SCHOOL OF ECONOMICS AND MANAGEMENT
FACULTY OF SOCIAL SCIENCES
AARHUS UNIVERSITY

CREATES
Center for Research in Econometric Analysis of Time Series

CREATES Research Paper 2009-16

Least Squares Inference on Integrated Volatility and the Relationship between Efficient Prices and Noise

Ingmar Nolte and Valeri Voev

School of Economics and Management
Aarhus University
Bartholins Allé 10, Building 1322, DK-8000 Aarhus C
Denmark

Least Squares Inference on Integrated Volatility and the Relationship between Efficient Prices and Noise*

Ingmar Nolte† Valeri Voev‡

Warwick Business School, FERC, CoFE

University of Aarhus CREATES, CoFE

This Version: April 27, 2009

JEL classification: $G10, F31, C32$

Keywords: High frequency data, Subsampling, Realized volatility, Market microstructure

*We would like to thank Peter Hansen, Asger Lunde, Mark Podolskij, Almut Veraart, Kevin Sheppard and Ilze Kalnina for helpful discussions. All remaining errors are ours.

†Warwick Business School, Finance Group, Coventry, CV4 7AL, United Kingdom. Phone +44-24765-72838, Fax -23779, email: Ingmar.Nolte@wbs.ac.uk. The work has been supported in part by the European Community's Human Potential Program under contract HPRN-CT-2002-00232, Microstructure of Financial Markets in Europe; and by the Fritz Thyssen Foundation through the project 'Dealer-Behavior and Price-Dynamics on the Foreign Exchange Market'.

‡School of Economics and Management, University of Aarhus, 8000 Aarhus C, Denmark. Phone +45-8942-1539, email: vvoev@creates.au.dk. Financial support by the Center for Research in Econometric Analysis of Time Series, CREATES, funded by the Danish National Research Foundation, is gratefully acknowledged.

Abstract

The expected value of sums of squared intraday returns (realized variance) gives rise to a least squares regression which adapts itself to the assumptions of the noise process and allows for a joint inference on integrated volatility (IV), noise moments and price-noise relations. In the iid noise case we derive the asymptotic variance of the regression parameter estimating the IV, show that it is consistent and compare its asymptotic efficiency against alternative consistent IV measures. In case of noise which is correlated with the efficient return process, we postulate a new "asymptotically increasing" type of dependence and analyze its ability to cope with the empirically observed price-noise dependence in quote data. In the empirical section of the paper we apply the LS methodology to estimate the integrated volatility as well as the noise properties of 25 liquid stocks both with midquote and transaction price data. We find that while iid noise is an oversimplification, its non-iid characteristics have a decidedly negligible effect on volatility estimation within our framework, for which we provide a sound theoretical reason. In terms of noise-price endogeneity, we are not able to find empirical support for simple ad hoc theoretical models and we provide an alternative explanation for the observed patterns in midquote data, based on market microstructure theory.

1 Introduction

We provide a least-squares (LS) estimation framework for the integrated variance (IV) of stochastic volatility martingale price processes in the presence of general market microstructure (MMS) noise. For the simplest case of iid noise we derive the asymptotic variance of the estimator, show how it can be minimized, and compare it to the variance of other consistent integrated volatility estimators. The analysis shows that the precision of our estimator compares very favorably to that of competing methodologies. While we do not achieve the fastest possible rate of convergence, the minimal variance of the asymptotic distribution is much smaller that that of estimators converging at the fastest rate.[1] In typical empirically-relevant situations our simulations confirm that the LS approach provides estimates which are at least as precise as compared to other consistent estimation techniques.

A further attractive feature of the LS framework is that it simultaneously provides inference on the dependence structure of the MMS noise and sheds light on the impact of non-iid noise on the estimation of IV, for which we find interesting differences between quote and trade data. We also allow for dependence between the noise and the latent returns (endogenous noise) and show that estimation of IV still remains possible, while at the same time we obtain some insights on the nature of the endogeneity.

The estimator we propose can be classified as a subsampling estimator, very much in the spirit of the two-scales realized volatility (TSRV) of Zhang, Mykland & Aït-Sahalia (2005) and the multi-scales realized volatility (MSRV) of Zhang (2006). We also consider the realized kernels (RK) of Barndorff-Nielsen, Hansen, Lunde & Shephard (2008*a*), which have been shown to be related to the multi-scale approach. The comparative attractiveness of our methodology, however, lies not only in its strong statistical properties, but mainly in its flexibility: it adapts naturally to a variety of noise specifications, while remaining easy to handle and estimate. Furthermore, we view the methodology not only as a way of estimating IV, but also as a powerful tool in analyzing MMS phenomena and their impact on stock prices at high frequencies, a research area which has generated a wealth of theoretical and empirical MMS literature and is also of interest in relation to volatility estimation as shown by the works of Hansen & Lunde (2006), Bandi & Russell (2006) and Oomen (2005).

[1] By asymptotic distribution, we mean the distribution properly standardized by a function of the number of observations so that it does not degenerate. Of course, faster converging estimators with a higher asymptotic variance will eventually dominate the slower converging one. Furthermore, comparing asymptotic distributions might be rather misleading for slowly converging estimators in "smallish" samples.

The LS approach is not only flexible in terms of accommodating various noise specifications, but it can also be easily adapted to covariance estimation with non-synchronous observations and MMS noise. To our knowledge, the multivariate extension of the above mentioned estimation techniques has only been considered for the realized kernel, which is provided by Barndorff-Nielsen, Hansen, Lunde & Shephard (2008*b*). We undertake a detailed analysis of LS-based covariance estimation in a subsequent paper. It should be noted, that the possibility of OLS estimation of IV has been addressed in an independent study of Corsi & Curci (2006). Their main focus, however, remains on the discrete sine transform of multi-scale volatility measures and on iid noise specifications.

The paper is structured as follows: in Section 2 we introduce the notation and the theoretical framework and the estimation methodology, Sections 3 and 4 contain our simulation and empirical results and Section 5 concludes. Proofs are collected in the Appendix.

2 Theoretical Setup

Our basic assumption is that we have irregularly spaced, non-synchronous observations of a one-dimensional continuous time process p_t, $t \geq 0$, which is a noisy signal for an underlying process p_t^*:

$$p_t = p_t^* + u_t,$$

where u_t is the noise term. The process p_t^* satisfies the following assumption:

Assumption 1. *The process p_t^* is a stochastic volatility martingale process satisfying*

$$p_t^* = \int_0^t \sigma_u dW_u$$

where σ is a cádlág[2] stochastic process and W is a standard Brownian motion.

The integrated variation process of p^* is given by

$$IV_t = \int_0^t \sigma_u^2 du.$$

[2]The acronym cádlág stands for "continue à droite, limite à gauche". This condition ensures that the integral with respect to W exists.

Our aim is to estimate the increment of integrated variation

$$IV_{(a,b)} = \int_a^b \sigma_u^2 du = IV_b - IV_a.$$

for some predetermined choice of (a, b), e.g., a trading day. Henceforth, we assume that the period of interest is a trading day with $a = 0$ and $b = 1$, and we will omit a and b in the notation.

2.1 IID Noise

With respect to the market microstructure noise process, we start off with the following assumption:

Assumption 2. *The noise process u_t satisfies the following conditions*

(*i*) $p_s^* \perp\!\!\!\perp u_t$, *for all s and t; (Exogeneity)*

(*ii*) $u_s \perp\!\!\!\perp u_t$, *for all $s \neq t$; (Independence)*

(*iii*) $\mathrm{E}[u_t] = 0$, $\omega^2 \equiv \mathrm{E}[u_t^2] < \infty$, *and* $\mu_4 \equiv \mathrm{E}[u_t^4] < \infty$, *for all t.*

While this assumption is highly unrealistic from an empirical point of view, it is a convenient starting point for analyzing our methodology, comparing it to other existing methods for estimation of IV, and establishing an asymptotic theory.
Consider an asset with N observations (ticks, transactions, quote updates) within the period of interest. The grid of observations $\{t_j\}_{j=1,\ldots,N}$ is divided into subgrids $\{t_{js+h}\}_{j=0,\ldots,\lfloor \frac{N-h}{s} \rfloor}$, where $s = 1, \ldots, S$ and $h = 1, \ldots, s$. Here $\{t_{js+h}\}_{j=0,\ldots,\lfloor \frac{N-h}{s} \rfloor}$ denotes the h-th subgrid for a sampling frequency of s ticks (e.g., with $s = 2$ we can have two subgrids, the first one comprising the times $\{t_1, t_3, t_5, \ldots\}$ and the second – the times $\{t_2, t_4, t_6, \ldots\}$). For each subgrid, we can define the corresponding observed and efficient s-tick returns as

$$\begin{aligned} r_{t_{js+h}} &= p_{t_{(j-1)s+h}} - p_{t_{js+h}}, \quad j = 1, \ldots, \left\lfloor \frac{N-h}{s} \right\rfloor \\ r^*_{t_{js+h}} &= p^*_{t_{(j-1)s+h}} - p^*_{t_{js+h}}, \quad j = 1, \ldots, \left\lfloor \frac{N-h}{s} \right\rfloor, \end{aligned}$$

and the noise returns as

$$e_{t_{js+h}} = u_{t_{(j-1)s+h}} - u_{t_{js+h}}, \quad j = 1, \ldots, \left\lfloor \frac{N-h}{s} \right\rfloor.$$

Denote the number of returns for the h-th s-tick subgrid as $N_{h,s} = \left\lfloor \frac{N-h}{s} \right\rfloor - 1$. We define the realized variance (RV) as a function of the number of returns on this subgrid as:

$$RV^{h,s}(N_{h,s}) = \sum_{j=1}^{N_{h,s}} r^2_{t_{js+h}}.$$

Under Assumptions 1 and 2 it holds that

$$\mathrm{E}\left[RV^{h,s}(N_{h,s})\right] \;=\; IV + 2N_{h,s}\omega^2, \tag{1}$$

a result which appears in Hansen & Lunde (2006). On the basis of the theoretical relationship in Equation (1), we can easily derive an OLS regression of the form

$$y_{h,s} = c + \beta_0 N_{h,s} + \varepsilon_{h,s}, \quad s = 1, \ldots, S, \quad h = 1, \ldots, s \tag{2}$$

where $y_{h,s} = RV^{h,s}(N^k_{h,s})$.[3]
It is interesting to note that the above OLS regression has a close relation to the so-called volatility signature plot introduced by Andersen, Bollerslev, Diebold & Labys (2003), which is a graphical description of the effect of sampling frequency on realized volatility. In the above regression, the estimated constant $\hat{c}$ is an estimate of the integrated variance IV, while $\hat{\beta}_0$ is an estimate of $2\omega^2$. Hence, as a by-product of this estimation we obtain the variance of the noise process, which is of interest in is own right.

Theorem 1. *Let $N \to \infty$ and $S = \alpha N^\beta$ for $\alpha > 0$ and $\beta \in [0.5, 1)$. Under Assumptions 1 and 2, the asymptotic variance of $\hat{c}$ is given by*

$$\mathrm{Var}[\hat{c}] = \underbrace{\frac{2(\pi^2 a - 6(\gamma_0^2 + 2\gamma_1)a^*)N}{3S^2(\ln S)^2}}_{\textit{noise term}} + \underbrace{\frac{8IQS}{15N}}_{\textit{discretization term}},$$

where $a = 12\kappa\omega^4$, $a^* = 12\kappa\omega^4 - 4\omega^4$, $IQ = \int_0^1 \sigma_s^4 ds$, γ_0 is the Euler-Mascheroni constant, γ_1 is the first Stieltjes constant, and $\kappa = \mu_4/3\omega^4$. The value of β determines which term dominates the expression.

Proof. See the Appendix. □

We discuss two corollaries of Theorem 1. First we look at the effect of the value of β.

[3]Note that in contrast to the MSRV estimator, we do not average over h for each s.

Corollary 1.

1. *Let $S = \alpha N^{1/2}$. Then the asymptotic variance of $\hat{c}$ is given by*

$$\text{Var}[\hat{c}] = \frac{8(\pi^2 a - 6(\gamma_0^2 + 2\gamma_1)a^*)}{3\alpha^2(\ln(N))^2}.$$

2. *Let $S = \alpha N^{2/3}$. Then the asymptotic variance of $\hat{c}$ is given by*

$$\text{Var}[\hat{c}] = \frac{8IQ\alpha}{15N^{1/3}}.$$

3. *Let $\tilde{N}$ solve $\tilde{N}(\ln(\tilde{N}))^{2/3} = N^{2/3}$ and set $S = \alpha\tilde{N}$. Define p such that $\frac{\tilde{N}}{N} = \frac{1}{N^p}$. Then the asymptotic variance of $\hat{c}$ is given by*

$$\text{Var}[\hat{c}] = \left(\frac{10(\pi^2 a - 6(\gamma_0^2 + 2\gamma_1)a^*) + 8IQ\alpha^3}{15\alpha^2}\right)\frac{1}{N^p}.$$

Note that $1/2 > p > 1/3$.

Choosing $S = \alpha\tilde{N}$ balances the order of the noise and discretization induced terms in the asymptotic variance and achieves the fastest speed of convergence for our estimator which is in this case $N^{p/2}$. In terms of its asymptotic efficiency our approach can be ranked between the two-scale RV (TSRV) of Zhang et al. (2005) which is $\sqrt[6]{N}$-consistent, and the multi-scale RV (MSRV) by Zhang (2006) as well as the realized kernels of Barndorff-Nielsen et al. (2008*a*) which are $\sqrt[4]{N}$-consistent. At the expense of some loss of asymptotic efficiency, the number of subgrids can be chosen proportional to $N^{2/3}$ achieving $N^{1/6}$-consistency. This has the appealing feature that the noise term becomes asymptotically negligible and thus the estimator achieves some robustness to misspecifications in the noise process.

Corollary 2. *Assume a normal distribution for the noise process, so that $\kappa = 1$. Set $S = \alpha\tilde{N}$. The asymptotic variance of $\hat{c}$ is given by*

$$\text{Var}[\hat{c}] = \left(\frac{8\omega^4(\pi^2 - 4(\gamma_0^2 + 2\gamma_1))}{\alpha^2} + \frac{8IQ\alpha}{15}\right)\frac{1}{N^p}.$$

The minimum of the expression in the brackets is attained for $\alpha^* = \sqrt[3]{\frac{30\omega^4(\pi^2 - 4(\gamma_0^2+2\gamma_1))}{IQ}}$ and is equal to $2.16IQ^{2/3}\omega^{4/3}$.

We are now in a position to compare the asymptotic variance of our OLS-IV estimator to other consistent estimators for IV. Comparing to the two-scale RV (TSRV) of Zhang et al. (2005) we have a faster rate of convergence, and interestingly the

same type of constant involving a factor times $IQ^{2/3}\omega^{4/3}$. The smallest factor they obtain is approximately 4.58, as given in the equation which immediately follows their Equation (63). Thus, even after ignoring the faster speed of convergence our estimator has an asymptotic variance which is more than twice lower than the variance of the TSRV. A comparison to the faster converging multi-scale RV (MSRV) by Zhang (2006) and the realized kernels of Barndorff-Nielsen et al. (2008*a*) is somewhat more involved. Barndorff-Nielsen et al. (2008*a*) show that the theoretically smallest asymptotic variance achievable for a $\sqrt[4]{N}$-consistent estimator in this setting is $8\omega IQ^{3/4}N^{-1/2}$ which corresponds to the efficiency of the parametric maximum likelihood estimator. The minimal asymptotic variance of our OLS-IV estimator has been shown to be $2.16\omega^{4/3}IQ^{2/3}N^{-p}$. The factors $8\omega IQ^{3/4}$ and $2.16\omega^{4/3}IQ^{2/3}$ are not directly comparable and depend on the values of ω and IQ. Furthermore, in finite samples, the variance is likely to be affected by smaller order terms which vanish in the asymptotic limit. Therefore, in order to have some meaningful comparison of the estimators we conduct a small Monte Carlo simulation study in section 3 which shows that for any reasonable sample size the OLS-IV estimator compares quite favorably to the other consistent estimation approaches.

2.2 Dependent Noise

We consider two types of dependence: in calendar (or physical) time, and in tick time.

2.2.1 Dependence in Calendar Time

The calendar time dependent noise satisfies the following assumption

Assumption 3. *The noise process u_t satisfies the following*

(*i*) $p_s^* \perp\!\!\!\perp u_t$, *for all s and t;*

(*ii*) $\mathrm{E}[u_t] = 0$ *for all t;*

(*iii*) *The noise process u is covariance stationary with autocovariance function given by $\gamma(q) = \mathrm{E}[u_t u_{t-q}]$.*

Under Assumptions 1 and 3 it holds that

$$\begin{aligned} \mathrm{E}\left[RV^{h,s}(N_{h,s})\right] &= IV + \sum_{j=1}^{N_{h,s}} \mathrm{Var}\left[e_{t_{js+h}}\right] = IV + 2\sum_{q=1}^{\infty} N_{h,s}(q)\left(\gamma(0) - \gamma(q)\right) \\ &\approx IV + 2N_{h,s}\gamma(0) - 2\sum_{q=1}^{Q} N_{h,s}(q)\gamma(q), \end{aligned} \tag{3}$$

where $N_{h,s}(q)$ counts the number of q-time-units (e.g., seconds) returns for the (h,s)-subgrid given by

$$N_{h,s}(q) = \sum_j \mathbb{1}_{\{t_{js+h} - t_{(j-1)s+h} = q\}}.$$

This result is a straightforward extension of the result in Equation (1) and the fact that under Assumption 3, $\text{Var}\left[e_{t_{js+h}}\right] = 2(\gamma(0) - \gamma(q))$, where $q = t_{js+h} - t_{(j-1)s+h}$. The approximation in Equation (3) results from truncating the autocorrelation function at lag Q. This is reasonable, since for a covariance stationary process the autocovariance function tends to zero for large lags. Alternatively, the equation could be made exact if one explicitly assumes $\gamma(q) = 0$ for $q > Q$, for some positive Q. In terms of estimation, Q has to be chosen by the econometrician. On the basis of the theoretical relationship in Equation (3) and the above assumptions, we can derive the corresponding pooled OLS regression

$$y_{h,s} = c + \beta' x_{h,s} + \varepsilon_{h,s}, \quad s = 1, \ldots, S, \quad h = 1, \ldots, s \tag{4}$$

where $y_{h,s} = RV^{h,s}(N_{h,s})$ and $x_{h,s}$ is the Q-dimensional vector given by $x_{h,s} = (N_{h,s}, N_{h,s}(1), \ldots, N_{h,s}(Q))'$. The estimated constant $\hat{c}$ is an estimate of the integrated variance IV, while $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_Q$ are estimates of $2\gamma(0), -2\gamma(1), \ldots, -2\gamma(Q)$.

2.2.2 Dependence in Tick Time

The tick time dependent noise satisfies the following assumption

Assumption 4. *The noise process u_t satisfies the following*

(*i*) $p_s^* \perp\!\!\!\perp u_t$, *for all s and t;*

(*ii*) $\text{E}[u_t] = 0$ *for all t;*

(*iii*) *The noise process u is covariance stationary with autocovariance function given by* $\gamma(q) = \text{E}\left[u_{t_j} u_{t_{j-q}}\right]$.

The assumption of tick time dependence creates some difficulties as now we have under Assumptions 1 and 4 that

$$\text{E}\left[RV^{h,s}(N_{h,s})\right] = IV + \sum_{j=1}^{N_{h,s}} \text{Var}\left[e_{t_{js+h}}\right] = IV + 2N_{h,s}\left(\gamma(0) - \gamma(s)\right),$$

and thus $\gamma(0) - \gamma(s)$ cannot be identified without additional assumptions, either as a sum or separately. A possible identifying assumption is to postulate that $\gamma(q) = 0$ for $q \geq \bar{Q}$ for some $\bar{Q} > 0$. Then for $s \geq \bar{Q}$ we will have that

$$\mathrm{E}\left[RV^{h,s}(N_{h,s})\right] = IV + \sum_{j=1}^{N_{h,s}} \mathrm{Var}\left[e_{t_{js+h}}\right] = IV + 2N_{h,s}\gamma(0), \quad s \geq \bar{Q}. \tag{5}$$

Thus, an OLS regression which can identify IV and $\gamma(0)$ is

$$y_{h,s} = c + \beta_0 N_{h,s} + \varepsilon_{h,s}, \quad s = \bar{Q}, \ldots, S, \quad h = 1, \ldots, s. \tag{6}$$

This essentially is the iid noise framework (since in this case $\gamma(0) = \omega^2$), where we have assumed that the noise in $\bar{Q}$-tick returns can be considered to be iid. Given the evidence in Hansen & Lunde (2006), $\bar{Q}$ can be chosen so that there is approximately one minute between returns. The iid noise theory can then be applied to the resulting $\bar{Q}$-tick returns. As an alternative, if if $\bar{Q}$ is too large compared to an optimally selected S, we recommend to use some sparse sampling (as the approximately 1 minute sampling in Barndorff-Nielsen et al. (2008*a*)) and apply the OLS regression for the iid noise case to the sparse returns.

2.3 Endogenous Noise

In order to introduce dependence between the noise and the price process, we follow an idea in Hansen & Lunde (2006), which we develop further. In Hansen & Lunde (2006) a possible way to generate such dependence is provided by assuming that the noise is given by:

$$u_{t_j} = \phi r^*_{t_j} + \nu_{t_j}, \tag{7}$$

where ν_{t_j} is a sequence of iid random variables with mean zero and variance ω^2. We augment this assumption as follows:

$$u_{t_j} = \phi \Delta_j^{-\delta/2} r^*_{t_j} + \nu_{t_j}, \tag{8}$$

where $\Delta_j = t_j - t_{j-1}$ and $\delta > 0$. The reason for this augmentation is that if Equation (7) holds, then as a direct implication of the result in Hansen & Lunde (2006) we would have that

$$\mathrm{E}\left[RV^{h,s}(N_{h,s})\right] = IV + 2\phi(1+\phi)IV + 2N_{h,s}\omega^2. \tag{9}$$

The resulting bias $2\phi(1+\phi)IV + 2N_{h,s}\omega^2$ would not disappear as the term $2N_{h,s}\omega^2$ becomes negligible as we sample less and less frequently. In the volatility signature

plots in their Figure 1, however, the RV calculated using midquotes appears to eventually settle to some unbiased value as the sampling frequency decreases. Therefore, it seems more plausible that the type of dependence between the noise and the efficient price process is what we will call "asymptotically increasing", i.e., it becomes stronger as Δ_j becomes smaller. This is achieved by adding the term $\Delta_j^{-\delta/2}$ in Equation (8). Then we obtain

$$\mathrm{E}\left[RV^{h,s}(N_{h,s})\right] = IV + 2\phi \sum_{j=1}^{N_{h,s}} \Delta_{j,s}^{-\delta/2} \sigma_{j,s}^2 + 2\phi^2 \sum_{j=1}^{N_{h,s}} \Delta_{j,s}^{-\delta} \sigma_{j,s}^2 + 2N_{h,s}\omega^2 + o_p(1)$$

where $\sigma_{j,s}^2 = \int_{t_{j-s}}^{t_j} \sigma_s^2 ds$ and $\Delta_{j,s} = t_{j-s} - t_j$. In order to continue, we make some assumptions on the regularity of Δ_j, namely that

$$H_N(s) := N^{-\delta/2} \sum_{j:t_j \leq s} \Delta_j^{1-\delta/2} \longrightarrow H(s) \quad \text{(uniformly)}$$

for some differentiable function H, and

$$\sum_{j=1}^{N} \int_{t_{j-1}}^{t_j} |H'(s) - (N\Delta_j)^{\delta/2}| ds = o(N^{-\delta/2}).$$

Then

$$2\phi \sum_{j=1}^{N} \Delta_j^{-\delta/2} \sigma_j^2 = 2\phi N^{\delta/2} \int_0^1 \sigma_s^2 H'(s) ds + o_p(1). \tag{10}$$

Similarly, assume that

$$G_N(s) := N^{-\delta} \sum_{j:t_j \leq s} \Delta_j^{1-\delta} \longrightarrow G(s) \quad \text{(uniformly)}$$

for some differentiable function G, and

$$\sum_{j=1}^{N} \int_{t_{j-1}}^{t_j} |G'(s) - (N\Delta_j)^{\delta}| ds = o(N^{-\delta}).$$

Then we have

$$2\phi^2 \sum_{j=1}^{N} \Delta_j^{-\delta} \sigma_j^2 = 2\phi^2 N^{\delta} \int_0^1 \sigma_s^2 G'(s) ds + o_p(1). \tag{11}$$

Under a regular sampling scheme for which $\Delta_j = 1/N$ (implying here that $G(x) = H(x) = x$) we would obtain as a direct correspondence to Equation (9):[4]

$$\mathrm{E}\left[RV^{h,s}(N_{h,s})\right] \;=\; IV + 2\phi^2 N_{h,s}^{\delta} IV + 2\phi N_{h,s}^{\delta/2} IV + 2N_{h,s}\omega^2 + o_p(1). \tag{12}$$

[4]We are indebted to Mark Podolskij for pointing to us the more general version, as it appears in Equations (10) and (11).

Since the parameters ϕ and δ are unknown and enter nonlinearly, we unfortunately lose the possibility of OLS estimation in this case, but we can still estimate IV, δ and ω^2 by non-linear least squares. If $\phi < 0$, then as N increases and ω^2 is very small (as found in quote data) there will be a range of values for N, for which the volatility signature plot will be decreasing before the dominating noise term eventually kicks in. Furthermore, we now have that as we sample less frequently (N goes to zero), the realized volatility tends to the integrated volatility, which we would expect. In the empirical section of the paper we estimate of equation (12) with trade and quote data. As in the case of exogenous noise, the iid assumption on the ν_{t_j}'s can be relaxed. If ν_{t_j} conforms to Assumption 3, for example, we could simply add $N_{h,s}(q)$ as regressors in the same way as in Equation (3).

3 Simulation Evidence

In order to compare the performance of the OLS-IV estimator to other consistent estimation techniques we run a set of simple Monte Carlo experiments. As alternative estimation techniques, we employ the MSRV of Zhang (2006), the TSRV of Zhang et al. (2005) and the realized kernels (RK) of Barndorff-Nielsen et al. (2008*a*) using the modified Tukey-Hanning$_2$ kernel. We employ an iid noise setup, since in this case we have an asymptotic theory for all estimators and a theoretically founded way of choosing an optimal number of subgrids or kernel length. The notation S is used to denote both the number of subgrids or the number of realized autocovariances (kernel length) in the realized kernel framework. We simulate

$$dp_t^* = \sigma_t dW_t, \tag{13}$$

where W_t is a standard Brownian motion. The volatility follows a GARCH diffusion processes: $d\sigma_t^2 = \theta(\varpi - \sigma_t^2)dt + \sqrt{2\lambda\theta\sigma_t}dW_{(\sigma)t}$, where $W_{(\sigma)t}$ is a Brownian motion independent of W_t. We use $\lambda = 0.296$, $\varpi = 0.636$ and $\theta = 0.035$. The noise is iid normal with mean zero and variance $\omega^2 = 0.001$ and we employ 5000 simulation runs. We run three experiments, which can be summarized as follows:

1. Set $N = 23400$ (corresponding to 6.5 hours of second-by-second data), vary S (the number of subgrids, or kernel length) from 2 to 50;
2. Set $N = 86400$ (corresponding to 24 hours of second-by-second data), vary S (the number of subgrids, or kernel length) from 2 to 100;
3. Let N vary from 1 000 to 100 000 with a step of 1 000, for a total of 100 values. For each of the four estimators choose S in an optimal way, given the

corresponding asymptotic theory.[5]

The results of the first two simulation experiments are summarized in Table 1, while the outcome of the third Monte Carlo experiment is illustrated in Figure 1. In the figure we plot the standard deviations, since root mean squared errors (RMSE's) are almost identical to the standard deviations due to the unbiasedness of all estimators, which is evident in Table 1. All simulation experiments show that the OLS methodology compares quite well with the other approaches. We can also conclude that there is evidence that the OLS-IV estimator outperforms the MSRV and TSRV (at least for sample sizes up to 100 000 observations per day), while it seems to behave similarly to the realized kernels. Generally, this confirms our expectations that with iid noise, the OLS-IV method provides very precise measures of ex-post integrated volatility. It is worth emphasizing, that we have not constructed the estimator in a way to explicitly minimize the RMSE (as is the case with the MSRV and TSRV), but we have rather been motivated by having a flexible method which easily adapts to various noise specification. Thus, the precision we obtain in this simple scenario comes as a nice complement to the flexibility of our estimation framework.

	$N = 23400$					$N = 86400$				
		S^*	$V(S^*)$	$\hat{S}^*$	$V(\hat{S}^*)$		S^*	$V(S^*)$	$\hat{S}^*$	$V(\hat{S}^*)$
St.dev.	OLS-IV	26	4.583	21	4.672	OLS-IV	60	3.421	52	3.456
	MSRV	19	4.835	17	4.882	MSRV	41	3.566	35	3.651
	TSRV	28	5.105	24	5.187	TSRV	71	4.081	63	4.120
	RK	30	4.697	33	4.703	RK	67	3.468	69	3.469
RMSE	OLS-IV	26	4.583	21	4.673	OLS-IV	60	3.422	52	3.457
	MSRV	19	4.889	17	4.932	MSRV	41	3.662	35	3.736
	TSRV	28	5.132	24	5.188	TSRV	71	4.083	63	4.132
	RK	30	4.697	33	4.703	RK	67	3.469	69	3.469

Table 1: Standard deviations (St.dev.) and RMSE's for the OLS-IV, MSRV, TSRV and RK estimators in percent of the true value of IV. S^* denotes the number of subsamples (kernel length) for which the corresponding minimum is achieved across the values of S considered in the simulation, $\hat{S}^*$ denotes the asymptotically optimal number of subsamples (kernel length), and $V(x)$ is the value of the statistic (St.dev. or RMSE) at x.

[5]In the iid noise case, the optimal S is provided explicitly for each estimator. Since it depends on unknown quantities, such as ω^2 and IQ, whose estimation could render the comparison less sharp, we set these quantities to their true values to avoid estimation noise affecting the optimal choice of S.

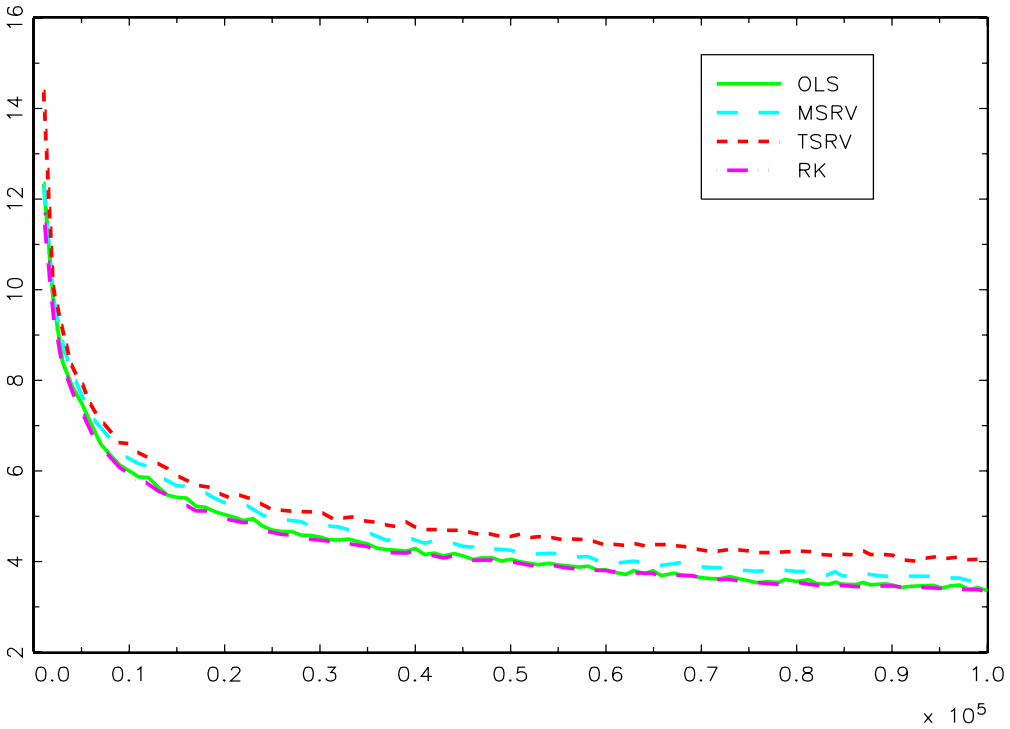

Figure 1: Standard deviations in percent of the true value of IV across different values for the sample size, N, ranging from 1000 to 100 000. For each N, the number of subsamples (kernel length) is chosen optimally according to the corresponding asymptotic theory for each estimator.

4 Empirical Analysis

In this section we apply the LS estimation framework to high-frequency data (trades and quotes) to a set of 25 stocks traded on the NYSE for the period 01.01.2004 to 31.07.2008.[6]

An empirical application of the LS estimation methodology as proposed in this paper requires a choice of Q and S. While we provide an indication of how to choose S in Corollary 2, the choice of Q should be data-driven as it depends on the strength of the serial dependence of the noise process. In order to analyze this dependence we propose a graphical tool very similar to the volatility signature plots mentioned in Section 2, which we name Q-plots. A Q-plot is a plot of the estimate of IV (the

[6]We are grateful to Asger Lunde for providing us with the data and refer the reader to Barndorff-Nielsen, Hansen, Lunde & Shephard (2009) for a description of the dataset and cleaning procedures. The ticker symbols of the stocks in our study are AA, AIG, AXP, BA, BAC, C, CAT, CVX, DD, DIS, GE, GM, HD, IBM, JNJ, JPM, KO, MCD, MMM, MRK, PG, UTX, VZ, WMT and XOM.

intercept) against Q from the regression in Equation 4, arising from the relation:

$$\begin{aligned} \mathrm{E}\left[RV^{h,s}(N_{h,s})\right] &= IV + \sum_{j=1}^{N_{h,s}} \mathrm{Var}\left[e_{t_{js+h}}\right] = IV + 2\sum_{q=1}^{\infty} N_{h,s}(q)\left(\gamma(0) - \gamma(q)\right) \\ &\approx IV + 2N_{h,s}\gamma(0) - 2\sum_{q=1}^{Q} N_{h,s}(q)\gamma(q). \end{aligned}$$

Here we use an arbitrary large enough value of $S = 50$, and we note that while a suboptimally chosen S might increase the variance of the estimator, it cannot lead to biases. The Q-plots are a guide for the value of Q which should be chosen in practice so that the resulting estimate of IV is (at least approximately) unbiased. If Q is chosen too large relative to the strength of the dependence in u, then the estimate will be unbiased but its standard error will be increased (inclusion of irrelevant regressors). If, on the contrary, Q is chosen too small relative to the strength of the dependence in u, then the estimate will be biased (omitted variable bias).

Figures 2 and 3 are collections of Q-plots for the 25 stocks in our study for the trade and quote data, respectively. For comparison, we also compute the realized Parzen kernel as recommended in Barndorff-Nielsen et al. (2009) with data sampled at ticks at approximately Q seconds apart for $Q = 1, 5, 10, 15, 20, 25, 30$ (Barndorff-Nielsen et al. (2009) recommend using all data, which corresponds in our plots to the kernel at $Q = 1$). While there is not a direct correspondence between the kernel with sampling at approximately Q seconds and the OLS-IV estimator with Q noise autocovariances, they are closely related. As a benchmark we also report the standard realized volatility using all data which is plotted for convenience at $Q = -1$.

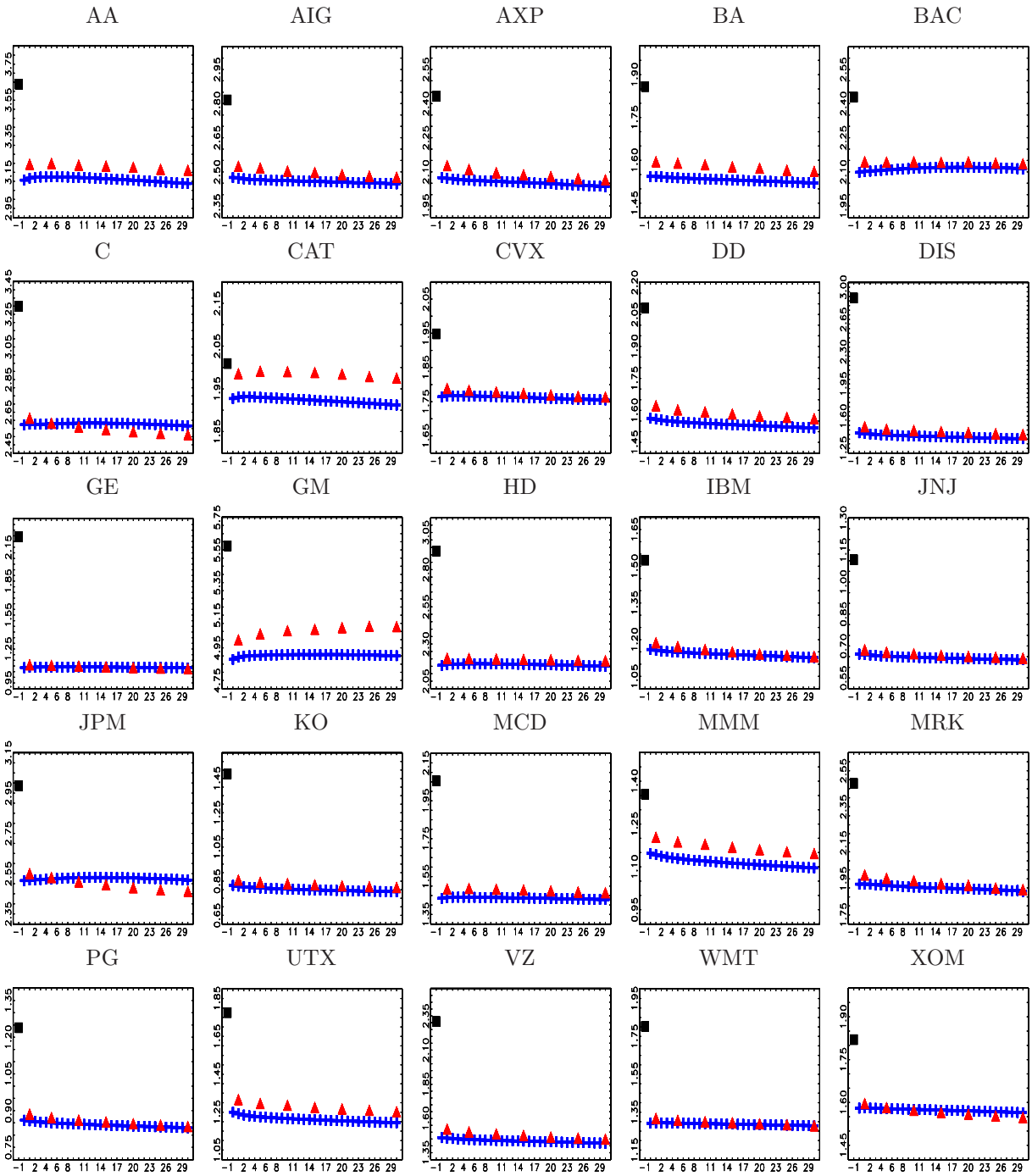

Figure 2: Q-plots with transaction data. Plots of the estimated IV from the regression in Equation 4 against values of Q ranging from 0 to 30 (pluses). The realized volatility estimator using all available data is plotted for comparison at -1 (square). Realized kernels with data sampled at approximately Q seconds using the Parzen kernel are plotted for values of $Q = 1, 5, 10, 15, 20, 25, 30$ (triangles). All plots are averages over the whole sample (1152 days).

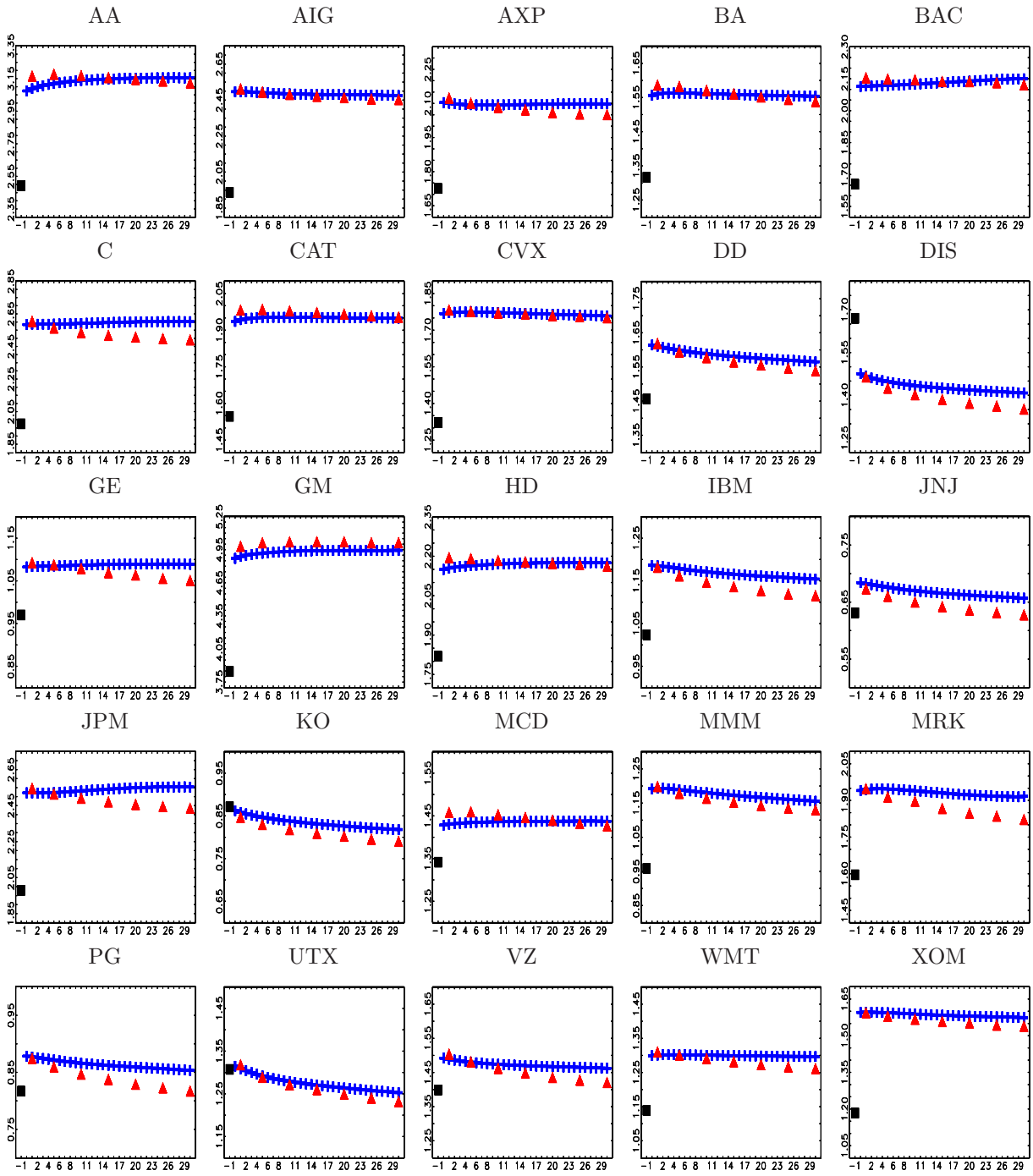

Figure 3: Q-plots with midquote data. Plots of the estimated IV from the regression in Equation 4 against values of Q ranging from 0 to 30 (pluses). The realized volatility estimator using all available data is plotted for comparison at -1 (square). Realized kernels with data sampled at approximately Q seconds using the Parzen kernel are plotted for values of $Q = 1, 5, 10, 15, 20, 25, 30$ (triangles). All plots are averages over the whole sample (1152 days).

The first striking difference between the plots based on transaction data against the plots based on quote data, is that both the OLS-IV estimator and the realized kernel are considerably more stable when applied to transaction data. Furthermore, the RV is always upward biased with trade data, while it is almost always downward biased with midquote data. A clear message from Figure 2 is that $Q = 0$ is a very reasonable choice with trade data, which is also confirmed by the close match with the realized kernel using all data. Choosing a larger value of Q hardly changes the estimate while it most likely increases the variance as discussed above. Thus, while noise in trade data might not be iid, it seems that its iid component is overwhelmingly dominating in terms of biasing the RV at very high frequencies. Fine-tuning the estimator to take account of potential non-iid components of the noise appears redundant.

Figure 3 is more puzzling. Firstly, the simple RV estimator is almost always downward biased, which is completely at odds with the exogenous noise assumption. Note, however, that while this renders the interpretation of the β_q coefficients as noise autocovariances meaningless, it does not suggest that the estimated constant fails as an estimator of IV. In fact, the OLS estimate of IV is more stable across Q than the realized kernel estimate. Barndorff-Nielsen et al. (2009) perform an analysis of the coherence between IV kernel estimates based on trade versus midquote data by regressing them on each other and evaluating the fit. We perform the same analysis for our estimator and compare the results to the kernel estimates. Scatterplots of the data, the OLS fit and the 45° line for the OLS-IV estimator and the realized Parzen kernel are plotted in figures 4 and 5, respectively. Table 2 contains descriptive statistics for the coherence between the midquote-based and trade-based IV estimates for the OLS-IV estimator and the Parzen kernel. As in Barndorff-Nielsen et al. (2009) we find that the realized kernel provides a very good match between both estimates. The OLS-IV estimates are not so well aligned but do agree with each other satisfactorily for all stocks. Generally, our recommendation would be to use trade data, as it seems to deliver clearer results both for the OLS-IV approach and the realized kernels.

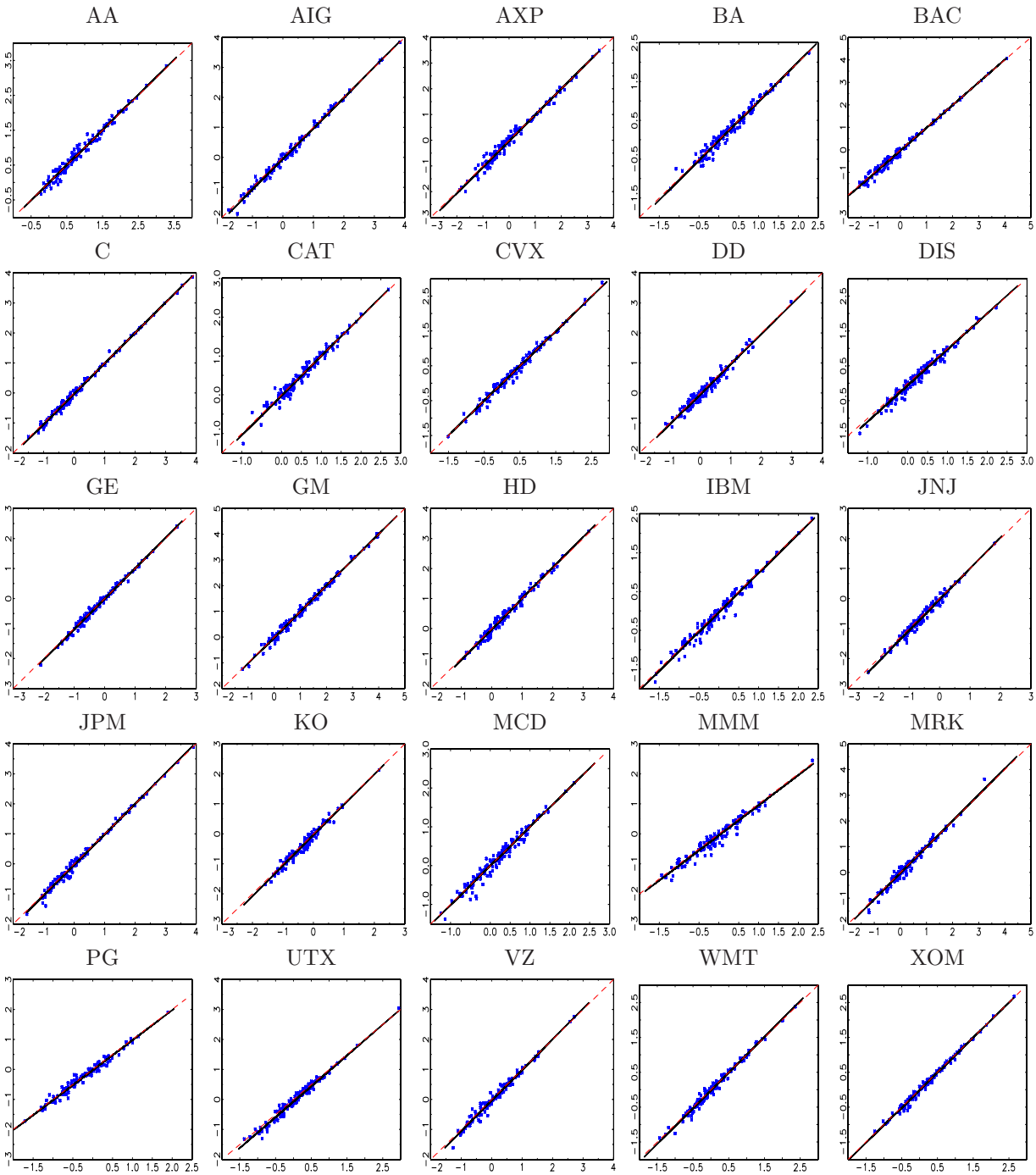

Figure 4: Plots of the logarithm of the OLS-IV estimate using quote data against the one using trade data along with the OLS fit (solid line) and the 45° line (dashed line). For clarity, only every tenth point is plotted.

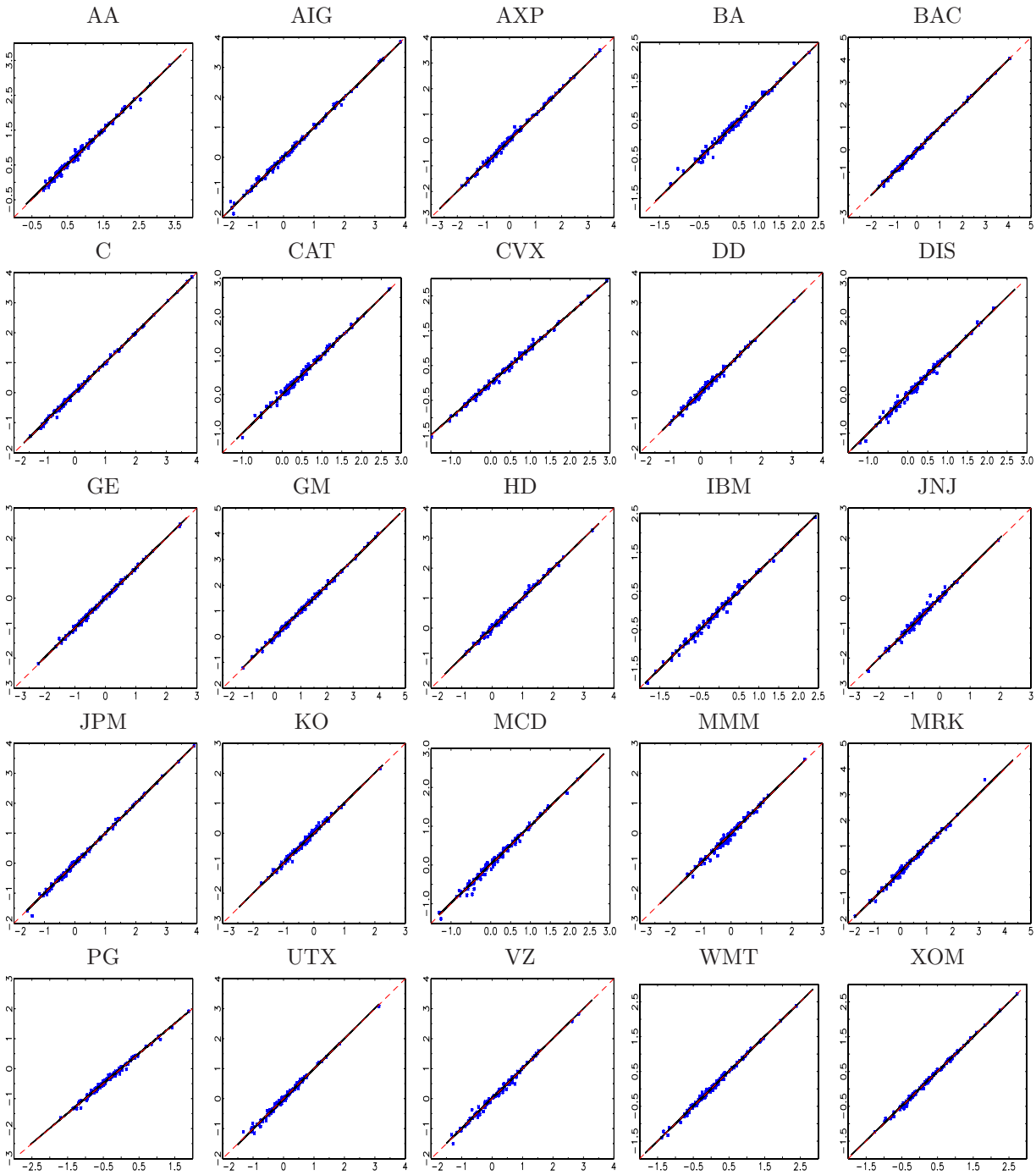

Figure 5: Plots of the logarithm of the Parzen kernel IV estimate using quote data against the one using trade data along with the OLS fit (solid line) and the 45° line (dashed line). For clarity, only every tenth point is plotted.

	OLS-IV				Parzen kernel			
Stock	R^2	Int.	Slope	Dist.	R^2	Int.	Slope	Dist.
AA	0.970	-0.007 (0.006)	1.010 (0.005)	0.170	0.990	0.027 (0.003)	0.987 (0.003)	0.096
AIG	0.990	-0.052 (0.003)	1.022 (0.003)	0.108	0.997	0.002 (0.002)	0.999 (0.002)	0.066
AXP	0.983	-0.038 (0.004)	1.015 (0.004)	0.111	0.987	-0.002 (0.003)	0.999 (0.003)	0.057
BA	0.958	-0.027 (0.004)	1.018 (0.006)	0.095	0.984	0.007 (0.002)	0.994 (0.004)	0.053
BAC	0.988	-0.033 (0.003)	1.012 (0.003)	0.070	0.997	-0.008 (0.002)	1.002 (0.002)	0.041
C	0.994	-0.025 (0.002)	1.012 (0.002)	0.064	0.998	0.004 (0.001)	1.000 (0.001)	0.042
CAT	0.961	-0.026 (0.004)	1.018 (0.006)	0.110	0.990	0.001 (0.002)	1.000 (0.003)	0.056
CVX	0.985	-0.002 (0.003)	0.994 (0.004)	0.074	0.996	0.007 (0.001)	0.996 (0.002)	0.036
DD	0.957	-0.045 (0.004)	1.003 (0.006)	0.107	0.989	0.000 (0.002)	0.998 (0.003)	0.050
DIS	0.962	-0.062 (0.004)	1.031 (0.006)	0.091	0.988	0.003 (0.002)	1.005 (0.003)	0.047
GE	0.987	-0.004 (0.002)	1.017 (0.003)	0.038	0.996	0.010 (0.001)	0.999 (0.002)	0.026
GM	0.987	-0.016 (0.005)	1.006 (0.003)	0.235	0.996	0.015 (0.003)	0.995 (0.002)	0.158
HD	0.980	-0.027 (0.004)	1.025 (0.004)	0.116	0.986	0.006 (0.003)	0.994 (0.004)	0.062
IBM	0.974	-0.034 (0.003)	1.009 (0.005)	0.061	0.991	0.000 (0.002)	1.004 (0.003)	0.034
JNJ	0.968	-0.032 (0.004)	1.036 (0.006)	0.041	0.988	0.010 (0.003)	1.012 (0.003)	0.022
JPM	0.991	-0.028 (0.003)	1.018 (0.003)	0.080	0.997	-0.001 (0.002)	1.002 (0.002)	0.050
KO	0.964	-0.048 (0.004)	1.017 (0.006)	0.051	0.988	0.007 (0.002)	1.003 (0.003)	0.028
MCD	0.956	-0.007 (0.004)	1.007 (0.006)	0.089	0.987	0.016 (0.002)	0.999 (0.003)	0.049
MMM	0.950	-0.051 (0.004)	1.004 (0.007)	0.081	0.985	-0.002 (0.002)	1.005 (0.004)	0.042
MRK	0.974	-0.029 (0.004)	1.017 (0.005)	0.122	0.991	0.003 (0.002)	1.009 (0.003)	0.075
PG	0.960	-0.025 (0.004)	1.000 (0.006)	0.049	0.988	0.009 (0.002)	0.999 (0.003)	0.027
UTX	0.945	-0.077 (0.004)	1.029 (0.007)	0.095	0.983	-0.010 (0.002)	1.009 (0.004)	0.047
VZ	0.972	-0.043 (0.003)	1.024 (0.005)	0.082	0.991	0.007 (0.002)	1.000 (0.003)	0.046
WMT	0.981	-0.019 (0.002)	1.026 (0.004)	0.057	0.994	0.005 (0.001)	1.000 (0.002)	0.030
XOM	0.989	-0.015 (0.002)	1.006 (0.003)	0.050	0.989	0.006 (0.002)	0.992 (0.003)	0.031

Table 2: Descriptive statistics for the coherence of the OLS-IV (left panel) and Parzen kernel (right panel) estimates using midquote and trade data. "R^2" is the R^2 of the a regression of log estimates based on trade data on log estimates based on midquotes. "Int." and "Slope" refer to the intercept and slope coefficients of this regression with standard errors in brackets. "Dist." is the average distance of the fit from the 45° line given by $\frac{1}{D}\sum_{d=1}^{D}\sqrt{(\widehat{IV}_{T,d}-\overline{\widehat{IV}}_d)^2+(\widehat{IV}_{Q,d}-\overline{\widehat{IV}}_d)^2}$, where $\widehat{IV}_{T,d}$ and $\widehat{IV}_{Q,d}$ are the IV estimates using trade and midquote data, respectively, $\overline{\widehat{IV}}_d$ is their average on day d and $D = 1152$ is the total number of days in the sample.

Another interesting quantity of interest in the regressions we suggest is the coefficient β_0 which under the assumption of exogenous noise is an estimator of $2\gamma_0 \equiv 2\omega^2$. If noise is iid, then a plot of β_0 against Q should remain flat, as it will be unbiased for all Q. If, for example, the noise follows an MA($\tilde{Q}$) process, then the estimate of β_0 will be biased in all regressions for which $Q < \tilde{Q}$, due to omission of relevant variables. Figures 6 and 7 are collections of plots of $\hat{\beta}_0/2$ against Q for the 25 stocks in our study for the trade and quote data, respectively.

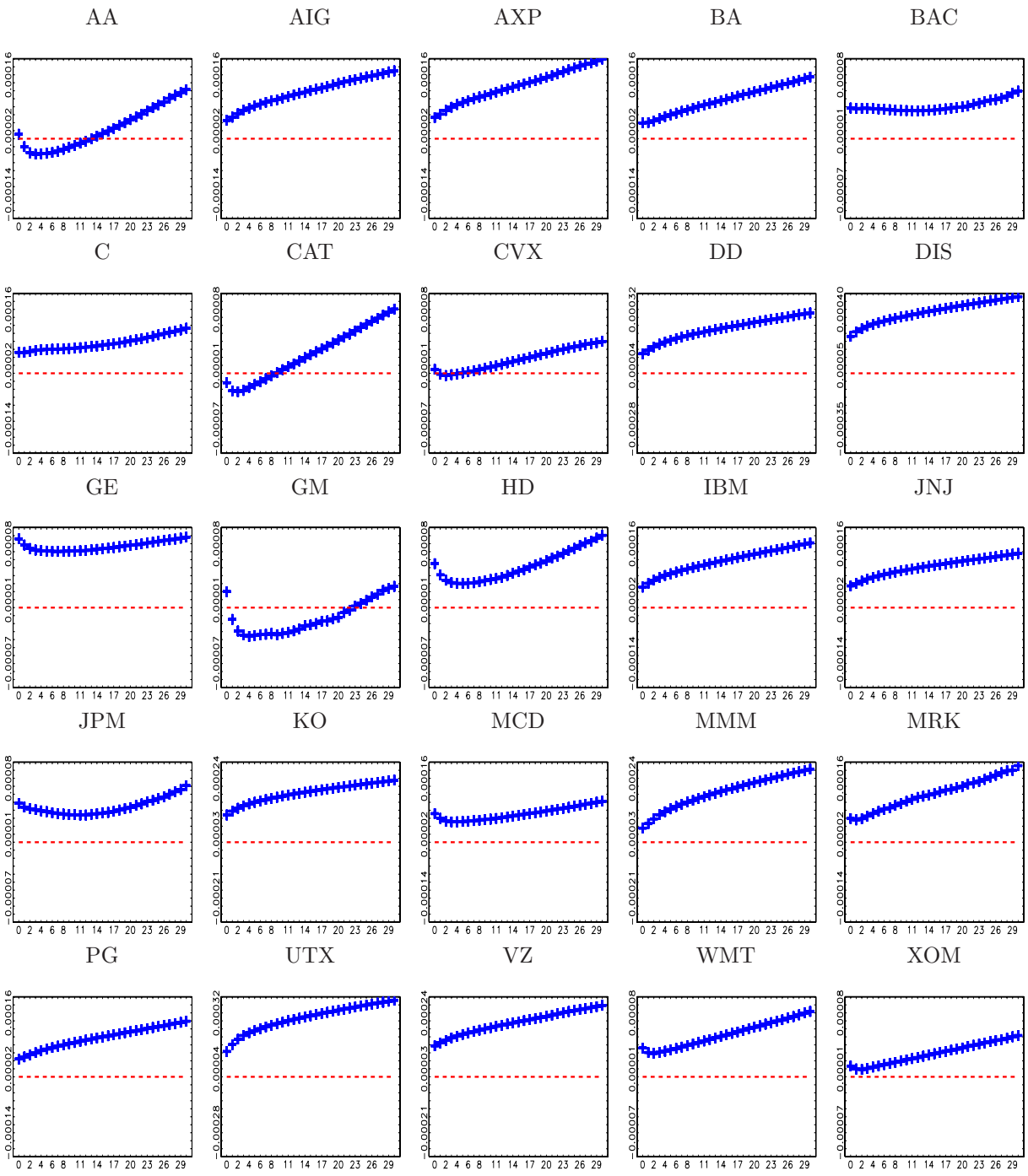

Figure 6: Plots of $\hat{\beta}_0/2$ from the regression in Equation 4 against values of Q ranging from 0 to 30 using transaction data. The dashed line is at zero for each plot. All plots are averages over the whole sample (1152 days).

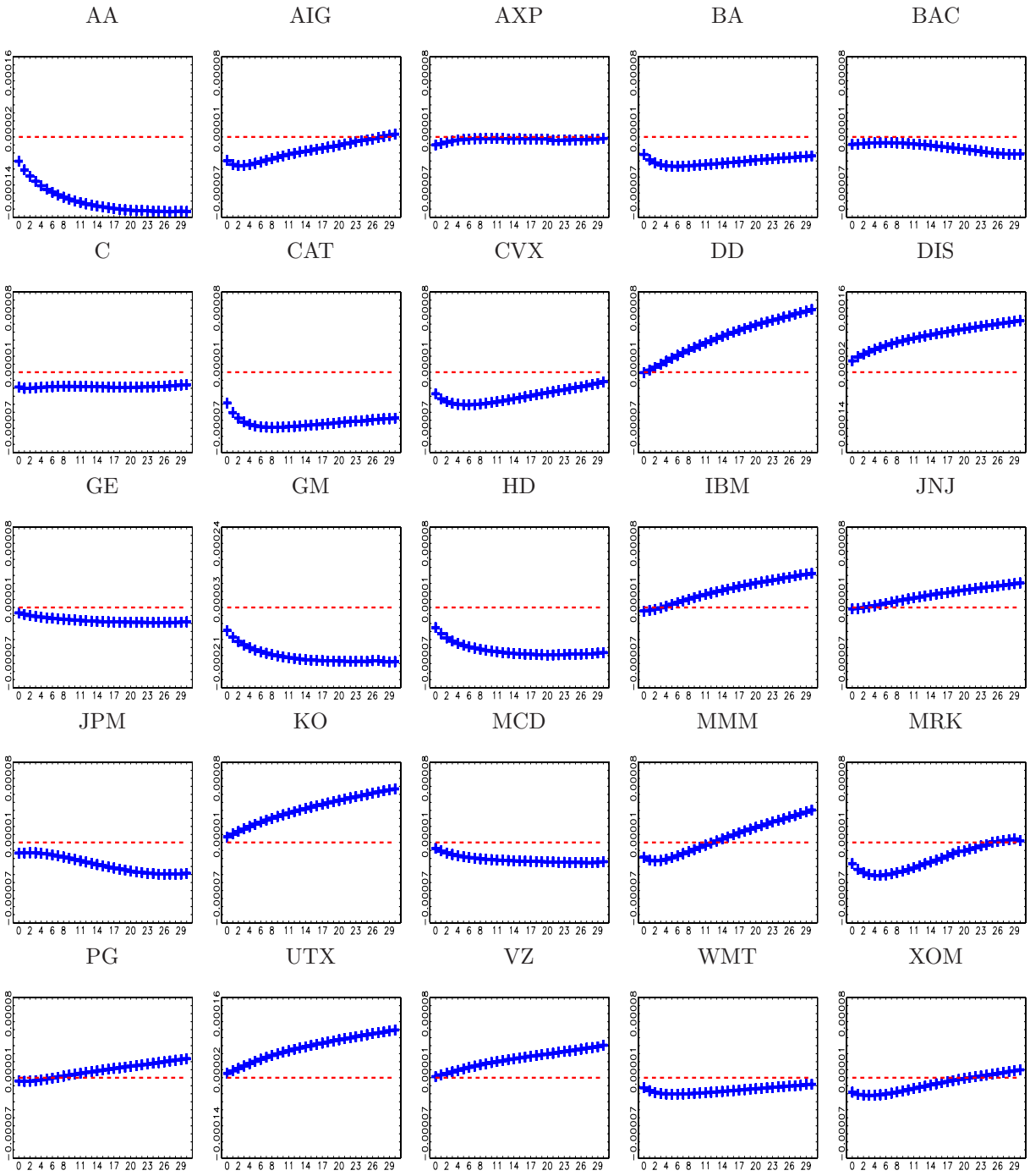

Figure 7: Plots of $\hat{\beta}_0/2$ from the regression in Equation 4 against values of Q ranging from 0 to 30 using midquote data. The dashed line is at zero for each plot. All plots are averages over the whole sample (1152 days).

From the plots it is evident that noise does not seem to be iid, and not even exogenous for the case of midquotes, where the estimates are consistently negative for many of the stocks. This is in fact not surprising given that with midquote data the RV is downward biased as we saw in Figure 3. Given these results we refrain from interpreting the plots using midquote data. The plots with transaction data reveal that the iid assumption is an oversimplification, as the estimate of ω^2 is increasing in Q. In particular for smaller values of Q, $\hat{\beta}_0$ is downward biased. From the theory of omitted variables, we know that the bias is determined by the sign of the correlation of the omitted variable with the included variable and the sign of the coefficient on the omitted variable. Given that the regressor related to β_0, $N_{h,s}$, is positively related to the regressors $N_{h,s}(q)$ (the more observations there are, the more counts there will be for each q), and the omitted variables have coefficients $-2\gamma(q)$, we conclude that $\gamma(q)$ is positive. While this seems to be the general pattern, for some of the stocks we observe a downward slope of the plot for small values of Q, indicating that for $\gamma(q)$ could also be negative for small q.

Comparing the Q-plots for the integrated variance against the Q-plots for ω^2 (Figures 2 and 6), there seems to be a contradiction: IV is largely unaffected by increasing Q beyond 0, while ω^2 is rather sensitive. We provide two explanations to resolve this issue: firstly, we argue that magnitude plays a role. While ω^2 is of the same order of magnitude as the autocovariances $\gamma(q)$, the integrated variance is of much larger magnitude; typically the ratio IV/ω^2 is larger than 10^4. The fact that MMS noise is "small" has been documented in a very comprehensive empirical study by Hansen & Lunde (2006). Interestingly, Barndorff-Nielsen et al. (2008*a*) have a section dedicated to local-to-0 ω^2 asymptotics, where they look at the case $\omega^2 = \omega_0^2 N^{-\alpha}$ for some $0 \leq \alpha < 1$ and a constant ω_0^2. Secondly, which is a more compelling theoretical argument, omitted variable bias can be decomposed as the product of the coefficients of the regression of the omitted variables on the included variables and the coefficients on the omitted variables. Let us consider what happens when we move from $Q = 0$ to $Q = 1$. The biases of $\hat{c}$ and $\hat{\beta}_0$ then depend on the coefficients of the regression of $N_{h,s}(1)$ (the omitted variable) on a constant and $N_{h,s}$ (the included variables). The intercept of this regression will in population be zero, as for $N_{h,s} = 0$, any $N_{h,s}(q)$, $q > 0$ will necessarily be zero as well, while the slope coefficient will be positive, since as mentioned above $N_{h,s}$, is positively related to the regressors $N_{h,s}(q)$. Thus, there is a strong theoretical reason why the estimate of ω^2 is biased, while the estimate of integrated variance, $\hat{c}$, is not. This type of argument naturally carries over for any $Q > 0$.

In order to analyze the endogeneity of MMS noise as described in Section 2.3 we employ non-linear least squares to estimate a regression of the form

$$RV^{h,s}(N_{h,s}) \quad = \quad c + \beta_0 N_{h,s} + \phi N_{h,s}^{\delta} + \varepsilon_{h,s}. \tag{14}$$

which corresponds to Equation (12), where we do not impose the parameter restrictions implied in the equation and include only the leading non-linear term in $N_{h,s}$. The parameters of interest here are in particular ϕ and δ, since these two parameters characterize the relationship between the noise process u and the efficient price p^*. In tables 3 and 4 we present summarized results for the parameter estimates of the above regressions for trade and quote data, respectively. For each coefficient we report the 5%-, median and 95%-quantile across the days in the sample. The 5%- to 95%-quantile range can be seen as a kind of 90% confidence interval under the assumption that the corresponding parameter is constant across days. While for $c \equiv IV$ and $\beta_0 \equiv 2\omega^2$ this does not seem a reasonable assumption, there is substantial evidence that ϕ and δ are quite stable across days and in fact also across stocks. One of the main findings is that the so-constructed confidence intervals for ϕ always contain zero, both for trades and quotes, and for all stocks which is an indication that we do not find endogeneity of the form we have assumed in Equation (8). It should be emphasized that we are not ruling out endogeneity, but rather that we are unable to confirm that it is generated as the model would suggest.

Parameter	c			ω^2			ϕ			δ		
Quantile	0.05	0.5	0.95	0.05	0.5	0.95	0.05	0.5	0.95	0.05	0.5	0.95
AA	0.95	2.50	10.16	-1.96	1.61	7.92	-0.46	-0.02	0.29	0.22	0.38	0.95
AIG	0.28	1.06	9.30	-1.49	0.39	2.59	-0.04	-0.01	0.29	0.22	0.33	0.66
AXP	0.28	1.00	8.46	-2.49	0.47	3.22	-0.04	-0.01	0.38	0.23	0.34	0.81
BA	0.51	1.26	3.99	-1.27	0.46	3.20	-0.04	-0.01	0.04	0.24	0.35	0.66
BAC	0.24	0.81	9.05	-0.33	0.55	2.53	-0.03	-0.01	0.04	0.21	0.32	0.72
C	0.37	1.00	11.30	-0.35	0.53	4.19	-0.03	-0.01	0.26	0.20	0.33	0.98
CAT	0.74	1.71	5.28	-1.92	0.61	4.83	-0.15	-0.02	0.04	0.25	0.40	0.72
CVX	0.55	1.62	5.10	-2.01	0.46	3.78	-0.05	-0.01	0.10	0.26	0.41	0.89
DD	0.49	1.30	4.41	-1.53	0.73	3.65	-0.04	-0.02	0.04	0.23	0.36	0.68
DIS	0.46	1.13	3.40	-0.53	1.29	3.22	-0.04	-0.01	0.04	0.21	0.32	0.61
GE	0.34	0.80	3.29	0.30	1.09	1.78	-0.03	-0.02	0.03	0.22	0.31	0.55
GM	0.80	3.96	20.28	-2.08	2.38	81.78	-1.23	-0.02	0.56	0.19	0.38	0.99
HD	0.62	1.59	7.68	-0.64	1.01	4.16	-0.12	-0.02	0.04	0.21	0.33	0.68
IBM	0.35	0.95	3.12	-0.66	0.32	1.42	-0.03	-0.01	0.03	0.24	0.33	0.61
JNJ	0.16	0.57	1.41	-0.02	0.41	0.81	-0.03	0.00	0.03	0.22	0.31	0.48
JPM	0.29	1.07	10.80	-1.85	0.73	2.22	-0.04	-0.01	0.50	0.21	0.32	0.65
KO	0.21	0.65	1.84	-0.13	0.68	1.17	-0.03	0.01	0.04	0.20	0.29	0.50
MCD	0.44	1.21	3.81	-0.67	0.98	3.75	-0.04	-0.02	0.03	0.22	0.35	0.65
MMM	0.36	1.04	3.04	-1.28	0.38	2.65	-0.04	-0.01	0.03	0.25	0.35	0.66
MRK	0.51	1.44	5.05	-0.75	0.99	4.24	-0.04	-0.02	0.03	0.23	0.35	0.67
PG	0.28	0.71	1.97	-0.52	0.42	1.44	-0.03	-0.01	0.03	0.24	0.33	0.59
UTX	0.35	1.07	2.89	-1.96	0.41	1.69	-0.03	0.01	0.04	0.23	0.34	0.63
VZ	0.33	1.15	4.02	-0.46	0.98	2.55	-0.04	-0.01	0.04	0.22	0.32	0.62
WMT	0.43	1.09	3.58	-0.25	0.65	2.18	-0.03	-0.02	0.03	0.21	0.33	0.62
XOM	0.44	1.42	4.37	-2.10	0.35	3.43	-0.03	-0.01	0.20	0.27	0.43	0.97

Table 3: Estimation results for the parameters in Equation 14 with transaction data. The estimates of ω^2 are scaled by 10^4. For each parameter we report the 5%, median and 95% quantile of the distribution across days for the full sample of 1152 days.

Parameter	c			ω^2			ϕ			δ		
Quantile	0.05	0.5	0.95	0.05	0.5	0.95	0.05	0.5	0.95	0.05	0.5	0.95
AA	0.82	2.22	8.56	-1.26	0.07	1.41	-0.20	-0.01	0.58	0.21	0.37	0.68
AIG	0.33	1.13	9.97	-1.15	0.01	1.14	-0.03	-0.01	0.22	0.22	0.35	0.65
AXP	0.27	0.96	7.99	-1.59	0.01	0.72	-0.02	-0.01	0.66	0.21	0.34	0.62
BA	0.55	1.31	4.01	-0.59	0.03	1.00	-0.02	-0.01	0.02	0.29	0.37	0.61
BAC	0.22	0.76	9.02	-0.69	0.00	0.75	-0.02	0.00	0.02	0.23	0.34	0.67
C	0.35	1.03	12.00	-0.93	0.02	23.91	-0.02	-0.01	0.19	0.22	0.35	0.99
CAT	0.69	1.67	4.94	-1.07	0.06	1.35	-0.09	-0.01	0.22	0.24	0.39	0.63
CVX	0.39	1.54	4.89	-1.10	0.03	1.01	-0.17	-0.01	0.47	0.22	0.38	0.62
DD	0.40	1.27	4.18	-1.05	0.03	0.94	-0.02	-0.01	0.26	0.25	0.36	0.60
DIS	0.45	1.13	3.33	-0.65	-0.01	0.46	-0.02	0.01	0.02	0.24	0.35	0.58
GE	0.28	0.73	3.29	-0.36	0.00	0.13	-0.02	0.00	0.02	0.22	0.34	0.51
GM	0.82	3.52	21.04	-1.99	0.11	40.14	-0.81	-0.01	0.63	0.20	0.37	0.99
HD	0.61	1.59	6.44	-0.92	0.03	0.96	-0.03	-0.01	0.24	0.24	0.36	0.61
IBM	0.39	1.00	3.25	-0.52	0.00	0.58	-0.02	0.00	0.02	0.27	0.35	0.59
JNJ	0.17	0.58	1.38	-0.23	-0.01	0.09	-0.02	0.00	0.02	0.24	0.34	0.47
JPM	0.31	1.07	11.05	-1.25	0.00	0.65	-0.02	0.00	0.43	0.21	0.35	0.64
KO	0.26	0.67	1.93	-0.33	-0.01	0.12	-0.02	0.00	0.02	0.29	0.34	0.48
MCD	0.38	1.11	3.48	-0.69	0.03	0.78	-0.02	0.00	0.02	0.26	0.35	0.57
MMM	0.36	1.06	2.96	-0.68	0.03	0.98	-0.02	-0.01	0.02	0.29	0.37	0.60
MRK	0.53	1.39	4.80	-0.76	0.00	0.83	-0.02	0.00	0.03	0.23	0.35	0.61
PG	0.26	0.71	2.01	-0.51	0.01	0.26	-0.02	0.00	0.02	0.26	0.35	0.54
UTX	0.35	1.11	2.88	-0.94	-0.02	0.77	-0.02	0.00	0.18	0.27	0.36	0.59
VZ	0.34	1.13	4.13	-0.71	0.00	0.50	-0.02	0.00	0.03	0.25	0.35	0.57
WMT	0.45	1.03	3.45	-0.60	0.02	0.33	-0.02	0.00	0.02	0.24	0.35	0.57
XOM	0.17	1.36	4.71	-1.06	0.01	0.95	-0.07	-0.01	0.45	0.23	0.39	0.67

Table 4: Estimation results for the parameters in Equation 14 with midquote data. The estimates of ω^2 are scaled by 10^4. For each parameter we report the 5%, median and 95% quantile of the distribution across days for the full sample of 1152 days.

While we did not expect to find strong endogeneity effects in the trade data, for which the bid-ask bounce dominates as a noise component, it is somewhat surprising to see that midquote data does not exhibit these effects either. We would like to put forward an alternative explanation for the behavior of midquotes, which we believe can be attributed to sluggish adjustment of the either or both of the bid and ask prices. A model in these lines is Lo & MacKinlay (1990) who show that sluggish information assimilation can lead to spurious autocorrelation in observed returns.

On the NYSE there are certain institutional features that can lead to sluggish adjustment of bid and ask quotes. In particular, NYSE Rule 104.10(1) states that the "maintenance of a fair and orderly market implies the maintenance of price continuity with reasonable depth, and the minimizing of the effects of temporary disparity

between supply and demand". Furthermore, NYSE Rule 104.10(2) states that "it is commonly desirable that a member acting as a specialist engage to a reasonable degree under existing circumstances in dealing for his own account when lack of price continuity, lack of depth, or disparity between supply and demand exists or is reasonably to be anticipated".

We illustrate the effect of sluggish quote adjustment on realized volatility by a simple example. Consider a case in which the fundamental price, p^* moves upward by one cent between two ticks. Assume that the ask price adjusts immediately (by exactly the same amount) due to immediate buying pressure, while the bid only adjust some time (say a tick) later. This is rather plausible, as it usually takes some time for the liquidity suppliers (limit order submissions or specialist posting a new quote) to post a competitive quote on the other side of the market. In general we often observe quote updates where only one of the bid or ask prices change. In our example, the move of 1 cent will be realized as two midquote moves of half a cent. Thus the realized variance of the midquote moves will be $2 \cdot 0.5^2 = 0.5$ which is only half of the contribution to the variance of the efficient price. In a more general model where ask and bid prices are allowed to adjust slowly to new information, one can expect that by sampling more and more frequently one would capture more and more of these fragmented midquote moves which will underestimate the variation of the true price process. This can naturally lead to the observed downward sloping volatility signature plots generated with midquote data. We believe that a model of sluggish adjustment of bid and ask prices can potentially be very interesting and informative about the behavior of midquotes but we do not pursue a formal analysis here.

5 Conclusion

In the present paper, we propose a flexible way of estimating the integrated volatility of noisy stochastic volatility martingales and analyze the dependence structure between efficient prices and market microstructure noise. We show that in the case of iid noise, the least-squares approach we propose has good statistical properties and compares very well to other existing consistent estimators of IV.

Besides the simple and straightforward way of implementing our estimator within a Least Squares framework, we argue that the value added of our approach is its flexibility of adapting to various noise dependence structures and its ability to simultaneously estimate second moments of the noise process as well as possible correlation between the process and the efficient return process.

The Monte Carlo studies we carry out confirm the statistical precision of our approach, while its ability to explain empirically relevant price-noise relationships is deferred for later study.

A Appendix: Proofs

A.1 Preliminaries

Under Assumptions 1 and 2, we have that

$$\mathrm{E}\left[RV^{h,s}(N_{h,s})\right] = IV + 2N_{h,s}\omega^2,$$

and hence we have the regression

$$y_{h,s} = c + \beta_0 N_{h,s} + \varepsilon_{h,s}, \quad s = 1, \dots S,\ h = 1, \dots, s,$$

where $y_{h,s} = RV^{h,s}(N_{h,s})$ and the total number of observations in the regression is $N_{tot} = S(S+1)/2$. Set $N_{h,s} = N_s$ as $N_{h,s} \approx \frac{N}{s}$, $s = 1, \dots, S$ up to a rounding error. The above regression can be written in a matrix form as

$$Y = X\theta + \varepsilon,$$

where $\theta = (c, \beta_0)'$. From now on, we condition on the trading times t_j, $j = 1, \dots, N$, which is equivalent to conditioning on the regressor matrix X.
Set $\mathrm{Var}[\varepsilon] = \Xi = \Xi(N, S)$ (we will usually suppress the dependence on N and S). Hansen & Lunde (2006) (Equation 2) show that

$$\mathrm{Var}[y_{h,s}] = \mathrm{Var}[\varepsilon_{h,s}] = 12\kappa\omega^4 N_s + 8\omega^2 \int_0^1 \sigma_s^2 ds - (6\kappa - 2)\omega^4 + \frac{2}{N_s}\int_0^1 \sigma_s^4 ds + o\left(\frac{1}{N_s}\right), \quad (15)$$

which is a diagonal element of Ξ. Denoting the OLS estimator $\hat{\theta} = (\hat{c}, \hat{\beta}_0)'$ we have that

$$\mathrm{Var}[\hat{\theta}] = (X'X)^{-1}X'\Xi X(X'X)^{-1}.$$

Denote by X_1 the first row of $(X'X)^{-1}X'$. Then

$$\mathrm{Var}[\hat{c}] = X_1 \Xi X_1'.$$

A.2 Auxiliary Lemma

Lemma 1. *It holds that*

$$\mathrm{Cov}[RV^{h,s}(N_{h,s}), RV^{h',s'}(N_{h',s'})] = \begin{cases} \frac{2IQ\min(s,r)}{N}, & \text{if } (\star) \\ \frac{2IQ\min(s,r)}{N} + 4\omega^2 \int_{\mathcal{O}} \sigma_s^2 ds + \frac{N\omega^4(12\kappa - 4)}{\mathrm{lcm}(s,r)}, & \text{otherwise} \end{cases},$$

where $(\star) : \{t_{js+h}\}_{j=1,\dots,N_{h,s}} \bigcap \{t_{is'+h'}\}_{i=1,\dots,N_{h',s'}} = \varnothing$ *and the set* $\mathcal{O}$ *is defined in the following proof.* $\mathrm{lcm}(s,r)$ *stands for the least common multiplier of* s *and* r.[7]

[7] It holds that $\max(s,r) \leq \mathrm{lcm}(s,r) \leq sr$. For coprime s and r, $\mathrm{lcm}(s,r) = sr$.

Proof. Write the covariance $\mathrm{Cov}[RV^{h,s}(N_{h,s}), RV^{h',s'}(N_{h',s'})]$ explicitly as

$$\mathrm{Cov}[RV^{h,s}(N_{h,s}), RV^{h',s'}(N_{h',s'})] = \mathrm{Cov}\left[\sum_{j=1}^{N_{h,s}} r^2_{t_{js+h}}, \sum_{i=1}^{N_{h',s'}} r^2_{t_{is'+h'}}\right].$$

This expression can be decomposed as

$$\begin{aligned}
&\mathrm{Cov}\left[\sum_{j=1}^{N_{h,s}} r^2_{t_{js+h}}, \sum_{i=1}^{N_{h',s'}} r^2_{t_{is'+h'}}\right] \\
&= \mathrm{Cov}\left[\sum_{j=1}^{N_{h,s}} r^{*2}_{t_{js+h}}, \sum_{i=1}^{N_{h',s'}} r^{*2}_{t_{is'+h'}}\right] + 2\,\mathrm{Cov}\left[\sum_{j=1}^{N_{h,s}} r^{*2}_{t_{js+h}}, \sum_{i=1}^{N_{h',s'}} r^{*}_{t_{is'+h'}} e_{t_{is'+h'}}\right] + \mathrm{Cov}\left[\sum_{j=1}^{N_{h,s}} r^{*2}_{t_{js+h}}, \sum_{i=1}^{N_{h',s'}} e^2_{t_{is'+h'}}\right] \\
&+ 2\,\mathrm{Cov}\left[\sum_{j=1}^{N_{h,s}} r^{*}_{t_{js+h}} e_{t_{js+h}}, \sum_{i=1}^{N_{h',s'}} r^{*2}_{t_{is'+h'}}\right] + 4\,\mathrm{Cov}\left[\sum_{j=1}^{N_{h,s}} r^{*}_{t_{js+h}} e_{t_{js+h}}, \sum_{i=1}^{N_{h',s'}} r^{*}_{t_{is'+h'}} e_{t_{is'+h'}}\right] + 2\,\mathrm{Cov}\left[\sum_{j=1}^{N_{h,s}} r^{*}_{t_{js+h}} e_{t_{js+h}}, \sum_{i=1}^{N_{h',s'}} e^2_{t_{is'+h'}}\right] \\
&+ \mathrm{Cov}\left[\sum_{j=1}^{N_{h,s}} e^2_{t_{js+h}}, \sum_{i=1}^{N_{h',s'}} r^{*2}_{t_{is'+h'}}\right] + 2\,\mathrm{Cov}\left[\sum_{j=1}^{N_{h,s}} e^2_{t_{js+h}}, \sum_{i=1}^{N_{h',s'}} r^{*}_{t_{is'+h'}} e_{t_{is'+h'}}\right] + \mathrm{Cov}\left[\sum_{j=1}^{N_{h,s}} e^2_{t_{js+h}}, \sum_{i=1}^{N_{h',s'}} e^2_{t_{is'+h'}}\right] \\
&= \mathrm{Cov}\left[\sum_{j=1}^{N_{h,s}} r^{*2}_{t_{js+h}}, \sum_{i=1}^{N_{h',s'}} r^{*2}_{t_{is'+h'}}\right] + 4\,\mathrm{Cov}\left[\sum_{j=1}^{N_{h,s}} r^{*}_{t_{js+h}} e_{t_{js+h}}, \sum_{i=1}^{N_{h',s'}} r^{*}_{t_{is'+h'}} e_{t_{is'+h'}}\right] + \mathrm{Cov}\left[\sum_{j=1}^{N_{h,s}} e^2_{t_{js+h}}, \sum_{i=1}^{N_{h',s'}} e^2_{t_{is'+h'}}\right],
\end{aligned}$$

where the last equation follows because all other terms are zero. Consider the first term, which is the covariance between two estimators for IV. Using Lemma 2.1 in Hausman (1978) it follows that the covariance between them is equal to the variance of the more efficient one, i.e.,

$$\begin{aligned}
\mathrm{Cov}\left[\sum_{j=1}^{N_{h,s}} r^{*2}_{t_{js+h}}, \sum_{i=1}^{N_{h',s'}} r^{*2}_{t_{is'+h'}}\right] &= \begin{cases} \mathrm{Var}\left[\sum_{j=1}^{N_{h,s}} r^{*2}_{t_{js+h}}\right] & \text{, if } N_{h,s} \geq N_{h',s'} \\ \mathrm{Var}\left[\sum_{i=1}^{N_{h',s'}} r^{*2}_{t_{is'+h'}}\right] & \text{, otherwise.} \end{cases} \\
&= \frac{2}{\max(N_{h,s}, N_{h',s'})} \int_0^1 \sigma_s^4 ds = \frac{c^* \min(s,r)}{N}
\end{aligned}$$

The second term vanishes if $\{t_{js+h}\}_{j=1,\ldots,N_{h,s}} \bigcap \{t_{is'+h'}\}_{i=1,\ldots,N_{h',s'}} = \varnothing$, since then the summands are uncorrelated. In the remaining cases we have $\{t_{js+h}\}_{j=1,\ldots,N_{h,s}} \bigcap \{t_{is'+h'}\}_{i=1,\ldots,N_{h',s'}} = \mathcal{A}$, which is a set with $\frac{N}{\mathrm{lcm}(s,r)}$ elements. For $\{\{\{t_{js+h}\} \in \mathcal{A}\} \bigcup \{\{t_{is'+h'}\} \in \mathcal{A}\}\}$, denote $t^* = \max(t_{(j-1)s+h}, t_{(i-1)s'+h'})$ and $t_* = \min(t_{js+h}, t_{is'+h'})$, where the dependence on i, j, s, h, s', h' is deliberately suppressed. Since $e_{t_{js+h}} = u_{t_{(j-1)s+h}} - u_{t_{js+h}}$, we have that for each individual summand in the second term, there are 3 possibilities:

$$\mathrm{Cov}\left[r^*_{t_{js+h}} e_{t_{js+h}}, r^*_{t_{is'+h'}} e_{t_{is'+h'}}\right] = \begin{cases} 0, & \text{if } t_{js+h} \neq t_{is'+h'} \text{ and } t_{(j-1)s+h} \neq t_{(i-1)s'+h'} \\ \omega^2 \int_{t^*}^{t_{js+h}} \sigma_s^2 ds, & \text{if } t_{js+h} = t_{is'+h'} \\ \omega^2 \int_{t_{(j-1)s+h}}^{t_*} \sigma_s^2 ds, & \text{if } t_{(j-1)s+h} = t_{(i-1)s'+h'} \end{cases}.$$

It follows that

$$\begin{aligned}
4\,\mathrm{Cov}\left[\sum_{j=1}^{N_{h,s}} r^*_{t_{js+h}} e_{t_{js+h}}, \sum_{i=1}^{N_{h',s'}} r^*_{t_{is'+h'}} e_{t_{is'+h'}}\right] &= 4\omega^2 \sum_{\mathcal{A}} \left(\int_{t^*}^{t_{js+h}} \sigma_s^2 ds + \int_{t_{(j-1)s+h}}^{t_*} \sigma_s^2 ds \right) \\
&= 4\omega^2 \sum_{\mathcal{A}} \left(\int_{t^*}^{t_*} \sigma_s^2 ds \right) = 4\omega^2 \int_{\mathcal{O}} \sigma_s^2 ds,
\end{aligned}$$

where $\mathcal{O} = \bigcup_{t_{js+h}\in\mathcal{A}, t_{is'+h'}\in\mathcal{A}}[t_*, t^*]$. Since the set $\mathcal{A}$ has $\frac{N}{\mathrm{lcm}(s,r)}$ elements and each of the integrals $\int_{t^*}^{t_*} \sigma_s^2 ds$ is of order $O\left(\frac{1}{\max(N_{h,s}, N_{h',s'})}\right)$ and $\frac{1}{\max(N_{h,s}, N_{h',s'})} = \frac{\min(r,s)}{N}$, it follows that $\int_{\mathcal{O}} \sigma_s^2 ds$ is of order $O\left(\frac{\min(s,r)}{\mathrm{lcm}(s,r)}\right)$.
The third term is also zero whenever $\{t_{js+h}\}_{j=1,\ldots,N_{h,s}} \bigcap \{t_{is'+h'}\}_{i=1,\ldots,N_{h',s'}} = \varnothing$. In the remaining cases we have that for each $j, i : t_{js+h} \in \mathcal{A}, t_{is'+h'} \in \mathcal{A}$ there are four correlated pairs of noise terms, e.g., if $t_{js+h} = t_{is'+h'}$, then the following four pairs are correlated: $e^2_{t_{js+h}}, e^2_{t_{is'+h'}}$; $e^2_{t_{(j-1)s+h}}, e^2_{t_{is'+h'}}$; $e^2_{t_{js+h}}, e^2_{t_{(i-1)s'+h'}}$ and $e^2_{t_{(j-1)s+h}}, e^2_{t_{(i-1)s'+h'}}$. Take, for example, the first pair and consider its covariance:

$$\begin{aligned}
\mathrm{Cov}\left[e^2_{t_{js+h}}, e^2_{t_{is'+h'}}\right] &= \mathrm{E}\left[e^2_{t_{js+h}} e^2_{t_{is'+h'}}\right] - \mathrm{E}\left[e^2_{t_{js+h}}\right]\mathrm{E}\left[e^2_{t_{is'+h'}}\right] \\
&= \mathrm{E}\left[u^2_{t_{js+h}} u^2_{t_{is'+h'}}\right] + \mathrm{E}\left[u^2_{t_{(j-1)s+h}} u^2_{t_{is'+h'}}\right] + \mathrm{E}\left[u^2_{t_{js+h}} u^2_{t_{(i-1)s'+h'}}\right] \\
&\quad + \mathrm{E}\left[u^2_{t_{(j-1)s+h}} u^2_{t_{(i-1)s'+h'}}\right] - \mathrm{E}\left[e^2_{t_{(js+h}}\right]\mathrm{E}\left[e^2_{t_{is'+h'}}\right] \\
&= \mu_4 + 3\omega^4 - 4\omega^4 = \mu_4 - \omega^4 = (3\kappa - 1)\omega^4.
\end{aligned}$$

The remaining three pairs can be similarly shown to have the same covariance. Thus it follows

$$\mathrm{Cov}\left[\sum_{j=1}^{N_{h,s}} e^2_{t_{js+h}}, \sum_{i=1}^{N_{h',s'}} e^2_{t_{is'+h'}}\right] = \frac{N\omega^4(12\kappa - 4)}{\mathrm{lcm}(s,r)}.$$

□

A.3 Proof of Theorem 1

Calculating X_1

We have

$$X'X = \begin{pmatrix} N_{tot} & \sum_{s,h} N_s \\ \sum_{s,h} N_s & \sum_{s,h} N_s^2 \end{pmatrix},$$

and in the following we suppress the double summation indices s, h when unambiguous. Then

$$\det(X'X) = N_{tot} \sum N_s^2 - \left(\sum N_s\right)^2,$$

and

$$(X'X)^{-1} = \frac{1}{\det(X'X)} \begin{pmatrix} \sum N_s^2 & -\sum N_s \\ -\sum N_s & N_{tot} \end{pmatrix}.$$

The first row of $(X'X)^{-1}$ is

$$\left(\frac{\sum N_s^2}{N_{tot}\sum N_s^2 - (\sum N_s)^2} \quad -\frac{\sum N_s}{N_{tot}\sum N_s^2 - (\sum N_s)^2}\right).$$

Set

$$A = \frac{1}{N_{tot}B - C^2}, \text{ with } B = \sum N_s^2 \text{ and } C = \sum N_s.$$

We then have:

$$X_1 = \begin{pmatrix} AB - ACN_1 \\ \\ \left.\begin{matrix} AB - ACN_2 \\ AB - ACN_2 \end{matrix}\right\} 2 \text{ times} \\ \vdots \\ \left.\begin{matrix} AB - ACN_S \\ \vdots \\ AB - ACN_S \end{matrix}\right\} S \text{ times} \end{pmatrix}'.$$

Calculating $\mathrm{Var}[\hat{c}]$

Given the block structure of X_1 and Ξ, we can write

$$X_1 \Xi X_1' = \sum_{s=1}^{S}\sum_{r=1}^{S}\sum_{i=1}^{s}\sum_{j=1}^{r} X_1^{(s)} X_1^{(r)} \xi_{ij}^{(s,r)}.$$

where $\xi_{ij}^{(s,r)}$ is the ij element in the (s,r)-block of Ξ. Let us look at the terms A, B and C. For B we have

$$\lim_{S\to\infty} B = \lim_{S\to\infty} \sum N_s^2 = \lim_{S\to\infty} \sum_{s=1}^{S}\sum_{h=1}^{s} N_s^2 = \lim_{S\to\infty} \sum_{s=1}^{S} sN_s^2 = N^2 \lim_{S\to\infty} \sum_{s=1}^{S} \frac{1}{s} = N^2 \lim_{S\to\infty} (\ln(S) + \gamma_0)$$

with γ_0 the Euler-Mascheroni constant. Similarly, we can derive $C = NS$. It follows that

$$\begin{aligned} \lim_{S\to\infty} A &= \lim_{S\to\infty} \frac{1}{\frac{S(S+1)}{2} N^2(\ln(S) + \gamma_0) - N^2 S^2} \\ &= \lim_{S\to\infty} \frac{2}{N^2(S^2 \ln(S) + S^2(\gamma_0 - 2) + S\ln(S) + S\gamma_0)} \end{aligned}$$

The expression

$$\mathrm{Var}[\hat{c}] = X_1 \Xi X_1' = \sum_{s=1}^{S}\sum_{r=1}^{S}\sum_{i=1}^{s}\sum_{j=1}^{r} X_1^{(s)} X_1^{(r)} \xi_{ij}^{(s,r)}$$

can be decomposed as

$$\begin{aligned} &\sum_{s=1}^{S}\sum_{r=1}^{S}\sum_{i=1}^{s}\sum_{j=1}^{r} X_1^{(s)} X_1^{(r)} \xi_{ij}^{(s,r)} \\ &= \sum_{s=1}^{S}\sum_{i=1}^{s}\sum_{j=1}^{s} \left(X_1^{(s)}\right)^2 \xi_{ij}^{(s,s)} + \sum_{s=1}^{S}\sum_{r\neq s}^{S}\sum_{i=1}^{s}\sum_{j=1}^{r} X_1^{(s)} X_1^{(r)} \xi_{ij}^{(s,r)} \\ &= \sum_{s=1}^{S}\sum_{i=1}^{s} \left(X_1^{(s)}\right)^2 \xi_{ii}^{(s,s)} + \sum_{s=1}^{S}\sum_{i=1}^{s}\sum_{j\neq i}^{s} \left(X_1^{(s)}\right)^2 \xi_{ij}^{(s,s)} + \sum_{s=1}^{S}\sum_{r\neq s}^{S}\sum_{i=1}^{s}\sum_{j=1}^{r} X_1^{(s)} X_1^{(r)} \xi_{ij}^{(s,r)} \end{aligned}$$

The Term $\sum_{s=1}^{S}\sum_{i=1}^{s}\left(X_1^{(s)}\right)^2\xi_{ii}^{(s,s)}$

Since $X_1^{(s)}$ does not depend on i we have

$$\sum_{s=1}^{S}\sum_{i=1}^{s}\left(X_1^{(s)}\right)^2\xi_{ii}^{(s,s)}=\sum_{s=1}^{S}\left(X_1^{(s)}\right)^2\sum_{i=1}^{s}\xi_{ii}^{(s,s)}$$

We have that (ignoring the $o\left(\frac{1}{N_s}\right)$ term)

$$\xi_{ii}^{(s,s)}=\underbrace{aN_s+b}_{\text{noise error}}+\underbrace{\frac{c}{N_s}}_{\text{discretization error}},$$

where by comparing to Equation (15) we see that

$$a=12\kappa\omega^4,\quad b=8\omega^2\int_0^1\sigma_s^2ds-(6\kappa-2)\omega^4,\quad c=2\int_0^1\sigma_s^4ds$$

The inner sum is

$$\sum_{i=1}^{s}\xi_{ii}^{(s,s)}=\sum_{i=1}^{s}\left(a\frac{N}{s}+b+\frac{cs}{N}\right)=aN+bs+\frac{cs^2}{N}.$$

Further

$$\left(X_1^{(s)}\right)^2=\left(AB-AC\frac{N}{s}\right)^2=A^2B^2-2A^2BC\frac{N}{s}+A^2C^2\frac{N^2}{s^2}$$

Finally, we have

$$\sum_{s=1}^{S}\sum_{i=1}^{s}\left(X_1^{(s)}\right)^2\xi_{ii}^{(s,s)}\;=\;\sum_{s=1}^{S}\left(aN+bs+\frac{cs^2}{N}\right)\left(A^2B^2-2A^2BC\frac{N}{s}+A^2C^2\frac{N^2}{s^2}\right)$$

Since

$$\begin{aligned}
\sum_{s=1}^{S} s^2 &= \frac{1}{6}(2S^3+3S^2+S) \\
\sum_{s=1}^{S} s &= \frac{1}{2}(S^2+S) \\
\lim_{S\to\infty}\left(\sum_{s=1}^{S}\frac{1}{s}-\ln(S)\right) &= \gamma_0 \\
\lim_{S\to\infty}\left(\sum_{s=1}^{S}\frac{1}{s^2}-\frac{\pi^2}{6}\right) &= 0 \\
A^2B^2 &\in O\left(\frac{1}{S^4}\right) \\
A^2BC &\in O\left(\frac{1}{NS^3\ln(S)}\right) \\
A^2C^2 &\in O\left(\frac{1}{N^2S^2(\ln(S))^2}\right),
\end{aligned}$$

we obtain that as $S\to\infty$ and $N\to\infty$, $\sum_{s=1}^{S}\sum_{i=1}^{s}\left(X_1^{(s)}\right)^2\xi_{ii}^{(s,s)}$ is dominated by $\frac{2\pi^2 aN}{3(S(\ln(S)+\gamma_0)+(\ln(S)+\gamma_0)-2S)^2}$ which is of order $O\left(\frac{N}{S^2(\ln(S))^2}\right)$.

The Term $\sum_{s=1}^{S}\sum_{i=1}^{s}\sum_{j\neq i}^{s}\left(X_1^{(s)}\right)^2\xi_{ij}^{(s,s)}$

For this term we need the covariance between two realized variances computed at the same sampling frequency (within an (s,s)-block) but with non-overlapping grids. As we are working under an i.i.d. noise framework, this covariance is not affected by the noise. Using the same arguments as Barndorff-Nielsen & Shephard (2002), it follows that this covariance is equal to

$$\xi_{ij}^{(s,s)} = \mathrm{Cov}\left[RV^h(N_s), RV^{h'}(N_s)\right] = \frac{2}{N_s}\int_0^1 \sigma_s^4 ds + o\left(\frac{1}{N_s}\right) = \frac{c}{N_s} + o\left(\frac{1}{N_s}\right), \quad h,h' = s.$$

Then we have

$$\sum_{s=1}^{S}\sum_{i=1}^{s}\sum_{j\neq i}^{s}\left(X_1^{(s)}\right)^2\xi_{ij}^{(s,s)} = \sum_{s=1}^{S}\left(X_1^{(s)}\right)^2\sum_{i=1}^{s}\sum_{j\neq i}^{s}\frac{cs}{N} = \sum_{s=1}^{S}\left(X_1^{(s)}\right)^2\frac{s^2(s-1)c}{N}$$

Substituting in $\left(X_1^{(s)}\right)^2$ yields

$$\begin{aligned}\sum_{s=1}^{S}\left(X_1^{(s)}\right)^2\frac{s^2(s-1)c}{N} &= \sum_{s=1}^{S}\left(A^2B^2-2A^2BC\frac{N}{s}+A^2C^2\frac{N^2}{s^2}\right)\frac{s^2(s-1)c}{N}\\ &= \sum_{s=1}^{S}cA^2B^2\frac{s^2(s-1)}{N}-2cA^2BCs(s-1)+cA^2C^2N(s-1).\end{aligned}$$

This sum is of order $O\left(\frac{1}{N}\right)$ and thus negligible.

The Term $\sum_{s=1}^{S}\sum_{r\neq s}^{S}\sum_{i=1}^{s}\sum_{j=1}^{r}X_1^{(s)}X_1^{(r)}\xi_{ij}^{(s,r)}$

For this term we use Lemma 1. The covariance $\xi_{ij}^{(s,r)}$ is affected by whether the numbers s and r are coprime or not. Consider first the case (I) when s and r are coprime. This implies that the number of common observations in an s-subgrid and r-subgrid is $\frac{N}{sr}$ for all s-subgrids and r-subgrids. From Lemma 1, it follows that in this case the covariance $\xi_{ij}^{(s,r)}$ can be written as

$$\xi_{ij}^{(s,r)}=a^*\frac{N}{sr}+b^*\int_{\mathcal{O}}\sigma_s^2ds+\frac{c^*\min(s,r)}{N},$$

where

$$a^*=12\kappa\omega^4-4\omega^4,\quad b^*=4\omega^2,\quad c^*=2\int_0^1\sigma_s^4ds.$$

There are two main differences to be mentioned with respect to a diagonal element $\xi_{ii}^{(s,s)}$. First a and a^* are slightly different, but this is just a technical result. More subtly, we have a term $4\omega^2\int_{\mathcal{O}}\sigma_s^2ds$ of order $O\left(\frac{1}{\max(s,r)}\right)$ which looks similar to b, but unlike it, is decreasing in s and r.

In the second case (II) s and r are not coprime. In such an (s,r)-block there are two possibilities: (II.1) in $\operatorname{lcm}(s,r)$ out of the sr elements in the block, the number of common points on both subgrids is $\frac{N}{\operatorname{lcm}(s,r)}$, (II.2) in the remaining $sr-\operatorname{lcm}(s,r)$ cases the subgrids do not share observations. In case (II.1) we have

$$\xi_{ij}^{(s,r)}=a^*\frac{N}{\operatorname{lcm}(s,r)}+b^*\int_{\mathcal{O}}\sigma_s^2ds+\frac{c^*\min(s,r)}{N},$$

while in case (II.2) it holds that

$$\xi_{ij}^{(s,r)}=\frac{c^*\min(s,r)}{N}.$$

As in all cases (I, II.1 and II.2), $\xi_{ij}^{(s,r)}$ does not depend on i and j and because for coprime s and r, $\text{lcm}(s,r) = sr$, we can write in general that

$$\begin{aligned}\sum_{i=1}^{s}\sum_{j=1}^{r}\xi_{ij}^{(s,r)} &= \left(a^*\frac{N}{\text{lcm}(s,r)} + b^*\int_{\mathcal{O}}\sigma_s^2 ds + \frac{c^*\min(s,r)}{N}\right)\text{lcm}(s,r) + \frac{c^*\min(s,r)}{N}(sr - \text{lcm}(s,r)) \\ &= a^*N + b^*\int_{\mathcal{O}}\sigma_s^2 ds\,\text{lcm}(s,r) + \frac{c^*sr\min(s,r)}{N} \\ &\approx a^*N + b^*\min(s,r) + \frac{c^*sr\min(s,r)}{N}\end{aligned}$$

where the last approximation is employed for operational reasons in the sense that $\int_{\mathcal{O}}\sigma_s^2 ds$ term is of order $O\left(\frac{\min(s,r)}{\text{lcm}(s,r)}\right)$ (and as we show in the sequel, terms involving b^* are asymptotically negligible). As the matrix Ξ is diagonal we express

$$\sum_{s=1}^{S}\sum_{r\neq s}^{S}X_1^{(s)}X_1^{(r)}\sum_{i=1}^{s}\sum_{j=1}^{r}\xi_{ij}^{(s,r)} = 2\sum_{s=1}^{S}\sum_{r>s}^{S}X_1^{(s)}X_1^{(r)}\sum_{i=1}^{s}\sum_{j=1}^{r}\xi_{ij}^{(s,r)}.$$

Substituting in the above derived equation for $\sum_{i=1}^{s}\sum_{j=1}^{r}\xi_{ij}^{(s,r)}$, $X_1^{(s)}$ and $X_1^{(r)}$ results in

$$\begin{aligned}2\sum_{s=1}^{S}\sum_{r>s}^{S}X_1^{(s)}X_1^{(r)}\sum_{i=1}^{s}\sum_{j=1}^{r}\xi_{ij}^{(s,r)} &= 2\sum_{s=1}^{S}\sum_{r>s}^{S}A^2\left(B - C\frac{N}{s}\right)\left(B - C\frac{N}{r}\right)\left(a^*N + b^*s + \frac{c^*s^2r}{N}\right) \\ &= 2\left(\frac{S(S-1)}{2}a^*A^2B^2N + b^*A^2B^2\sum_{s=1}^{S}\sum_{r>s}^{S}s + \frac{c^*A^2B^2}{N}\sum_{s=1}^{S}\sum_{r>s}^{S}s^2r\right. \\ &- a^*A^2BCN^2\sum_{s=1}^{S}\sum_{r>s}^{S}\left(\frac{1}{r}+\frac{1}{s}\right) - b^*A^2BCN\sum_{s=1}^{S}\sum_{r>s}^{S}\left(1+\frac{s}{r}\right) - c^*A^2BC\sum_{s=1}^{S}\sum_{r>s}^{S}\left(s^2+sr\right) \\ &\left.+a^*A^2C^2N^3\sum_{s=1}^{S}\sum_{r>s}^{S}\frac{1}{rs} + b^*A^2C^2N^2\sum_{s=1}^{S}\sum_{r>s}^{S}\frac{1}{r} + c^*A^2C^2N\sum_{s=1}^{S}\sum_{r>s}^{S}s\right).\end{aligned}$$

We first show that the terms involving b^* are asymptotically negligible. This can be confirmed by considering that $\sum_{s=1}^{S}\sum_{r>s}^{S}s \in O\left(S^3\right)$, $\sum_{s=1}^{S}\sum_{r>s}^{S}\left(1+\frac{s}{r}\right) \in O\left(S^2\right)$ and $\sum_{s=1}^{S}\sum_{r>s}^{S}\frac{1}{r} \in O\left(S\right)$. The term $b^*A^2B^2\sum_{s=1}^{S}\sum_{r>s}^{S}s$ is dominant and of order $O\left(\frac{1}{S}\right)$ and hence asymptotically negligible. Next, we look at limits (we implicitly mean $S\to\infty$ in all equations below) of terms involving a^*. To this end consider the sums

$$\begin{aligned}\sum_{s=1}^{S}\sum_{r>s}^{S}\left(\frac{1}{r}+\frac{1}{s}\right) &= \sum_{s=1}^{S}\left(\sum_{r=1}^{S}\frac{1}{r} - \sum_{r=1}^{s}\frac{1}{r}\right) + \sum_{s=1}^{S}\frac{1}{s}(S-s) \\ &= 2\sum_{s=1}^{S}(\ln(S)+\gamma_0) - \sum_{s=1}^{S}(\ln(s)+\gamma_0) - S = S(\ln(S)+\gamma_0) - 0.5\ln(S) - 0.5\ln(2\pi).\end{aligned}$$

$$\begin{aligned}\sum_{s=1}^{S}\sum_{r>s}^{S}\frac{1}{rs} &= \sum_{s=1}^{S}\frac{1}{s}\left(\sum_{r=1}^{S}\frac{1}{r} - \sum_{r=1}^{s}\frac{1}{r}\right) = \sum_{s=1}^{S}\frac{1}{s}(\ln(S)+\gamma_0) - \sum_{s=1}^{S}\frac{1}{s}(\ln(s)+\gamma_0) \\ &= (\ln(s)+\gamma_0)^2 - 0.5\left(\ln(S)\right)^2 - \gamma_1 - \gamma_0(\ln(S)+\gamma_0) = 0.5\left(\ln(S)\right)^2 + \gamma_0\ln(S) - \gamma_1.\end{aligned}$$

where we have used that $\lim_{S\to\infty}\left(\sum_{s=1}^{S}\ln(s)-\ln\left(\sqrt{2\pi S}\left(\frac{S}{e}\right)^{S}\right)\right)=0$ by Sterling's approximation and $\lim_{S\to\infty}\left(\sum_{s=1}^{S}\frac{\ln(s)}{s}-0.5\left(\ln(S)\right)^2\right)=\gamma_1$, where γ_1 is the first Stieltjes constant equal to approximately -0.0728 (see, e.g., Havil (2003)). Thus we obtain

$$\begin{aligned}\frac{S(S-1)}{2}A^2B^2N &= \frac{1}{2}\frac{4N(S^2-S)(\ln(S)+\gamma_0)^2}{\left(S^2(\ln(S)+\gamma_0)+S(\ln(S)+\gamma_0)-2S^2\right)^2}\\ &= \frac{2N(S^2-S)}{\left(S^2+S-\frac{2S^2}{\ln(S)+\gamma_0}\right)^2}=\frac{2N}{\left(S+1-\frac{2S}{\ln(S)+\gamma_0}\right)^2}+O\left(\frac{N}{S^3}\right).\end{aligned}$$

$$\begin{aligned}-A^2BCN^2\sum_{s=1}^{S}\sum_{r>s}^{S}\left(\frac{1}{r}+\frac{1}{s}\right) &= -\frac{4NS(\ln(S)+\gamma_0)\left(S(\ln(S)+\gamma_0)-0.5\ln(S)-0.5\ln(2\pi)\right)}{\left(S^2(\ln(S)+\gamma_0)+S(\ln(S)+\gamma_0)-2S^2\right)^2}\\ &= -\frac{4N}{\left(S+1-\frac{2S}{\ln(S)+\gamma_0}\right)^2}+O\left(\frac{N}{S^3}\right).\end{aligned}$$

$$\begin{aligned}a^*A^2C^2N^3\sum_{s=1}^{S}\sum_{r>s}^{S}\frac{1}{rs} &= \frac{4NS^2\left(0.5\left(\ln(S)\right)^2+\gamma_0\ln(S)-\gamma_1\right)}{\left(S^2(\ln(S)+\gamma_0)+S(\ln(S)+\gamma_0)-2S^2\right)^2}\\ &= \frac{NS^2\left(2\left(\ln(S)+\gamma_0\right)^2-2\gamma_0^2-4\gamma_1\right)}{\left(S^2(\ln(S)+\gamma_0)+S(\ln(S)+\gamma_0)-2S^2\right)^2}\\ &= \frac{2N}{\left(S+1-\frac{2S}{\ln(S)+\gamma_0}\right)^2}-\frac{N(2\gamma_0^2+4\gamma_1)}{\left(S(\ln(S)+\gamma_0)+(\ln(S)+\gamma_0)-2S\right)^2}.\end{aligned}$$

Summing up the three terms we obtain $-\frac{(2\gamma_0^2+4\gamma_1)N}{(S(\ln(S)+\gamma_0)+(\ln(S)+\gamma_0)-2S)^2}+O\left(\frac{N}{S^3}\right)$. It remains to calculate the terms with c^*. We have $\sum_{s=1}^{S}\sum_{r>s}^{S}s^2r=1/15S^5+1/24S^4-1/12S^3-1/24S^2+1/60S$, $\sum_{s=1}^{S}\sum_{r>s}^{S}\left(s^2+sr\right)=5/24S^4+1/12S^3-5/24S^2-1/12S$, and $\sum_{s=1}^{S}\sum_{r>s}^{S}s=1/6S^3-1/6S$. Considering the order of the terms A^2B^2, A^2BC and A^2C^2, the leading term turns out to be

$$\frac{c^*A^2B^2}{N}\sum_{s=1}^{S}\sum_{r>s}^{S}s^2r=\frac{4Sc^*}{15N\left(1-\frac{2}{\ln(S)+\gamma_0}+\frac{1}{S}\right)^2}+O\left(\frac{1}{N}\right).$$

Final Result

Let $N\to\infty$ and $S=\alpha N^{\beta}$ for $\alpha>0$ and $\beta\in[0.5,1)$. Summing everything up together results in

$$\mathrm{Var}[\hat{c}]=\frac{2(\pi^2a-6(\gamma_0^2+2\gamma_1)a^*)N}{3\left(S(\ln(S)+\gamma_0)+(\ln(S)+\gamma_0)-2S\right)^2}+\frac{4Sc^*}{15N\left(1-\frac{2}{\ln(S)+\gamma_0}+\frac{1}{S}\right)^2}+O\left(N^{-1/2}\right)$$

Recalling that $a = 12\kappa\omega^4$, $a^* = 12\kappa\omega^4 - 4\omega^4$, $c^* = 2\int_0^1 \sigma_s^4 ds$, denoting $IQ = \int_0^1 \sigma_s^4 ds$ and setting $\kappa = 1$ (normal noise) we can rewrite the above equation as

$$\mathrm{Var}[\hat{c}] = \frac{8(\pi^2 - (4\gamma_0^2 + 8\gamma_1))\omega^4 N}{\left(S(\ln(S) + \gamma_0) + (\ln(S) + \gamma_0) - 2S\right)^2} + \frac{8SIQ}{15N\left(1 - \frac{2}{\ln(S)+\gamma_0} + \frac{1}{S}\right)^2} + O\left(N^{-1/2}\right).$$

References

Andersen, T. G., Bollerslev, T., Diebold, F. X. & Labys, P. (2003), 'Modeling and forecasting realized volatility', *Econometrica* **71**, 579–625.

Bandi, F. M. & Russell, J. R. (2006), 'Separating microstructure noise from volatility', *Journal of Financial Economics* **79**(3), 655–692.

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

Barndorff-Nielsen, O. E., Hansen, P., Lunde, A. & Shephard, N. (2008*b*), Multivariate realised kernels: consistent positive semi-definite estimators of the covariation of equity prices with noise and non-synchronous trading. CREATES Working Paper 2008-63.

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

Barndorff-Nielsen, O. E. & Shephard, N. (2002), 'Econometric analysis of realized volatility and its use in estimating stochastic volatility models', *Journal of the Royal Statistical Society Series B* **64**(2), 253–280.

Corsi, F. & Curci, G. (2006), Discrete sine transform for multi-scales realized volatility measures. Working paper, University of Lugano.

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

Hausman, J. A. (1978), 'Specification test in econometrics', *Econometrica* **46**, 1251–1272.

Havil, J. (2003), *Exploring Euler's Constant*, Princeton University Press, New Jersey.

Lo, A. & MacKinlay, A. C. (1990), 'An econometric analysis of nonsynchronous trading', *Journal of Econometrics* **45**, 181–212.

Oomen, R. C. A. (2005), 'Properties of bias-corrected realized variance under alternative sampling schemes', *Journal of Financial Econometrics* **3**, 555–577.

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

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

Research Papers 2009

2009-02: Morten Ørregaard Nielsen: Nonparametric Cointegration Analysis of Fractional Systems With Unknown Integration Orders

2009-03: Andrés González, Kirstin Hubrich and Timo Teräsvirta: Forecasting inflation with gradual regime shifts and exogenous information

2009-4: Theis Lange: First and second order non-linear cointegration models

2009-5: Tim Bollerslev, Natalia Sizova and George Tauchen: Volatility in Equilibrium: Asymmetries and Dynamic Dependencies

2009-6: Anders Tolver Jensen and Theis Lange: On IGARCH and convergence of the QMLE for misspecified GARCH models

2009-7: Jeroen V.K. Rombouts and Lars Stentoft: Bayesian Option Pricing Using Mixed Normal Heteroskedasticity Models

2009-8: Torben B. Rasmussen: Jump Testing and the Speed of Market Adjustment

2009-9: Dennis Kristensen and Andrew Ang: Testing Conditional Factor Models

2009-10: José Fajardo and Ernesto Mordecki: Skewness Premium with Lévy Processes

2009-11: Lasse Bork: Estimating US Monetary Policy Shocks Using a Factor-Augmented Vector Autoregression: An EM Algorithm Approach

2009-12: Konstantinos Fokianos, Anders Rahbek and Dag Tjøstheim: Poisson Autoregression

2009-13: Peter Reinhard Hansen and Guillaume Horel: Quadratic Variation by Markov Chains

2009-14: Dennis Kristensen and Antonio Mele: Adding and Subtracting Black-Scholes: A New Approach to Approximating Derivative Prices in Continuous Time Models

2009-15: Charlotte Christiansen, Angelo Ranaldo and Paul Söderlind: The Time-Varying Systematic Risk of Carry Trade Strategies

2009-16: Ingmar Nolte and Valeri Voev: Least Squares Inference on Integrated Volatility and the Relationship between Efficient Prices and Noise