,m}(y^*_{i,m} + y^*_{i+1,m})y^*_{i+1,m}(y^*_{i,m} + y^*_{i+1,m})] + 0 \\ &\quad = 2E[y^{*2}_{i,m}y^{*2}_{i+1,m}] = 2\sigma^2_{i,m}\sigma^2_{i+1,m},\end{aligned}$$

such that $\operatorname{cov}(Y_{i,m}, Y_{i+1,m}) = \sigma^2_{i,m}\sigma^2_{i+1,m}$, whereas $\operatorname{cov}(Y_{i,m}, Y_{i+h,m}) = 0$ for $|h| \geq 2$. Thus

$$\begin{aligned}(1) &= \sum_{i=1}^{m}(2\sigma^4_{i,m} + \sigma^2_{i,m}\sigma^2_{i-1,m} + \sigma^2_{i,m}\sigma^2_{i+1,m}) \\ &\quad + \sum_{i=2}^{m}\sigma^2_{i,m}\sigma^2_{i-1,m} + \sum_{i=1}^{m-1}\sigma^2_{i,m}\sigma^2_{i+1,m} \\ &= 2\sum_{i=1}^{m}\sigma^4_{i,m} + 2\sum_{i=1}^{m}\sigma^2_{i,m}\sigma^2_{i-1,m} + 2\sum_{i=1}^{m}\sigma^2_{i,m}\sigma^2_{i+1,m} \\ &\quad - \sigma^2_{1,m}\sigma^2_{0,m} - \sigma^2_{m,m}\sigma^2_{m+1,m} \\ &= 6\sum_{i=1}^{m}\sigma^4_{i,m} - 2\sum_{i=1}^{m}\sigma^2_{i,m}(\sigma^2_{i,m} - \sigma^2_{i-1,m}) \\ &\quad + 2\sum_{i=1}^{m}\sigma^2_{i,m}(\sigma^2_{i+1,m} - \sigma^2_{i,m}) - \sigma^2_{1,m}\sigma^2_{0,m} - \sigma^2_{m,m}\sigma^2_{m+1,m} \\ &= 6\sum_{i=1}^{m}\sigma^4_{i,m} - 2\sum_{i=2}^{m}\sigma^2_{i,m}(\sigma^2_{i,m} - \sigma^2_{i-1,m}) \\ &\quad + 2\sum_{i=1}^{m-1}\sigma^2_{i,m}(\sigma^2_{i+1,m} - \sigma^2_{i,m}) \\ &\quad - \sigma^2_{1,m}\sigma^2_{0,m} - \sigma^2_{m,m}\sigma^2_{m+1,m} - 2\sigma^2_{1,m}(\sigma^2_{1,m} - \sigma^2_{0,m}) \\ &\quad + 2\sigma^2_{m,m}(\sigma^2_{m+1,m} - \sigma^2_{m,m}) \\ &= 6\sum_{i=1}^{m}\sigma^4_{i,m} - 2\sum_{i=1}^{m-1}(\sigma^2_{i+1,m} - \sigma^2_{i,m})^2 - 2(\sigma^4_{1,m} + \sigma^4_{m,m}) \\ &\quad + \sigma^2_{1,m}\sigma^2_{0,m} + \sigma^2_{m,m}\sigma^2_{m+1,m}.\end{aligned}$$

2. $U_{i,m} = (u_{i,m} - u_{i-1,m})(u_{i+1,m} - u_{i-2,m})$, and, from $E(U^2_{i,m}) = E(u_{i,m} - u_{i-1,m})^2E(u_{i+1,m} - u_{i-2,m})^2$, it follows that $\operatorname{var}(U_{i,m}) = 4\omega^4$. The first- and second-order autocovariances are given by

$$\begin{aligned}E(U_{i,m}U_{i+1,m}) &= E\big[(u_{i,m} - u_{i-1,m})(u_{i+1,m} - u_{i-2,m}) \\ &\qquad \times (u_{i+1,m} - u_{i,m})(u_{i+2,m} - u_{i-1,m})\big] \\ &= E[u_{i-1,m}u_{i+1,m}u_{i+1,m}u_{i-1,m}] + 0 \\ &= \omega^4\end{aligned}$$

and

$$\begin{aligned}E(U_{i,m}U_{i+2,m}) &= E\big[(u_{i,m} - u_{i-1,m})(u_{i+1,m} - u_{i-2,m}) \\ &\qquad \times (u_{i+2,m} - u_{i+1,m})(u_{i+3,m} - u_{i,m})\big] \\ &= E[u_{i,m}u_{i+1,m}u_{i+1,m}u_{i,m}] + 0 \\ &= \omega^4,\end{aligned}$$

whereas $E(U_{i,m}U_{i+h,m}) = 0$ for $|h| \geq 3$. Thus $(2) = m4\omega^4 + 2(m-1)\omega^4 + 2(m-2)\omega^4 = 8\omega^4 m - 6\omega^4$.

3. $V_{i,m} = y^*_{i,m}(u_{i+1,m} - u_{i-2,m})$ such that $E(V^2_{i,m}) = \sigma^2_{i,m} \times 2\omega^2$ and $E[V_{i,m}V_{i+h,m}] = 0$ for all $h \neq 0$. Thus $(3) = \operatorname{var}(\sum_{i=1}^{m} V_{i,m}) = 2\omega^2 \sum_{i=1}^{m} \sigma^2_{i,m}$.

4. $W_{i,m} = (u_{i,m} - u_{i-1,m})(y^*_{i-1,m} + y^*_{i,m} + y^*_{i+1,m})$ such that $E(W^2_{i,m}) = 2\omega^2(\sigma^2_{i-1,m} + \sigma^2_{i,m} + \sigma^2_{i+1,m})$. The first-order autocovariance equals

$$\begin{aligned}\operatorname{cov}(W_{i,m}, W_{i+1,m}) &= E[-u^2_{i,m}(y^{*2}_{i,m} + y^{*2}_{i+1,m})] \\ &= -\omega^2(\sigma^2_{i,m} + \sigma^2_{i+1,m}),\end{aligned}$$

whereas $\operatorname{cov}(W_{i,m}, W_{i+h,m}) = 0$ for $|h| \geq 2$. Thus

$$\begin{aligned}(4) &= \sum_{i=1}^{m}\Bigg[2\omega^2(\sigma_{i-1,m}^2+\sigma_{i,m}^2+\sigma_{i+1,m}^2) \\ &\qquad -\sum_{i=2}^{m}\omega^2(\sigma_{i,m}^2+\sigma_{i-1,m}^2)-\sum_{i=1}^{m-1}\omega^2(\sigma_{i,m}^2+\sigma_{i+1,m}^2)\Bigg] \\ &= \omega^2\sum_{i=1}^{m}(\sigma_{i-1,m}^2+\sigma_{i+1,m}^2) \\ &\quad +\omega^2[\sigma_{1,m}^2+\sigma_{0,m}^2+\sigma_{m,m}^2+\sigma_{m+1,m}^2] \\ &= 2\omega^2\sum_{i=1}^{m}\sigma_{i,m}^2+\omega^2[\sigma_{0,m}^2-\sigma_{m,m}^2+\sigma_{m+1,m}^2-\sigma_{1,m}^2] \\ &\quad +\omega^2[\sigma_{1,m}^2+\sigma_{0,m}^2+\sigma_{m,m}^2+\sigma_{m+1,m}^2] \\ &= 2\omega^2\sum_{i=1}^{m}\sigma_{i,m}^2+2\omega^2[\sigma_{0,m}^2+\sigma_{m+1,m}^2].\end{aligned}$$

5. The autocovariances between the last two terms are given by

$$\begin{aligned}E[V_{i,m}W_{i+h,m}] = E\big[&y_{i,m}^*(u_{i+1,m}-u_{i-2,m})(u_{i+h,m}-u_{i-1+h,m}) \\ &\times(y_{i-1+h,m}^*+y_{i+h,m}^*+y_{i+1+h,m}^*)\big],\end{aligned}$$

showing that $\operatorname{cov}(V_{i,m}, W_{i\pm1,m}) = \omega^2\sigma_{i,m}^2$, whereas all other covariances are 0. From this, we conclude that

$$\begin{aligned}(5) &= 2\Bigg[2\sum_{i=1}^{m}\omega^2\sigma_{i,m}^2-\omega^2(\sigma_{1,m}^2+\sigma_{m,m}^2)\Bigg] \\ &= 4\omega^2\sum_{i=1}^{m}\sigma_{i,m}^2-2\omega^2(\sigma_{1,m}^2+\sigma_{m,m}^2).\end{aligned}$$

Adding up the five terms, we find that

$$\begin{aligned}&6\sum_{i=1}^{m}\sigma_{i,m}^4-2\sum_{i=1}^{m-1}(\sigma_{i+1,m}^2-\sigma_{i,m}^2)^2-2(\sigma_{1,m}^4+\sigma_{m,m}^4) \\ &\quad+\sigma_{1,m}^2\sigma_{0,m}^2+\sigma_{m,m}^2\sigma_{m+1,m}^2+8\omega^4m-6\omega^4 \\ &\quad+2\omega^2\sum_{i=1}^{m}\sigma_{i,m}^2+2\omega^2\sum_{i=1}^{m}\sigma_{i,m}^2+2\omega^2[\sigma_{0,m}^2+\sigma_{m+1,m}^2] \\ &\quad+4\omega^2\sum_{i=1}^{m}\sigma_{i,m}^2-2\omega^2[\sigma_{1,m}^2+\sigma_{m,m}^2] \\ &= 8\omega^4m+8\omega^2\sum_{i=1}^{m}\sigma_{i,m}^2-6\omega^4+6\sum_{i=1}^{m}\sigma_{i,m}^4+R_m,\end{aligned}$$

where

$$\begin{aligned}R_m \equiv &-2\sum_{i=1}^{m}(\sigma_{i+1,m}^2-\sigma_{i,m}^2)^2-2(\sigma_{1,m}^4+\sigma_{m,m}^4) \\ &+\sigma_{1,m}^2\sigma_{0,m}^2+\sigma_{m,m}^2\sigma_{m+1,m}^2 \\ &+2\omega^2(\sigma_{0,m}^2-\sigma_{1,m}^2+\sigma_{m+1,m}^2-\sigma_{m,m}^2).\end{aligned}$$

Under BTS, it follows immediately that $R_m = O(m^{-2})$. Under CTS, we use the Lipschitz condition, which states that $\exists\epsilon > 0$ such that $|\sigma^2(t)-\sigma^2(t+h)| \leq \epsilon h$ for all t and all h. This shows that $|\sigma_{i,m}^2| = |\int_{t_{i-1,m}}^{t_{i,m}}\sigma^2(s)\,ds| \leq \delta\sup_{t_{i-1,m}\leq s\leq t_{i,m}}\sigma^2(s) = O(m^{-1})$, because $\delta = \delta_{i,m} = (b-a)/m = O(m^{-1})$ under CTS, and

$$\begin{aligned}|\sigma_{i,m}^2-\sigma_{i-1,m}^2| &= \left|\int_{t_{i-1,m}}^{t_{i,m}}\sigma^2(s)-\sigma^2(s-\delta)\,ds\right| \\ &\leq \int_{t_{i-1,m}}^{t_{i,m}}|\sigma^2(s)-\sigma^2(s-\delta)|\,ds \\ &\leq \delta\sup_{t_{i-1,m}\leq s\leq t_{i,m}}|\sigma^2(s)-\sigma^2(s-\delta)| \\ &\leq \delta^2\epsilon = O(m^{-2}).\end{aligned}$$

Thus $\sum_{i=1}^{m}(\sigma_{i+1,m}^2-\sigma_{i,m}^2)^2 \leq m\cdot(\delta\frac{\epsilon}{m})^2 = O(m^{-3})$, which proves that $R_m = O(m^{-2})$, under CTS.

The asymptotic normality is established by expressing $RV_{\mathrm{AC}_1}^{(m)}$ as a sum of a martingale difference sequence. Let $u_{i,m} = u(t_{i,m})$, and define the sigma algebra $\mathcal{F}_{i,m} = \sigma(y_{i,m}^*, y_{i-1,m}^*, \ldots, u_{i,m}, u_{i-1,m}, \ldots)$. First, note that $y_{i,m}(y_{i-1,m}+y_{i,m}+y_{i+1,m}) = \sigma_{i,m}^2+\xi_{i-1,m}^{(1)}+\xi_{i,m}^{(2)}+\xi_{i+1,m}^{(3)}$, where

$$\begin{aligned}\xi_{i-1,m}^{(1)} &\equiv -u_{i-1,m}y_{i-1,m}^*+u_{i-1,m}u_{i-2,m}, \\ \xi_{i,m}^{(2)} &\equiv y_{i,m}^*y_{i,m}^*-\sigma_{i,m}^2+y_{i,m}^*y_{i-1,m}^*-y_{i,m}^*u_{i-2,m} \\ &\qquad +u_{i,m}y_{i-1,m}^*+u_{i,m}y_{i,m}^*-u_{i,m}u_{i-2,m}-u_{i-1,m}y_{i,m}^*,\end{aligned}$$

and

$$\begin{aligned}\xi_{i+1,m}^{(3)} \equiv\ &y_{i,m}^*y_{i+1,m}^*+y_{i,m}^*u_{i+1,m}+u_{i,m}y_{i+1,m}^* \\ &+u_{i,m}u_{i+1,m}-u_{i-1,m}y_{i+1,m}^*-u_{i-1,m}u_{i+1,m}.\end{aligned}$$

Thus, if we define $\xi_{i,m} \equiv (\xi_{i,m}^{(1)}+\xi_{i,m}^{(2)}+\xi_{i,m}^{(3)})$ (and use the conventions $\xi_{0,m}^{(2)} = \xi_{0,m}^{(3)} = \xi_{1,m}^{(3)} = \xi_{m,m}^{(1)} = \xi_{m+1,m}^{(1)} = \xi_{m+1,m}^{(2)} = 0$), then it follows that $[RV_{\mathrm{AC}_1}^{(m)}-IV] = \sum_{i=0}^{m+1}\xi_{i,m}$, where $\{\xi_{i,m}, \mathcal{F}_{i,m}\}_{i=0}^{m+1}$ is a martingale difference sequence that is squared-integrable, because

$$E(\xi_{i,m}^2) = \begin{cases}\omega^2\sigma_0^2+\omega^4<\infty & \text{for } i=0 \\ 2\omega^4+\sigma_0^2\sigma_1^2+\omega^2\sigma_0^2+2\omega^2\sigma_1^2+\sigma_1^4<\infty & \text{for } i=1 \\ 2\sigma_{i,m}^4+4\sigma_{i,m}^2\sigma_{i-1,m}^2+4\sigma_{i,m}^2\omega^2 & \\ \quad+4\sigma_{i-1,m}^2\omega^2+8\omega^4<\infty & \text{for } 1<i<m \\ \sigma_m^4+4\sigma_m^2\sigma_{m-1}^2+6\omega^2\sigma_m^2+4\omega^2\sigma_{m-1}^2+5\omega^4<\infty & \text{for } i=m \\ \sigma_m^2\sigma_{m+1}^2+2\omega^2\sigma_{m+1}^2+2\omega^4<\infty & \text{for } i=m+1.\end{cases}$$

Because $m^{-1/2}[RV_{\mathrm{AC}_1}^{(m)}-IV] = m^{-1/2}\sum_{i=0}^{m+1}\xi_{i,m}$, we can apply the central limit theorem for squared-integrable martingales (see Shiryaev 1995, p. 543, thm. 4) where the only remaining condition to be verified is the conditional Lindeberg con-

dition,

$$\sum_{i=0}^{m+1} E\big[m^{-1}\xi_{i,m}^2 \mathbb{1}_{\{|m^{-1/2}\xi_{i,m}|>\varepsilon\}} \big| \mathcal{F}_{i-1,m}\big] \xrightarrow{p} 0 \quad \text{as } m \to \infty.$$

Because $E[\xi_{i,m}^2 \mathbb{1}_{\{|m^{-1/2}\xi_{i,m}|>\varepsilon\}}]$ is bounded by $E[\xi_{i,m}^2] < \infty$ and $\sup_i P(\mathbb{1}_{\{|\xi_{i,m}|>\varepsilon\sqrt{m}\}} = 0) \to 0$, for all $\varepsilon > 0$, it follows that

$$\begin{aligned} E\left| m^{-1} \sum_{i=0}^{m+1} \xi_{i,m}^2 \mathbb{1}_{\{|m^{-1/2}\xi_{i,m}|>\varepsilon\}} - 0 \right| \\ \le m^{-1} \sum_{i=0}^{m+1} \left| E\big[\xi_{i,m}^2 \mathbb{1}_{\{|m^{-1/2}\xi_{i,m}|>\varepsilon\}}\big] - 0 \right| \\ \to 0 \quad \text{as } m \to \infty. \end{aligned}$$

The Lindeberg condition now follows because convergence in $\mathcal{L}_1$ implies convergence in probability.

Proof of Corollary 2

The MSEs are given from Lemmas 2 and 3, because BTS implies that the $O(m^{-2})$ term in Lemma 3 (see the proof of Lemma 3) is given by

$$R_m = 0 - 2\left(\frac{IV^2}{m^2} + \frac{IV^2}{m^2}\right) + \frac{IV^2}{m^2} + \frac{IV^2}{m^2} + 0 = -\frac{2IV^2}{m^2}.$$

Equating $\partial \operatorname{MSE}(RV^{(m)})/\partial m \propto 4\lambda^2 m + 6\lambda^2 - m^{-2}$ with 0 yields the first-order condition of the corollary, and the second result follows similarly from $\partial \operatorname{MSE}(RV_{\mathrm{AC}_1}^{(m)})/\partial m \propto 4\lambda^2 - 3m^{-2} + 2m^{-3}$.

Proof of Theorem 2

Under CTS, we can define $\delta_m = (b-a)/m = t_{i,m} - t_{i-1,m}$ for all $i = 1, \ldots, m$. To simplify our notation, let $u_{i,m} \equiv u(t_{i,m})$ and note that

$$\begin{aligned} E(e_{i,m}e_{i+h,m}) &= E[u_{i,m} - u_{i-1,m}][u_{i+h,m} - u_{i+h-1,m}] \\ &= 2\pi(h\delta_m) - \pi((h-1)\delta_m) - \pi((h+1)\delta_m) \\ &= \big[\pi(h\delta_m) - \pi((h+1)\delta_m)\big] \\ &\quad - \big[\pi((h-1)\delta_m) - \pi(h\delta_m)\big], \end{aligned}$$

such that

$$\begin{aligned} \sum_{h=1}^{q_m} E(e_{i,m}e_{i+h,m}) \\ = \big[\pi(q_m\delta_m) - \pi((q_m+1)\delta_m)\big] - [\pi(0) - \pi(\delta_m)], \end{aligned}$$

where the first term equals 0 given Assumption 4. Further, by Assumption 4, we have that $y_{i,m}^*$ is uncorrelated with the noise process when separated in time by at least θ_0. Therefore,

$$\begin{aligned} 0 &= E\big[y_{i,m}^*\big(u_{i+q_m,m} - u_{i-q_m-1,m}\big)\big] \\ &= E\big[y_{i,m}^*\big(e_{i+q_m,m} + \cdots + e_{i-q_m,m}\big)\big] \end{aligned}$$

and, similarly,

$$0 = E\big[e_{i,m}\big(y_{i+q_m,m}^* + \cdots + y_{i-q_m,m}^*\big)\big],$$

which implies that

$$\rho_m = \sum_{i=1}^{m} E[y_{i,m}^* e_{i,m}] = -\sum_{i=1}^{m}\sum_{h=1}^{q_m} E[y_{i,m}^* e_{i+h,m} + y_{i,m}^* e_{i-h,m}]$$

and

$$\rho_m = \sum_{i=1}^{m} E[y_{i,m}^* e_{i,m}] = -\sum_{i=1}^{m}\sum_{h=1}^{q_m} E[e_{i,m} y_{i+h,m}^* + e_{i,m} y_{i-h,m}^*].$$

For $h \neq 0$, we have that $E(y_{i,m}y_{i+h,m}) = (E[y_{i,m}^* e_{i+h,m} + e_{i,m} \times y_{i+h,m}^*] + E(e_{i,m}e_{i+h,m})$, because $E(y_{i,m}^* y_{i+h,m}^*) = 0$ by Assumption 4. It now follows that

$$\begin{aligned} E\left[\sum_{h=1}^{q_m}\sum_{i=1}^{m} y_{i,m}y_{i-h,m} + y_{i,m}y_{i+h,m}\right] \\ = -2\rho_m - 2m[\pi(0) - \pi(\delta_m)], \end{aligned}$$

which proves that $RV_{\mathrm{AC}_{q_m}}^{(m)}$ is unbiased.

APPENDIX B: CONFIDENCE INTERVAL FOR VOLATILITY SIGNATURE PLOTS

In our empirical analysis, we seek a measure that is informative about the precision of our bias-corrected RVs. A minimum requirement is that the sample average of RV_{AC} is in a neighborhood of average integrated variance,

$$\bar{\sigma}^2 \equiv n^{-1} \sum_{t=1}^{n} IV_t,$$

which may be estimated by

$$\widehat{\bar{\sigma}}^2 = n^{-1} \sum_{t=1}^{n} \widehat{IV}_t,$$

where we use $RV_{\mathrm{ACNW}_{30},t}^{(1\text{ tick})} \equiv (RV_{\mathrm{ACNW}_{30},t}^{(1\text{ tick}),\text{ tr}} + RV_{\mathrm{ACNW}_{30},t}^{(1\text{ tick}),\text{ bid}} + RV_{\mathrm{ACNW}_{30},t}^{(1\text{ tick}),\text{ ask}})/3$ as our choice for $\widehat{IV}_t$, because it is likely to be conditionally unbiased for IV_t. Next, we consider the problem of constructing a confidence interval for $\bar{\sigma}^2$. Andersen, Bollerslev, Diebold, and Labys (2000a, 2003) have shown that the realized variance is approximately log-normally distributed. Here we follow this idea and assume that $\log \widehat{IV}_t \sim \mathrm{N}(\xi, \omega^2)$, such that the unconditional expected value is given by $E[\widehat{IV}_t] = \bar{\sigma}^2 = \exp(\xi + \omega^2/2)$ and $\operatorname{var}(\widehat{IV}_t) = [\exp(\omega^2) - 1]\exp(\omega^2 + 2\xi)$.

Now $\hat{\xi} \pm \sigma_{\hat{\xi}} c_{1-\alpha/2}$ is a $(1-\alpha)$-confidence interval for ξ, where $\hat{\xi} = n^{-1}\sum_{t=1}^{n} \log \widehat{IV}_t$, $\sigma_{\hat{\xi}}^2 \equiv \operatorname{var}(\hat{\xi})$, and $c_{1-\alpha/2}$ is the appropriate quantile from the standard normal distribution. Because $\xi = \log(\mu) - \omega/2$, it follows that

$$\begin{aligned} P[l \le \xi \le u] &= P[l + \omega/2 \le \log(\bar{\sigma}^2) \le u + \omega/2] \\ &= P\big[\exp(l + \omega/2) \le \bar{\sigma}^2 \le \exp(u + \omega/2)\big], \end{aligned}$$

such that the confidence interval for ξ can be converted into one for $\bar{\sigma}^2$. These calculations require that ω^2 be known, whereas in practice we must estimate ω^2. Thus we define $\eta_t \equiv \log \widehat{IV}_t - \hat{\xi}$

and use the Newey–West estimator,

$$\hat{\omega}^2 \equiv \frac{1}{n-1}\sum_{t=1}^{n}\eta_t^2 + 2\sum_{h=1}^{q}\left(1-\frac{h}{q+1}\right)\frac{1}{n-h}\sum_{t=1}^{n-h}\eta_t\eta_{t+h},$$

$$\text{where } q = \operatorname{int}\left[4(n/100)^{2/9}\right],$$

and subsequently set $\hat{\sigma}_{\hat{\xi}}^2 \equiv \hat{\omega}^2/n$. An approximate confidence interval for $\bar{\sigma}^2$ is now given by

$$\mathrm{CI}(\bar{\sigma}^2) \equiv \exp\left(\log(\widehat{\bar{\sigma}}^2) \pm \hat{\sigma}_{\hat{\xi}} c_{1-\alpha/2}\right),$$

which we recenter about $\log(\widehat{\bar{\sigma}}^2)$ (rather than $\hat{\xi} + \hat{\omega}/2$), because this ensures that the sample average of $\widehat{IV}_t$ is inside the confidence interval, $\widehat{\bar{\sigma}}^2 \in \mathrm{CI}(\bar{\sigma}^2)$.

APPENDIX C: ESTIMATION OF COINTEGRATION VECTOR AUTOREGRESSION

To avoid a linear deterministic trend in the price series, we impose the constraint $\boldsymbol{\mu} = \boldsymbol{\alpha}\rho$, where $\boldsymbol{\rho} = (\rho_1, \rho_2)'$. (Although a linear deterministic trend may be quite sensible over a very long period, an estimate of such would be mostly spurious within a single day.) On the other hand, we do not want to entirely exclude the constant, because we want to allow for a bid–ask spread to have a nontrivial expected value, and this is fully accommodated by the restriction $\boldsymbol{\mu} = \boldsymbol{\alpha}\boldsymbol{\rho}$.

Although the model is simple to estimate by least squares when the constant in unrestricted or set to 0, the estimation under a restricted constant is slightly more complicated.

1. Regress $\Delta\mathbf{p}_{t_i}$ on $(\mathbf{p}'_{t_{i-1}}\boldsymbol{\beta}, \boldsymbol{\iota}, \Delta\mathbf{p}'_{t_{i-1}}, \ldots, \Delta\mathbf{p}'_{t_{i-k+1}})'$ by least squares and obtain $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Gamma}}_1, \ldots, \hat{\boldsymbol{\Gamma}}_{k-1})$.
2. Now set $\hat{\boldsymbol{\rho}} = (\hat{\boldsymbol{\alpha}}'\hat{\boldsymbol{\alpha}})^{-1}\hat{\boldsymbol{\alpha}}'\hat{\boldsymbol{\mu}}$. ($-\hat{\rho}_1$ captures average difference between transaction prices and the mid-quotes, and $-\hat{\rho}_2$ captures "average" spread between ask and bid quotes.)
3. Regress $\Delta\mathbf{p}_{t_i}$ on $(\mathbf{p}'_{t_{i-1}}\boldsymbol{\beta} + \hat{\boldsymbol{\rho}}', \Delta\mathbf{p}'_{t_{i-1}}, \ldots, \Delta\mathbf{p}'_{t_{i-k+1}})'$ by least squares and obtain $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\Gamma}}_1, \ldots, \hat{\boldsymbol{\Gamma}}_{k-1})$.

Define $\hat{\boldsymbol{\alpha}}_\perp \equiv \frac{1}{\varsigma}[\boldsymbol{\iota} - \hat{\boldsymbol{\alpha}}(\hat{\boldsymbol{\alpha}}'\hat{\boldsymbol{\alpha}})^{-1}\hat{\boldsymbol{\alpha}}'\boldsymbol{\iota}]$, where $\varsigma \equiv \boldsymbol{\iota}'\boldsymbol{\iota} - \boldsymbol{\iota}'\hat{\boldsymbol{\alpha}} \times (\hat{\boldsymbol{\alpha}}'\hat{\boldsymbol{\alpha}})^{-1}\hat{\boldsymbol{\alpha}}'\boldsymbol{\iota}$. Then it follows that $\hat{\boldsymbol{\alpha}}'\hat{\boldsymbol{\alpha}}_\perp = \frac{1}{\varsigma}[\hat{\boldsymbol{\alpha}}'\boldsymbol{\iota} - \hat{\boldsymbol{\alpha}}'\hat{\boldsymbol{\alpha}}(\hat{\boldsymbol{\alpha}}'\hat{\boldsymbol{\alpha}})^{-1} \times \hat{\boldsymbol{\alpha}}'\boldsymbol{\iota}] = \frac{1}{\varsigma}[\hat{\boldsymbol{\alpha}}'\iota - \hat{\boldsymbol{\alpha}}'\iota] = 0$ and $\boldsymbol{\iota}'\hat{\boldsymbol{\alpha}}_\perp = \frac{1}{\varsigma}[\boldsymbol{\iota}'\boldsymbol{\iota} - \boldsymbol{\iota}'\hat{\boldsymbol{\alpha}}(\hat{\boldsymbol{\alpha}}'\hat{\boldsymbol{\alpha}})^{-1}\hat{\boldsymbol{\alpha}}'\boldsymbol{\iota}] = 1$. In the impulse response analysis, we use the estimator $\hat{\boldsymbol{\Omega}} \equiv \frac{1}{m}\sum_{i=1}^{m}(\hat{\boldsymbol{\varepsilon}}_{t_i} - \bar{\hat{\boldsymbol{\varepsilon}}})(\hat{\boldsymbol{\varepsilon}}_{t_i} - \bar{\hat{\boldsymbol{\varepsilon}}})'$.

[Received February 2004. Revised September 2005.]

REFERENCES

Aït-Sahalia, Y., Mykland, P. A., and Zhang, L. (2005a), "How Often to Sample a Continuous-Time Process in the Presence of Market Microstructure Noise," *Review of Financial Studies*, 18, 351–416.

——— (2005b), "Ultra-High–Frequency Volatility Estimation With Dependent Microstructure Noise," Working Paper 11380, National Bureau of Economic Research.

Amihud, Y., and Mendelson, H. (1987), "Trading Mechanisms and Stock Returns: An Empirical Investigation," *Journal of Finance*, 42, 533–553.

Andersen, T. G., and Bollerslev, T. (1997), "Intraday Periodicity and Volatility Persistence in Financial Markets," *Journal of Empirical Finance*, 4, 115–158.

——— (1998), "Answering the Skeptics: Yes, Standard Volatility Models Do Provide Accurate Forecasts," *International Economic Review*, 39, 885–905.

Andersen, T. G., Bollerslev, T., and Diebold, F. X. (2003), "Some Like It Smooth and Some Like It Rough: Untangling Continuous and Jump Components in Measuring, Modeling and Forecasting Asset Return Volatility," working paper, Northwestern University, Duke University, and University of Pennsylvania.

Andersen, T. G., Bollerslev, T., Diebold, F. X., and Ebens, H. (2001), "The Distribution of Realized Stock Return Volatility," *Journal of Financial Economics*, 61, 43–76.

Andersen, T. G., Bollerslev, T., Diebold, F. X., and Labys, P. (2000a), "Exchange Rate Return Standardized by Realized Volatility Are (Nearly) Gaussian," *Multinational Finance Journal*, 4, 159–179.

——— (2000b), "Great Realizations," *Risk*, 13, 105–108.

——— (2003), "Modeling and Forecasting Realized Volatility," *Econometrica*, 71, 579–625.

Andersen, T. G., Bollerslev, T., and Meddahi, N. (2004), "Analytic Evaluation of Volatility Forecasts," *International Economic Review*, 45, 1079–1110.

Andreou, E., and Ghysels, E. (2002), "Rolling-Sample Volatility Estimators: Some New Theoretical, Simulation, and Empirical Results," *Journal of Business & Economic Statistics*, 20, 363–376.

Andrews, D. W. K. (1991), "Heteroskedasticity- and Autocorrelation-Consistent Covariance Matrix Estimation," *Econometrica*, 59, 817–858.

Awartani, B., Corradi, V., and Distaso, W. (2004), "Testing and Modelling Market Microstructure Effects With an Application to the Dow Jones Industrial Average," unpublished manuscript, Queen Mary, University of London, and University of Exeter.

Bai, X., Russell, J. R., and Tiao, G. C. (2004), "Effects of Non-Normality and Dependence on the Precision of Variance Estimates Using High-Frequency Financial Data," working paper, University of Chicago, Graduate School of Business.

Baillie, R. T., Booth, G. G., Tse, Y., and Zabotina, T. (2002), "Price Discovery and Common Factor Models," *Journal of Financial Markets*, 5, 309–321.

Bandi, F. M., and Phillips, P. C. (2004), "A Simple Approach to the Parametric Estimation of Potentially Nonstationary Diffusions," unpublished manuscript, University of Chicago and Yale University.

Bandi, F. M., and Russell, J. R. (2005), "Microstructure Noise, Realized Volatility, and Optimal Sampling," working paper, University of Chicago, Graduate School of Business.

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., and Shephard, N. (2004), "Regular and Modified Kernel-Based Estimators of Integrated Variance: The Case With Independent Noise," working paper, Stanford University, Dept. of Economics, *http://www.stanford.edu/people/peter.hansen*.

Barndorff-Nielsen, O. E., Nielsen, B., Shephard, N., and Ysusi, C. (1996), "Measuring and Forecasting Financial Variability Using Realised Variance With and Without a Model," in *State Space and Unobserved Components Models*: *Theory and Application*, eds. A. C. Harvey, S. J. Koopman, and N. Shephard, Cambridge, U.K.: Cambridge University Press, pp. 205–235.

Barndorff-Nielsen, O. E., and Shephard, N. (2002), "Econometric Analysis of Realised Volatility and Its Use in Estimating Stochastic Volatility Models," *Journal of the Royal Statistical Society*, Ser. B, 64, 253–280.

——— (2003), "Realized Power Variation and Stochastic Volatility," *Bernoulli*, 9, 243–265.

——— (2004), "Power and Bipower Variation With Stochastic Volatility and Jumps" (with discussion), *Journal of Financial Econometrics*, 2, 1–48.

——— (2006a), "Econometrics of Testing for Jumps in Financial Economics Using Bipower Variation," *Journal of Financial Econometrics*, 4, 41–30.

——— (2006b), "Impact of Jumps on Returns and Realised Variances: Econometric Analysis of Time-Deformed Levy Processes," *Journal of Econometrics*, 131, 217–252.

——— (2007), "Variation, Jumps, Market Frictions and High-Frequency Data in Financial Econometrics," in *Advances in Economics and Econometrics*. *Theory and Applications*, *Ninth World Congress*, eds. R. Blundell, T. Persson, and W. K. Newey, Econometric Society Monographs, Cambridge, U.K.: Cambridge University Press.

Billingsley, P. (1995), *Probability and Measure* (3nd ed.), New York: Wiley.

Black, F. (1976), "Noise," *Journal of Finance*, 41, 529–543.

Bollen, B., and Inder, B. (2002), "Estimating Daily Volatility in Financial Markets Utilizing Intraday Data," *Journal of Empirical Finance*, 9, 551–562.

Bollerslev, T., Kretschmer, U., Pigorsch, C., and Tauchen, G. (2005), "The Dynamics of Bipower Variation, Realized Volatility and Returns," unpublished manuscript, Duke University, Dept. of Economics.

Christensen, K., and Podolskij, M. (2005), "Asymptotic Theory for Range-Based Estimation of Integrated Variance of a Continuous Semi-Martingale," working paper, Aarhus School of Business and Ruhr University of Bochum.

Corsi, F., Zumbach, G., Müller, U., and Dacorogna, M. (2001), "Consistent High-Precision Volatility From High-Frequency Data," *Economic Notes*, 30, 183–204.

Curci, G., and Corsi, F. (2004), "Discrete Sine Transform Approach for Realized Volatility Measurement," unpublished manuscript, University of Southern Switzerland.

Dacorogna, M. M., Gencay, R., Müller, U., Olsen, R. B., and Pictet, O. V. (2001), *An Introduction to High-Frequency Finance*, London: Academic Press.

de Jong, F. (2002), "Measures of Contributions to Price Discovery: A Comparison," *Journal of Financial Markets*, 5, 323–327.

Easley, D., and O'Hara, M. (1987), "Price, Trade Size, and Information in Securities Markets," *Journal of Financial Economics*, 16, 69–90.

——— (1992), "Time and the Process of Security Price Adjustment," *Journal of Finance*, 47, 576–605.

Ebens, H. (1999), "Realized Stock Volatility," Working Paper 420, Johns Hopkins University, Dept. of Economics.

Engle, R., and Sun, Z. (2005), "Forecasting Volatility Using Tick by Tick Data," unpublished manuscript, New York University, Stern School of Business.

Fang, Y. (1996), "Volatility Modeling and Estimation of High-Frequency Data With Gaussian Noise," unpublished doctoral thesis, MIT, Sloan School of Management.

French, K. R., Schwert, G. W., and Stambaugh, R. F. (1987), "Expected Stock Returns and Volatility," *Journal of Financial Economics*, 19, 3–29.

Frijns, B., and Lehnert, T. (2004), "Realized Variance in the Presence of Non-iid Microstructure Noise: A Structural Approach," Working Paper 04-008, Limburg Institute of Financial Economics.

Ghysels, E., Santa-Clara, P., and Valkanov, R. (2006), "Predicting Volatility: Getting the Most Out of Return Data Sampled at Different Frequencies," *Journal of Econometrics*, 131, 59–95.

Glosten, L., and Milgrom, P. (1985), "Bid, Ask, and Transaction Prices in a Specialist Market With Heterogeneously Informed Traders," *Journal of Financial Economics*, 13, 71–100.

Gonçalves, S., and Meddahi, N. (2005), "Bootstrapping Realized Volatility," unpublished manuscript, Département de Sciences Économiques, CIREQ and CIRANO, Université de Montréal.

Gonzalo, J., and Granger, C. W. J. (1995), "Estimation of Common Long-Memory Components in Cointegrated Systems," *Journal of Business & Economic Statistics*, 13, 27–36.

Hansen, P. R. (2005), "Granger's Representation Theorem: A Closed-Form Expression for $I(1)$ Processes," *Econometrics Journal*, 8, 23–38.

Hansen, P. R., and Johansen, S. (1998), *Workbook on Cointegration*, Oxford, U.K.: Oxford University Press.

Hansen, P. R., and Lunde, A. (2003), "An Optimal and Unbiased Measure of Realized Variance Based on Intermittent High-Frequency Data," mimeo prepared for the CIREQ–CIRANO Conference: Realized Volatility, Montreal, November 2003.

——— (2005a), "A Forecast Comparison of Volatility Models: Does Anything Beat a GARCH(1,1)?" *Journal of Applied Econometrics*, 20, 873–889.

——— (2005b), "A Realized Variance for the Whole Day Based on Intermittent High-Frequency Data," *Journal of Financial Econometrics*, 13, 525–544.

——— (2006), "Consistent Ranking of Volatility Models," *Journal of Econometrics*, 131, 97–121.

Hansen, P. R., Lunde, A., and Nason, J. M. (2003), "Choosing the Best Volatility Models: The Model Confidence Set Approach," *Oxford Bulletin of Economics and Statistics*, 65, 839–861.

Harris, F. H. D., McInish, T. H., Shoesmith, G. L., and Wood, R. A. (1995), "Cointegration, Error Correction, and Price Discovery on Informationally Linked Security Markets," *Journal of Financial and Quantitative Analysis*, 30, 563–579.

Harris, F. H. D., McInish, T. H., and Wood, R. A. (2002), "Security Price Adjustment Across Exchanges: An Investigation of Common Factor Components for Dow Stocks," *Journal of Financial Markets*, 5, 277–308.

Harris, L. (1990), "Estimation of Stock Variance and Serial Covariance From Discrete Observations," *Journal of Financial and Quantitative Analysis*, 25, 291–306.

——— (1991), "Stock Price Clustering and Discreteness," *Review of Financial Studies*, 4, 389–415.

Hasbrouck, J. (1995), "One Security, Many Markets: Determining the Contributions to Price Discovery," *Journal of Finance*, 50, 1175–1198.

——— (2002), "Stalking the 'Efficient Price' in Market Microstructure Specifications: An Overview," *Journal of Financial Markets*, 5, 329–339.

——— (2004), "Empirical Market Microstructure: Economic and Statistical Perspectives on the Dynamics of Trade in Securities Markets," lecture notes, New York University, Stern School of Business.

Huang, X., and Tauchen, G. (2005), "The Relative Contribution of Jumps to Total Price Variance," *Journal of Financial Econometrics*, 3, 456–499.

Jacod, J. (1994), "Limit of Random Measures Associated With the Increments of a Brownian Semimartingale," unpublished manuscript, Laboratorie de Probabilities, Universite P. and M. Curie, Paris.

Jacod, J., and Protter, P. (1998), "Asymptotic Error Distributions for the Euler Method for Stochastic Differential Equations," *The Annals of Probability*, 26, 267–307.

Jansson, M. (2004), "The Error in Rejection Probability of Simple Autocorrelation Robust Tests," *Econometrica*, 72, 937–946.

Johansen, S. (1988), "Statistical Analysis of Cointegration Vectors," *Journal of Economic Dynamics and Control*, 12, 231–254.

——— (1996), *Likelihood Based Inference in Cointegrated Vector Autoregressive Models* (2nd ed.), Oxford, U.K.: Oxford University Press.

Kiefer, N. M., Vogelsang, T. J., and Bunzel, H. (2000), "Simple Robust Testing of Regression Hypotheses," *Econometrica*, 68, 695–714.

Koopman, S. J., Jungbacker, B., and Hol, E. (2005), "Forecasting Daily Variability of the S&P100 Stock Index Using Historical, Realised and Implied Volatility Measurements," *Journal of Empirical Finance*, 12, 445–475.

Large, J. (2005), "Estimating Quadratic Variation When Quoted Prices Jump by a Constant Increment," Economics Group Working Paper W05, Nuffield College.

Lehmann, B. (2002), "Some Desiderata for the Measurement of Price Discovery Across Markets," *Journal of Financial Markets*, 5, 259–276.

Maheu, J. M., and McCurdy, T. H. (2002), "Nonlinear Features of Realized FX Volatility," *Review of Economics & Statistics*, 84, 668–681.

Meddahi, N. (2002), "A Theoretical Comparison Between Integrated and Realized Volatility," *Journal of Applied Econometrics*, 17, 479–508.

Merton, R. C. (1980), "On Estimating the Expected Return on the Market: An Exploratory Investigation," *Journal of Financial Economics*, 8, 323–361.

Müller, U. A. (1993), "Statistics of Variables Observed Over Overlapping Intervals," discussion paper, O&A Research Group.

Müller, U. A., Dacorogna, M. M., Olsen, R. B., Pictet, O. V., Schwarz, M., and Morgenegg, C. (1990), "Statistical Study of Foreign Exchange Rates, Empirical Evidence of a Price Change Scaling Law, and Intraday Analysis," *Journal of Banking and Finance*, 14, 1189–1208.

Mykland, P. A., and Zhang, L. (2006), "ANOVA for Diffusions and Ito Processes," *The Annals of Statistics*, 34, forthcoming.

Newey, W., and West, K. (1987), "A Simple Positive Semi-Definite, Heteroskedasticity- and Autocorrelation-Consistent Covariance Matrix," *Econometrica*, 55, 703–708.

O'Hara, M. (1995), *Market Microstructure Theory*, London: Blackwell.

Oomen, R. A. C. (2002), "Modelling Realized Variance When Returns Are Serially Correlated," unpublished manuscript, University of Warwick, Warwick Business School.

——— (2005), "Properties of Bias-Corrected Realized Variance Under Alternative Sampling Schemes," *Journal of Financial Econometrics*, 3, 555–577.

——— (2006), "Properties of Realized Variance Under Alternative Sampling Schemes," *Journal of Business & Economic Statistics*, 24, 219–237.

Owens, J. P., and Steigerwald, D. G. (2005), "Noise-Reduced Realized Volatility: A Kalman Filter Approach," unpublished manuscript, Victoria University of Wellington and University of California Santa Barbara.

Patton, A. (2005), "Volatility Forecast Evaluation and Comparison Using Imperfect Volatility Proxies," unpublished manuscript, London School of Economics.

Protter, P. (2005), *Stochastic Integration and Differential Equations*, New York: Springer-Verlag.

Roll, R. (1984), "A Simple Implicit Measure of the Effective Bid–Ask Spread in an Efficient Market," *Journal of Finance*, 39, 1127–1139.

Shiryaev, A. N. (1995), *Probability* (2nd ed.), New York: Springer-Verlag.

Stein, M. L. (1987), "Minimum Norm Quadratic Estimation of Spatial Variograms," *Journal of the American Statistical Association*, 82, 765–772.

Tauchen, G., and Zhou, H. (2004), "Identifying Realized Jumps on Financial Markets," unpublished manuscript, Duke University, Dept. of Economics.

Wasserfallen, W., and Zimmermann, H. (1985), "The Behavior of Intraday Exchange Rates," *Journal of Banking and Finance*, 9, 55–72.

West, K. D. (1997), "Another Heteroskedasticity- and Autocorrelation-Consistent Covariance Matrix Estimator," *Journal of Econometrics*, 76, 171–191.

Zhang, L. (2004), "Efficient Estimation of Stochastic Volatility Using Noisy Observations: A Multi-Scale Approach," research paper, Carnegie Mellon University, Dept. of Statistics.

Zhang, L., Mykland, P. A., and Aït-Sahalia, Y. (2005), "A Tale of Two Time Scales: Determining Integrated Volatility With Noisy High-Frequency Data," *Journal of the American Statistical Association*, 100, 1394–1411.

Zhou, B. (1996), "High-Frequency Data and Volatility in Foreign-Exchange Rates," *Journal of Business & Economic Statistics*, 14, 45–52.

——— (1998), "Parametric and Nonparametric Volatility Measurement, in *Nonlinear Modelling of High-Frequency Financial Time Series*, eds. C. L. Dunis and B. Zhou, New York: Wiley, pp. 109–123.