Realised Kernels in Practice: Trades and Quotes

OLE E. BARNDORFF-NIELSEN[a], PETER REINHARD HANSEN[b], ASGER LUNDE[c] AND NEIL SHEPHARD[d]

[a] *The T.N. Thiele Centre for Mathematics in Natural Science, Department of Mathematical Sciences, University of Aarhus, Ny Munkegade, DK-8000 Aarhus C, Denmark & CREATES, University of Aarhus*
E-mail: oebn@imf.au.dk

[b] *Department of Economics, Stanford University, Landau Economics Building, 579 Serra Mall, Stanford, CA 94305-6072, USA*
E-mail: peter.hansen@stanford.edu

[c] *Department of Marketing and Statistics, Aarhus School of Business, University of Aarhus, Fuglesangs Alle 4, DK-8210 Aarhus V, Denmark & CREATES, University of Aarhus*
E-mail: alunde@asb.dk

[d] *Oxford-Man Institute, University of Oxford, Blue Boar Court, 9 Alfred Street, Oxford OX1 4EH, UK & Department of Economics, University of Oxford*
E-mail: neil.shephard@economics.ox.ac.uk

Received: May 2008

Summary Realised kernels use high frequency data to estimate daily volatility of individual stock prices. They can be applied to either trade or quote data. Here we provide the details of how we suggest implementing them in practice. We compare the estimates based on trade and quote data for the same stock and find a remarkable level of agreement.

We identify some features of the high frequency data which are challenging for realised kernels. They are when there are local trends in the data, over periods of around 10 minutes, where the prices and quotes are driven up or down. These can be associated with high volumes. One explanation for this is that they are due to non-trivial liquidity effects.

Keywords: *HAC estimator, Long run variance estimator; Market frictions; Quadratic variation; Realised variance.*

JEL classification: *C01, C14, C32*

1. INTRODUCTION

The class of realised kernel estimators, introduced by Barndorff-Nielsen, Hansen, Lunde, and Shephard (2008a), can be used to estimate the quadratic variation of an underlying efficient price process from high frequency noisy data. This method, together with alternative techniques such as subsampling and pre-averaging[1], extends the influential realised variance literature which has recently been shown to significantly improve our understanding of time-varying volatility and our ability to predict future volatility — see Andersen, Bollerslev, Diebold, and Labys (2001) and Barndorff-Nielsen and Shephard (2002) and the reviews of that literature by, for example, Andersen, Bollerslev, and Diebold (2008) and Barndorff-Nielsen and Shephard (2007). In this paper we detail the implementation of our recommended realised kernel estimator in practice, focusing on end effects, bandwidth selection and data cleaning across different types of financial databases.

We place emphasis on methods which deliver similar estimates of volatility when applied to either quote data or trade data. This is difficult as they have very different microstructure properties. We show realised kernels perform well on this test. We identify a feature of some datasets which causes these methods difficulties — gradual jumps. These are rare in financial markets, they are when prices exhibit strong linear trends for periods of quite a few minutes. We discuss this issue at some length.

In order to focus on the core issue we represent the period over which we wish to measure the variation of asset prices as the single interval $[0, T]$. We consider the case where Y is a Brownian semimartingale plus jump process ($\mathcal{BMSJ}$) given from

$$Y_t = \int_0^t a_u \mathrm{d}u + \int_0^t \sigma_u \mathrm{d}W_u + J_t, \tag{1}$$

where $J_t = \sum_{i=1}^{N_t} C_i$ is a finite activity jump process (meaning it has a finite number of jumps in any bounded interval of time). So N_t counts the number of jumps that have occurred in the interval $[0, t]$ and $N_t < \infty$ for any t. We assume that a is a predictable locally bounded drift, σ is a càdlàg volatility process and W is a Brownian motion, all adapted to some filtration $\mathcal{F}$. For reviews of the econometrics of processes of the type Y see, for example, Shephard (2005).

Our object of interest is the quadratic variation of Y,

$$[Y] = \int_0^T \sigma_u^2 \mathrm{d}u + \sum_{i=1}^{N_T} C_i^2,$$

where $\int_0^T \sigma_u^2 \mathrm{d}u$ is the integrated variance. We estimate it from the observations

$$X_{\tau_0}, ..., X_{\tau_n}, \quad 0 = \tau_0 < \tau_1 < ... < \tau_n = T,$$

where X_{τ_j} is a noisy observation of Y_{τ_j},

$$X_{\tau_j} = Y_{\tau_j} + U_{\tau_j}.$$

We initially think of U as noise and assume $\mathrm{E}(U_{\tau_j}) = 0$, $\mathrm{Var}\left(U_{\tau_j}\right) = \omega^2$. It can be due to, for example, liquidity effects, bid/ask bounce and misrecording. Specific models for U have

[1] Leading references on this include Zhang, Mykland, and Aït-Sahalia (2005), Zhang (2006) and Jacod, Li, Mykland, Podolskij, and Vetter (2007).

been suggested in this context by, for example, Zhou (1996), Hansen and Lunde (2006), Li and Mykland (2007) and Diebold and Strasser (2007). We will write $U \in \mathcal{WN}$ to denote the case where $(U_{\tau_0}, ..., U_{\tau_n})$ are mutually independent and jointly independent of Y.

There has been substantial recent interest in learning about the integrated variance and the quadratic variation in the presence of noise. Leading references include Zhou (1996), Andersen, Bollerslev, Diebold, and Labys (2000), Zhang, Mykland, and Aït-Sahalia (2005), Andersen, Bollerslev, and Meddahi (2006), Zhang (2006), Bandi and Russell (2008), Hansen and Lunde (2006), Kalnina and Linton (2006), Jacod, Li, Mykland, Podolskij, and Vetter (2007), Barndorff-Nielsen, Hansen, Lunde, and Shephard (2006) and Barndorff-Nielsen, Hansen, Lunde, and Shephard (2008a).

Our recommended way of carrying out estimation based on realised kernels is spelt out in Barndorff-Nielsen, Hansen, Lunde, and Shephard (2008b). Their non-negative estimator takes on the following form

$$K(X) = \sum_{h=-H}^{H} k\left(\tfrac{h}{H+1}\right)\gamma_h, \quad \gamma_h = \sum_{j=|h|+1}^{n} x_j x_{j-h}, \tag{2}$$

where $k(x)$ is a kernel weight function. We focus on the Parzen kernel, because it satisfies the smoothness conditions, $k'(0) = k'(1) = 0$, and is guaranteed to produce a non-negative estimate.[2] The Parzen kernel function is given by

$$k(x) = \begin{cases} 1 - 6x^2 + 6x^3 & 0 \leq x \leq 1/2 \\ 2(1-x)^3 & 1/2 \leq x \leq 1 \\ 0 & x > 1. \end{cases}$$

Here x_j is the j-th high frequency return calculated over the interval τ_{j-1} to τ_j in a way which is detailed in Section 2.2. The method by which these returns are calculated is not trivial, for the accuracy and depth of data cleaning is important, as are the influence of end conditions.

This realised kernel has broadly the same form as a standard heteroskedasticity and autocorrelated (HAC) covariance matrix estimator familiar in econometrics (e.g. Andrews (1991)), but unlike them the statistics are not normalised by the sample size. This makes their analysis more subtle and the influence of end effects important.

Barndorff-Nielsen, Hansen, Lunde, and Shephard (2008b) show that as $n \to \infty$ if $K(U) \xrightarrow{p} 0$ and $K(Y) \xrightarrow{p} [Y]$ then

$$K(X) \xrightarrow{p} [Y] = \int_0^T \sigma_u^2 \mathrm{d}u + \sum_{i=1}^{N_T} C_i^2.$$

The dependence between U and Y is asymptotically irrelevant. They need H to increase with n in order to eliminate the noise in such a way that

$$K(U) \xrightarrow{p} 0.$$

[2]The more famous Bartlett kernel has $k(x) = 1 - |x|$, for $|x| \leq 1$. This kernel is used in the Newey and West (1987) estimator. An estimator based on this kernel function will have a slower rate of convergence because the Bartlett kernel does not satisfy the smoothness conditions.

With $H \propto n^{\eta}$ we will need $\eta > 1/3$ to eliminate the variance and $\eta > 1/2$ to eliminate the bias of $K(U)$, when $U \in \mathcal{WN}$.[3] For $K(Y) \xrightarrow{p} [Y]$ we simply need $\eta < 1$. Barndorff-Nielsen, Hansen, Lunde, and Shephard (2008b) show that $H \propto n^{3/5}$ is the best trade-off between asymptotic bias and variance[4].

Their preferred choice of bandwidth is

$$H^* = c^* \xi^{4/5} n^{3/5}, \quad \text{with } c^* = \left\{ \frac{k''(0)^2}{k_{\bullet}^{0,0}} \right\}^{1/5} \quad \text{and} \quad \xi^2 = \frac{\omega^2}{\sqrt{T \int_0^T \sigma_u^4 \mathrm{d}u}}, \tag{3}$$

where $c^* = ((12)^2/0.269)^{1/5} = 3.5134$ for the Parzen kernel. The bandwidth H^* depends on the unknown quantities ω^2 and $\int_0^T \sigma_u^4 \mathrm{d}u$, where the latter is called the integrated quarticity. In the next section we define an estimator of ξ, which leads to a bandwidth, $\hat{H}^* = c^* \hat{\xi}^{4/5} n^{3/5}$, that can be implemented in practice.

Although the assumption that $U \in \mathcal{WN}$ is a strong one, it is not needed for consistency. Previously $K(U) \xrightarrow{p} 0$ has been shown under quite wide conditions, allowing, for example, the U to be a weakly dependent covariance stationary process.

The structure of the paper is as follows. In Section 2 we discuss the selection of the bandwidth H and the important role of end effects for these statistics. This is followed by Section 3 which is on the data we used in our analysis and the data cleaning we employed. We then look at our data analysis in Section 4, suggesting there are some days where our methods are really challenged, while on most days we have a pretty successful analysis. Overall we produce the empirically important result that realised kernels applied to quote and trade data produce very similar results. Hence for applied workers they can use these methods on either type of data source with some comfort. This analysis is followed by a Conclusion.

2. PRACTICAL IMPLEMENTATION

2.1. Bandwidth selection in practice

Initially Barndorff-Nielsen, Hansen, Lunde, and Shephard (2008a) studied flat-top, unbiased realised kernels but their flat-top estimator is not guaranteed to be non-negative. This work has been extended to the non-negative realised kernels (2) by Barndorff-Nielsen, Hansen, Lunde, and Shephard (2008b) and it is their results we use here. Their optimal bandwidth depends on the

[3] This assumes a smooth kernel, such as the Parzen kernel. If we use a "kinked" kernel, such as the Bartlett kernel, then we need $\eta > 1/2$ to eliminate the variance and the impractical requirement that $H/n \to \infty$ in order to eliminate the bias. Flat-top realised kernels are unbiased and converge at a faster rate, but are not guaranteed to be non-negative. The latter point is crucial in the multivariate case. In the univariate case having a non-negative estimator is attractive but the flat-top kernel is only rarely negative with modern data. However, if $[Y]$ is very small and the ω^2 very large, which we saw on slow days on the NYSE when the tick size was \$1/8, then it can happen quite often when the flat-top realised kernel is used. Of course our non-negative realised kernels do not have this problem. We are grateful to Kevin Sheppard for pointing out these negative days.

[4] This means that $K(X) \xrightarrow{p} [Y]$ at rate $n^{1/5}$, which is not the optimal rate obtained by Barndorff-Nielsen, Hansen, Lunde, and Shephard (2008a) and Zhang (2006), but has the virtue of $K(X)$ being non-negative with probability one, which is not the case of the other estimators available in the literature.

unknown parameters ω^2 and $\int_0^T \sigma_u^4 \mathrm{d}u$, through ξ as spelt out in (3). We estimate ξ very simply by

$$\hat{\xi}^2 = \hat{\omega}^2 \big/ \widehat{\mathrm{IV}} \, ,$$

where $\hat{\omega}^2$ is an estimator of ω^2 and $\widehat{\mathrm{IV}}$ is a preliminary estimate of $\mathrm{IV} = \int_0^T \sigma_u^2 \mathrm{d}u$. The latter is motivated by the fact that $\mathrm{IV}^2 \simeq T \int_0^T \sigma_u^4 \mathrm{d}u$ when σ_u^2 does not vary too much over the interval $[0, T]$, and it is far easier to obtain a precise estimate of IV than of $\sqrt{T \int_0^T \sigma_u^4 \mathrm{d}u}$. In our implementation we use

$$\widehat{\mathrm{IV}} = \mathrm{RV}_{\mathrm{sparse}},$$

which is a subsampled realised variance based on 20 minute returns. More precisely, we compute a total of $1,200$ realised variances by shifting the time of the first observation in 1-second increments. $\mathrm{RV}_{\mathrm{sparse}}$ is simply the average of these estimators[5]. This is a reasonable starting point, because market microstructure effects have negligible effects on the realised variance at this frequency[6]. To estimate ω^2 we compute the realised variance using every q-th trade or quote. By varying the starting point, we obtain q distinct realised variances, $\mathrm{RV}^{(1)}_{\mathrm{dense}}, \ldots, \mathrm{RV}^{(q)}_{\mathrm{dense}}$ say. Next we compute

$$\hat{\omega}^2_{(i)} = \frac{\mathrm{RV}^{(i)}_{\mathrm{dense}}}{2n_{(i)}}, \qquad i = 1, \ldots, q,$$

where $n_{(i)}$ is the number of non-zero returns that were used to compute $\mathrm{RV}^{(i)}_{\mathrm{dense}}$. Finally, our estimate of ω^2 is the average of these q estimates,

$$\hat{\omega}^2 = \frac{1}{q} \sum_{i=1}^{q} \hat{\omega}^2_{(i)}.$$

For the case $q = 1$, this estimator was first proposed by Bandi and Russell (2008) and Zhang, Mykland, and Aït-Sahalia (2005). The reason that we choose $q > 1$ is robustness. For $\hat{\omega}^2_{(i)}$ to be a sensible estimator of $\mathrm{E}(U^2_\tau)$ it is important that $\mathrm{E}(U_{\tau_j} U_{\tau_{j+q}}) = 0$. There is overwhelming evidence against this assumption when $q = 1$, particularly for quote data. See Hansen and Lunde (2006) and the Figures presented later in this paper. So we choose q such that every q-th observation is, on average, 2 minutes apart. On a typical day in our empirical analysis in Section 4, we have $q \approx 25$ for transaction data and $q \approx 70$ for mid-quote data. These values for q are deemed sufficient for $\mathrm{E}(U_{\tau_j} U_{\tau_{j+q}}) = 0$ to be a sensible assumption[7].

So in our empirical analysis we use the expression $\hat{H} = 3.5134 \hat{\xi}^{4/5} n^{3/5}$ to choose the bandwidth parameter for the realised kernel estimator that is based on the Parzen kernel function.

[5]The initial two scale estimator of Zhang, Mykland, and Aït-Sahalia (2005) takes this type of average RV statistic and subtracts a positive multiple of a non-negative estimator of ω^2 — to try to bias adjust for the presence of noise (assuming $Y \perp\!\!\!\perp U$). Hence this two scale estimator must be below the average RV statistic. This makes it unsuitable, by construction, for mid-quote data where RV is typically below integrated variance due to its particular form of noise. Their bias corrected two scale estimator is renormalised and so maybe useful in this context.

[6]$\mathrm{RV}_{\mathrm{sparse}}$ was suggested by Zhang, Mykland, and Aït-Sahalia (2005) and has a smaller sampling variance than a single RV statistic and is more objective, for it does not depend upon the arbitrary choice of where to start computing the returns.

[7]Another issue in using $\mathrm{RV}^{(i)}_{\mathrm{dense}}/(2n_{(i)})$ as an estimator of ω^2, is an implicit assumption that ω^2 is large relative to $[Y]/(2n_{(i)})$. This problem was first emphasised by Hansen and Lunde (2006), who showed that the variance of the noise is very small after the decimalisation, in particular for actively traded assets where they found $\omega^2 \ll 0.001 \cdot [Y]$. The main reason being that the decimalisation has reduced some of the main sources for the noise, U, such as the magnitude of "rounding errors" in the observed prices, and the bid-ask bounces in transaction prices. So our estimator, $\hat{\omega}^2$ is likely to

2.2. End effects

In this section we discuss end-effects. From a theoretical angle we will explain why they show up in this estimation problem, why they are important, and how these effects are eliminated in the computation of the realised kernel. From an empirical perspective, we will then argue they can largely be ignored in practice.

The realised autocovariances, γ_h, $h = 0, 1, \ldots, H$ are not divided by the sample size. This means that the realised kernel is influenced by the noise components of the first and last observations in the sample, U_0 and U_T, respectively. The problem is that $K(U) \xrightarrow{p} U_0^2 + U_T^2 \neq 0$ as $n \to \infty$. The important theoretical implication is that $K(X)$ would be inconsistent if applied to raw price observations. Fortunately, this end-effect problem is easily resolved by replacing the first and last observation by local averages. The implication is that $K(U) = \bar{U}_0^2 + \bar{U}_T^2 + o_p(1)$, where $\bar{U}_0$ and $\bar{U}_T$ both are averages of m, say, observations. If U_t is ergodic with $\mathrm{E}(U_t) = 0$, then it follows that $K(U) \xrightarrow{p} 0$ as $m \to \infty$. So the local averaging at the two end-points eliminates the end-effects.

While the contribution from end-effects are dampened by the local averaging (jittering), a drawback from increasing m is that fewer observations are available for computing the realised kernel. This follows from the fact that $2m$ observations are used up for the two local averages. This trade-off defines a mean-squared optimal choice for m. In practice, the optimal choice for m is often $m = 1$, as shown in Barndorff-Nielsen, Hansen, Lunde, and Shephard (2008b). This is the reason that end-effects can safely be ignored in practice, despite their important theoretical implications for the asymptotic properties of the realised kernel estimator.

Loosely speaking, end-effects can safely be ignored whenever the quadratic variation, $[Y]$, is thought to dominate the size of $U_0^2 + U_T^2$. This is the case for actively traded equities. However for less liquid assets this could be a problem, e.g. on days where the squared spread is, say, 5% of the daily variance of returns. In any case, we now discuss how this local averaging is carried out in practice, for the case $m = 2$, which is the value we use in our empirical analysis.

Write the times at which the price process, X, is being recorded as $0 = \tau_0 \leq \cdots \leq \tau_N = T$. When the recording is being carried out regularly in time, we have $\tau_j - \tau_{j-1} = T/N$, for $j = 1, \ldots, N$, but in practice we typically have irregularly spaced observations. Define the discrete time observations X_0, X_1, ..., X_n where

$$X_0 = \frac{1}{2}\left(X_{\tau_0} + X_{\tau_1}\right),$$

$$X_j = X_{\tau_{j+1}}, \quad j = 1, 2, ..., n-1,$$

and

$$X_n = \frac{1}{2}\left(X_{\tau_{N-1}} + X_{\tau_N}\right).$$

Thus the end points, X_0 and X_n, are local averages of two available prices over a small interval of time. These prices allow us to define the high frequency returns as

$$x_j = X_j - X_{j-1}, \quad j = 1, 2, ..., n,$$

which are used in (2).

be upwards biased, which results in a conservative choice of bandwidth parameter. But there are a couple of advantages in using a conservative value of H. One is that a too small value for H will, in theory, cause more harm than a too large value for H, another is that a larger value of H increases the robustness of the realised kernel to serial dependence in U_τ.

3. PROCEDURE FOR CLEANING THE HIGH-FREQUENCY DATA

Careful data cleaning is one of the most important aspects of volatility estimation from high-frequency. The cleaning of high-frequency data have been given special attension in e.g. Dacorogna, Gencay, Müller, Olsen, and Pictet (2001, chapter 4), Falkenberry (2001), Hansen and Lunde (2006) and Brownless and Gallo (2006). Specifically, Hansen and Lunde (2006) show that tossing out a large number of observations can in fact improve volatility estimators. This result may seem counter intuitive at first, but the reasoning is fairly simple. An estimator that makes optimal use of all data, will typically put high weight on accurate data and be less influenced by the least accurate observations. The generalized least squares (GLS) estimator in the classical regression model is a good analogy. On the other hand, the precision of the standard least squares estimator can deteriorate when relatively noisy observations are included in the estimation. So the inclusion of poor quality observations can cause more harm than good to the least squares estimator and this is the relevant comparison to the present situation. The realised kernel and related estimators "treat all observations equally"and a few outliers can severely influence these estimators.

3.1. Step-by-step cleaning procedure

In our empirical analysis we use trade and quote data from the TAQ database, with the objective of estimating the quadratic variation for the period between 9:30am and 4:00pm. The cleaning of the TAQ high frequency data was carried out in the following steps. P1-P3 was applied to both trade and quote data, T1-T4 are only applicable to trade data, while Q1-Q4 is only applicable to quotation data.

All data

P1. Delete entries with a time stamp outside the 9:30 am to 4 pm window when the exchange is open.

P2. Delete entries with a bid, ask or transaction price equal to zero.

P3. Retain entries originating from a single exchange (NYSE in our application). Delete other entries.

Quote data only

Q1. When multiple quotes have the same timestamp, we replace all these with a single entry with the median bid and median ask price.

Q2. Delete entries for which the spread is negative.

Q3. Delete entries for which the spread is more that 50 times the median spread on that day.

Q4. Delete entries for which the mid-quote deviated by more than 5 mean absolute deviation from a centered mean (excluding the observation under consideration) of 50 observations.

Table 1: Summary statistics for the cleaning and aggregation procedures when applied to data from different exchanges. The first column gives the number of observations observed between 9:30 am and 4:00 pm (P1). Subsequent columns state the reductions in the number of observations due to each of the cleaning/aggregation rules. A blank entry means that the filter was not applied in the particular case. NYSE(N): New York Stock Exchange, PACIF(P): Pacific Exchange, NASD(D): National Association of Security Dealers, NASDAQ(T): National Association of Security Dealers Automated Quotient, in each case the letter in parenthesis is the TAQ identifier.

	Trade date						*Quote data*					
		P2	T1	T2	T3	T4		P2	Q1	Q2	Q3	Q4
January 24, 2007												
NYSE	7,276	0	0	0	2,299	5	42,120	0	28,210	0	0	84
PACIF	6,847	0	0	0	4,678	1	15,910	0	7,768	0	0	14
NASD	9,813	0	0	14	6,365	1	30,230	15	20,630	0	87	109
NASDAQ	0						0					
Other	142	0	0	3			32	3				
January 26, 2007												
NYSE	8,787	0	0	0	3,454	4	51,120	0	36,840	0	0	14
PACIF	4,606	0	0	0	2,824	6	21,510	0	12,020	0	0	2
NASD	10,740	0	0	2	6,728	11	40,130	26	28,920	0	197	104
NASDAQ	0						0					
Other	479	0	0	3			36	3				
May 4, 2007												
NYSE	8,487	0	0	0	3,234	8	48,810	0	34,180	0	0	43
PACIF	4,795	0	0	0	3,117	4	28,680	0	19,250	0	0	2
NASD	1,402	0	0	16	372	2	2,394	0	1,491	0	6	8
NASDAQ	10,130	0	0	0	7,155	0	49,720	0	39,750	0	0	11
Other	485	0	0	1			34,930	88				
May 8, 2007												
NYSE	24,350	0	0	1	14,480	56	109,200	0	90,770	0	0	30
PACIF	24,840	0	0	0	19,100	13	76,900	0	62,390	0	0	3
NASD	6,643	0	4	15	2,384	1	17,000	0	12,910	0	108	3
NASDAQ	42,160	0	0	0	34,480	28	138,100	0	122,600	0	0	8
Other	1,897	0	0	3			102,800	7				

Trade data only

T1. Delete entries with corrected trades. (Trades with a *Correction Indicator*, CORR $\neq 0$).

T2. Delete entries with abnormal *Sale Condition*. (Trades where COND has a letter code, except for "E" and "F"). See the TAQ 3 User's Guide for additional details about sale conditions.

T3. If multiple transactions have the same time stamp: use the median price.

T4. Delete entries with prices that are above the *ask* plus the bid-ask spread. Similar for entries with prices below the *bid* minus the bid-ask spread.

3.2. Discussion of filter rules

The first step P1 identifies the entries that are relevant for our analysis, which focuses on volatility in the 9:30 am to 4 pm interval.

Steps P2 and T1 removes very serious errors in the database, such as misrecording of prices (e.g. zero prices or misplaced decimal point), and timestamps that may be way off. T2 rules out datapoints which the TAQ database is flagging up as a problem. Table 1 gives a summary of the counts of data deleted or aggregated using these filter rules for the database used in Section 4, which analyses the Alcoa share price.

By far the most important rules here are P3, T3 and Q1. In our empirical work we will see the impact of suspending P3. It is used to reduce the impact of time-delays in the reporting of trades and quote updates. Some form of T3 and Q1 rule seems inevitable here, and it is these rules which lead to the largest deletion of data.

T4 is an attractive rule, as it disciplines the trade data using quotes. However, it has the disadvantage that it cannot be applied when quote data is not available.[8] We see from Table 1 that it is rarely activated in practice, while later results we will discuss in Table 2 on realised kernels demonstrate the RK estimator (unlike the RV statistic) is not very sensitive to the use of T4.

It is interesting to compare some of our filtering rules to those advocated by Falkenberry (2001) and Brownless and Gallo (2006). In such a comparison it is mainly the rules designed to purge outliers/misrecordings that could be controversial.

Among our rules Q4 and T4 are the relevant ones. Q4 is very closely related to the procedure Brownless and Gallo (2006, pp. 2237) advocate for removing outliers. They remove observation i if the condition; $|p_i - \bar{p}_i(k)| < 3s_i(k) + \gamma$ is true. Here $\bar{p}_i(k)$ and $s_i(k)$ denote respectively the δ-trimmed sample mean and sample standard deviation of a neighborhood of k observations around i and γ is a granularity parameter. We use the median in place of the trimmed sample mean, $\bar{p}_i(k)$, and the mean absolute deviation from the median in place of $s_i(k)$. By not using the sample standard deviation we become less sensitive to runs of outliers.

Falkenberry (2001) also use a threshold approach to determine if a certain observation is an outlier. But instead of using a "Search and Purge" approach he applies a "Search and Modify" methodology. Prices that deviate with a certain amount from a moving filter of all prices are modified to the filter value. For transactions this has the advantage of maintaining the volume of a trade even if the associated price is bad.

Finally, we note that our approach to discipline the trade data using quotes, T4, has formerly be applied in only Hansen and Lunde (2006), Barndorff-Nielsen, Hansen, Lunde, and Shephard (2006) and Barndorff-Nielsen, Hansen, Lunde, and Shephard (2008a).

[8] In situations where quote data is not available, Q4 can be applied in place of T4, replacing the word mid-quote with price.

Table 2: Sensitivity of RV and RK to our filtering rules P2, T3 and T4 for trade data from Alcoa Inc. on 3 specific days, and averaged across the full sample. Analysis based on data from the common exchanges (NYSE, PACIF, NASD and NASDAQ) and all exchanges (denoted ALL). T3A-E vary how multiple data on single seconds are aggregated. Our preferred method is T3.E, which takes the median prices. The first three columns report the observation count at each stage. T3.● signify that T3A-E all result in the same number of observations.

	No of Observations			*Realised variance*			*Realised kernel*						
	P2	T3.●	T4.E	P2	T3.E	T4.E	P2	T3.A	T3.B	T3.C	T3.D	T3.E	T4.E
January 24, 2007													
NYSE	7,275	4,976	4,971	3.25	2.18	2.12	0.91	0.81	0.83	0.83	0.83	0.82	0.82
PACIF	6,846	2,168	2,167	1.34	1.23	1.04	0.97	0.83	0.83	0.84	0.83	0.83	0.76
NASD	9,812	3,433	3,432	2.65	1.63	1.46	0.95	0.84	0.84	0.83	0.83	0.83	0.83
All	48,155	7,814		14.48	2.94		1.79	0.96	0.95	0.92	0.92	0.92	
January 26, 2007													
NYSE	8,168	5,093	5,086	6.95	5.56	5.61	5.10	5.30	5.31	5.31	5.31	5.31	5.31
PACIF	4,159	1,662	1,658	4.85	4.77	4.76	5.27	5.14	5.14	5.13	5.14	5.13	5.13
NASD	9,827	3,814	3,804	6.20	5.21	5.06	4.79	5.08	5.08	5.08	5.08	5.08	5.08
All	45,259	7,756		24.00	6.15		11.44	5.16	5.17	5.17	5.17	5.18	
May 8, 2007													
NYSE	24,346	9,870	9,798	14.27	6.90	7.14	6.25	6.82	6.73	6.70	6.71	6.70	6.70
PACIF	24,839	5,743	5,731	7.94	5.17	5.13	7.08	7.10	7.09	7.09	7.09	7.09	7.08
NASD	6,642	4,239	4,238	23.69	11.83	8.57	7.57	6.99	7.02	7.02	7.01	7.02	7.04
NASDAQ	42,161	7,678	7,654	7.57	4.89	4.96	6.51	6.89	6.87	6.84	6.87	6.87	6.88
All	199,777	13,584		131.20	6.34		21.25	6.90	6.88	6.88	6.87	6.88	
Averages over full sample													
NYSE	9,737	5,480	5,464	4.89	3.12	3.09	2.45	2.41	2.41	2.40	2.41	2.41	2.41
NASD	4,008	2,150	2,148	13.15	4.00	3.74	2.42	2.36	2.36	2.36	2.36	2.36	2.36
PACIF	7,549	2,343	2,338	2.79	2.40	2.40	2.52	2.43	2.43	2.43	2.43	2.43	2.43
NASDAQ	12,629	3,487	3,411	8.04	2.29	2.38	2.67	2.56	2.55	2.55	2.55	2.55	2.57
All	61,264	8,331		164.30	14.59		6.05	2.53	2.53	2.52	2.52	2.53	

4. DATA ANALYSIS

We analyze high-frequency stock prices for Alcoa Inc., which has the ticker symbol AA. It is the leading producer of aluminum and its stock is currently part of the Dow Jones Industrial Average (DJIA).

We have estimated daily volatility for each of the 123 days in the six-month period from January 3 to June 29, 2007. Much of our discussion will focus on four days that highlight some challenging empirical issues.

The data are transaction prices and quotations from NYSE and all data are from the TAQ database extracted from the Wharton Research Data Services (WRDS). We present empirical results for both transaction and mid-quote prices that are observed between 9:30am and 4:00pm.

We first present results for a regular day, by which we mean a day where the high frequency returns are such that it is straightforward to compute the realised kernel. Then we present empirical results on the use of realised kernels using the entire sample of 123 separate days, indicating the realised kernels behave very well and better than any available realised variance statistic. Then we turn our attention to days where the high-frequency data have some unusual and puzzling features that potentially could be harmful for the realised kernel.

4.1. Sensitivity to data cleaning methods

In Table 2 we give a summary of the various effects of aggregating and excluding observations in different manners.

We have carried out the analysis along two dimensions. First, we have separated data from different exchanges. Specifically, we consider trades on NYSE, PACIF, NASD and NASDAQ in isolation. We also investigate the performance of the estimator when all exchanges are consider simultaneously, which is the same as dropping P3 entirely. This defines the first dimension that is displayed in the rows of Table 2, for three of the four days we give special attention, and averaged over the full sample for AA..

Our second dimension is the amount of cleaning, aggregation and filtering which we apply to the data. With reference to the cleaning and filtering step in section 3.1, the columns of Table 2 have the following information.

P2: This is the data with a time stamp inside the 9:30 am to 4 pm window when most the exchanges are open. We have deleted entries with a bid, ask or transaction price equal to zero. So this is basically the raw data, with the only purged observations being clearly nonsense ones.

T3.A-E: This is what is left after step T.3. The different letters represent five different ways of aggregating transactions that have the same time stamp:

A. First single out unique prices and aggregate volume. Then use the price which has the largest volume.

B. First single out unique prices and aggregate volume. Then use the price by volume weighted average price.

C. First single out unique prices and aggregate volume. Then use the price by log(volume) weighted average price.

D. First single out unique prices and aggregate volume. Then use the price by number of trades weighted average price.

E. Use the median price. This is the method which we used in the paper.

T4.E This is what is left after rounding step T.4 on the data left after T3.E.

In Table 2 we present observation counts, realised variances and realised kernels. Two things are particular conspicuous. On January 24th at PACIF only one observations was filtered out by T4.E, still both the realised variance and the realised kernels are quite sensitive to whether this observation is excluded — it is the only day and exchange where this is the case. In the left panel of Figure 1 we display the data around this observation, and it is clear that it is out of line with the rest. Also May 8th at NASD only one observations was filtered out by T4.E, here only the realised variance is quite sensitive to whether this observation is excluded. In the right panel of Figure 1 we display the data around this observation, and again it is clear that it is out of line with the rest. Hence we conclude that T4 is useful when it can be applied in practice, but it does not usually make very much difference in practice when RK estimators are used.

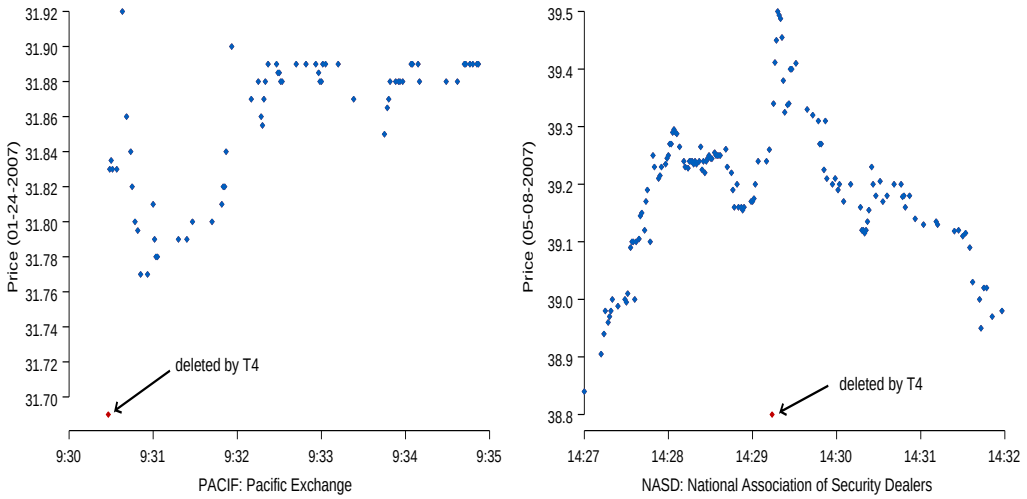

Figure 1: This figure shows transaction prices for Alcoa Inc. over a period of 5 minutes surrounding one observation deleted by T4.E. The left panel display January 24th on PACIF, and the right panel show the scenario at May 8th on NASD.

An noteworthy feature of Table 2 is how badly RV does when we aggregate data across exchanges and only apply P2 — basically only implementing trivial cleaning. The upward bias we see for RV when based on trade-by-trade data is dramatically magnified. Some of this is even picked up by the RK statistic, which significantly benefits from the application of T3. It is clear from this table that if one wanted to use information across exchanges, then it is better to carry out RK on each exchange separately and then average the answers across the exchanges rather than treat all the data as if it was from a single source.

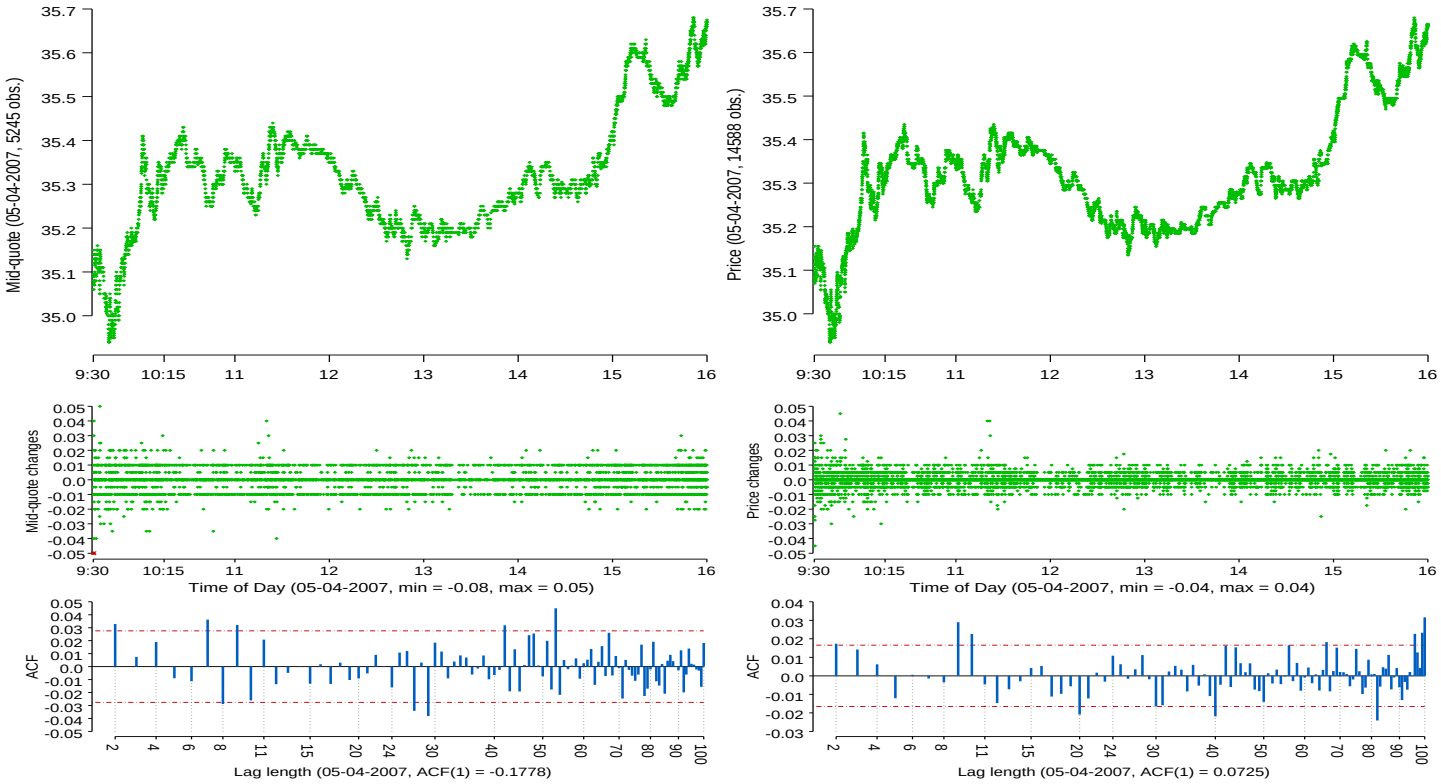

Figure 2: High-frequency prices and returns for Alcoa Inc. (AA) on May 4, 2007, and the first 100 autocorrelations for tick-by-tick returns. Left panels are for transaction prices and right panels are for mid-quote prices. Returns larger than 5 cents in absolute value are marked by red dots in the middle panels. The largest and smallest (most negative) returns are reported below the middle panels. Lower panels display the autocorrelations for tick-by-tick returns, starting with the second-order autocorrelation. The numerical value of the first-order autocorrelation is given below these plots. A log-scale is used for the x-axis such that the values for lower-order autocorrelations are easier to read of the plots.

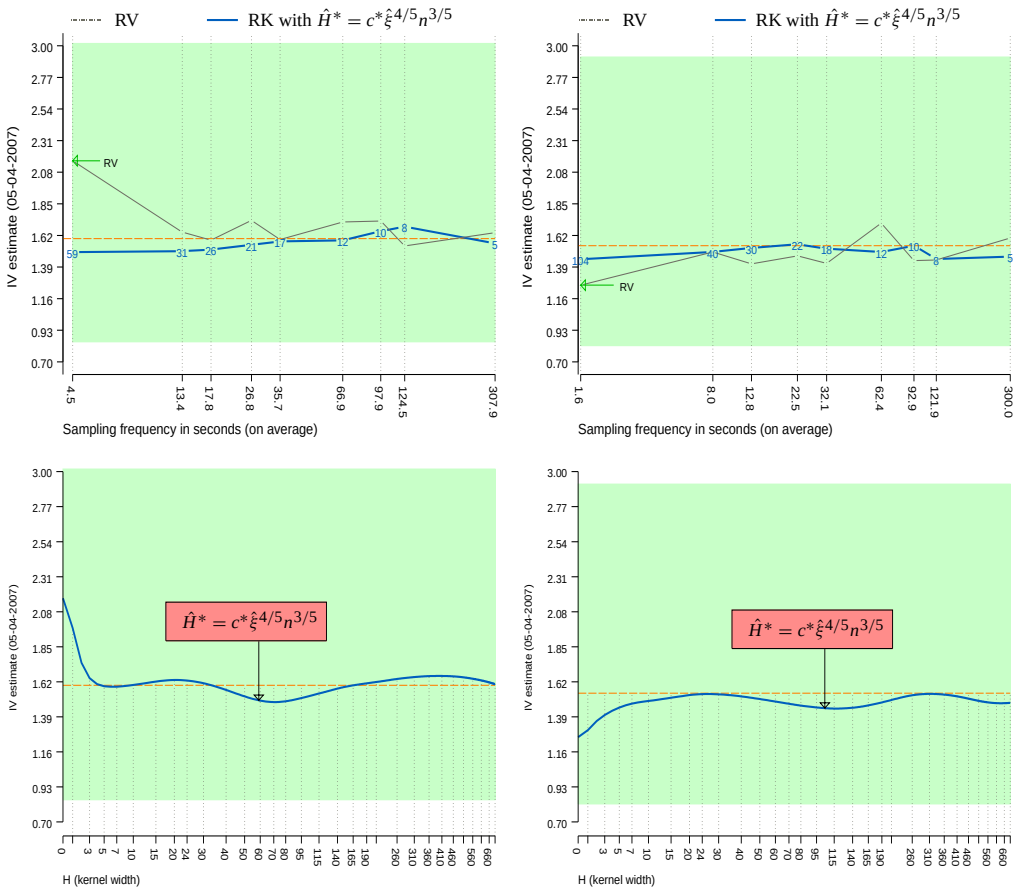

Figure 3: Signature plots for the realised kernel and realised variance on May 4, 2007 for Alcoa Inc.. Those based on transaction prices are plotted in left panels and those based on mid-quote prices are plotted in right panels. The horizontal line in these plots is the subsampled realised variances based 20-minute returns. The thicker dark line in the upper panels represents the realised kernels using the bandwidth $\hat{H}^* = c^*\hat{\xi}^{4/5}n^{3/5}$, and the thin line is the usual realised variance. The lower panels is a different sort of signature plot for the realised kernel. Here we plot the point estimates of the realised kernel as a function of the bandwidth, H, where the sampling frequency is the same (tick-by-tick returns) for all realised kernels. Our estimate of the optimal bandwidth is highlighted in the lower panels.

4.2. A regular day: May 4, 2007

Figure 2 shows the prices that were observed in our database after being cleaned. They are based on the irregularly spaced times series of transaction (left) and mid-quote (right) prices on May 4, 2007. The two upper plots show the actual tick-by-tick series, comprising 5, 246 transactions and 14, 631 quotations recorded on distinct seconds. Hence for transactions data we have a new observation on average every five seconds, while for mid-quotes it is more often than every

couple of seconds. In the middle panel the corresponding price changes are displayed, changes above 5 cents and below minus 5 cents are marked by a large star (red) and are truncated (in the picture) at ± 5 cents. May 4 was a quite tranquil day with only a couple of changes outside the range of the plot. The lower panel gives the autocorrelation function of the log-returns. The acf(1) is omitted from the plot, but its value is given in the subtext. For the transaction series the acf(1) is about -0.24, which is fundamentally different from the one found for the mid-quote series that equals 0.088. This difference is typically for NYSE data as first noted in Hansen and Lunde (2006). It is caused by the more smooth character of most mid-quote series, that induces a negative correlation between the innovations in Y and the innovations in U. The negative correlation results in a smaller, possibly negative, bias for the RV, and this feature of mid-quote data will be evident from Figure 5, which we discuss in the next subsection.

May 4, 2007 is an exemplary day. The upper panels of Figure 3 present volatility signature plots[9] for irregularly spaced times series of transaction prices (left) and mid-quote prices (right)[10]. The dark line is the Parzen kernel with $H = c^* \xi^{4/5} n^{3/5}$ and the light line is the simple realised variance. The lower panel present a kernel signature plot where the realised kernel computed on tick-by-tick data is plotted against increasing values of H. In these plots we have indicated the optimal choices of H. In both plots the horizontal line is an average of simple realised variances based on 20 minute returns sampled with different offsets. The shaded areas denote the 95% confidence interval based on 20 minute returns using the (Barndorff-Nielsen and Shephard 2002) feasible realised variance inference method. We characterize May 4, 2007 as an exemplary day, because the signature plots are almost horizontal. This shows that the realised kernel is insensitive to the choice of sampling frequency. An erratic signature plot indicates potential data issues, although pure chance is also a possible explanation.

4.3. General Features of Results Across Many Days

Transaction prices and mid-quote prices are both noisy measures of the latent "efficient prices", polluted by market microstructure effects. Thus, a good estimator is one that produces almost the same estimate with transaction data and mid-quote data. This is challenging as we have seen the noise has very different characteristics in these two series.

Figure 4 presents scatterplots where estimates based on transaction data are plotted against the corresponding estimates based on mid–quote data. The upper two panels are scatterplots for the realised kernel using tick-by-tick data (left) and the upper right plot is the realised kernel based on 1-minute returns, and both scatter plots are very close to the 45° suggesting that the realised kernel produce accurate estimates at this sampling frequencies, with little difference between the two graphs. The lower four panels are scatterplot for the realised variance using different sampling frequencies: Tick-by-tick returns (middle left), 1-minute returns (middle right), 5-minute returns (lower left), and 20-minute returns (lower right). These plots strongly suggest

[9] To construct volatility signature plots we use activity fixed tick time where the sampling frequency is chosen such that we get approximately the same number of observations each day. To explain it assume that the first trade at the i'th day occurred at time t_{i0} and the last trade on the i'th day occurred at time t_{in_i}. So approximate *60 second* sampling is constructed as follows. We get the tick time sampling frequency on day i as $\lceil 1 + n_i 60/(t_{in_i} - t_{i0}) \rceil$. In this way there will be approximately 60 seconds between observations when one takes the intraday average over the sampled intratrade durations. The actual sampled durations will in general be more or less widely dispersed.

[10] These pictures extend the important volatility signature plots for realised volatility introduced by Andersen, Bollerslev, Diebold, and Labys (2000).

that the realised variance is substantially less precise than the realised kernel. The realised variance based on tick-by-tick returns is strongly influenced by market microstructure noise. But the characteristics of market microstructure noise in transaction prices are very different from those of mid-quote prices. Thus, as already indicated, the trade data causes the realised variances to be upward biased, while for quote data it is typically downward bias. This explains that the scatterplot for tick-by-tick data (middle left) is shifted away from the 45° degree line.

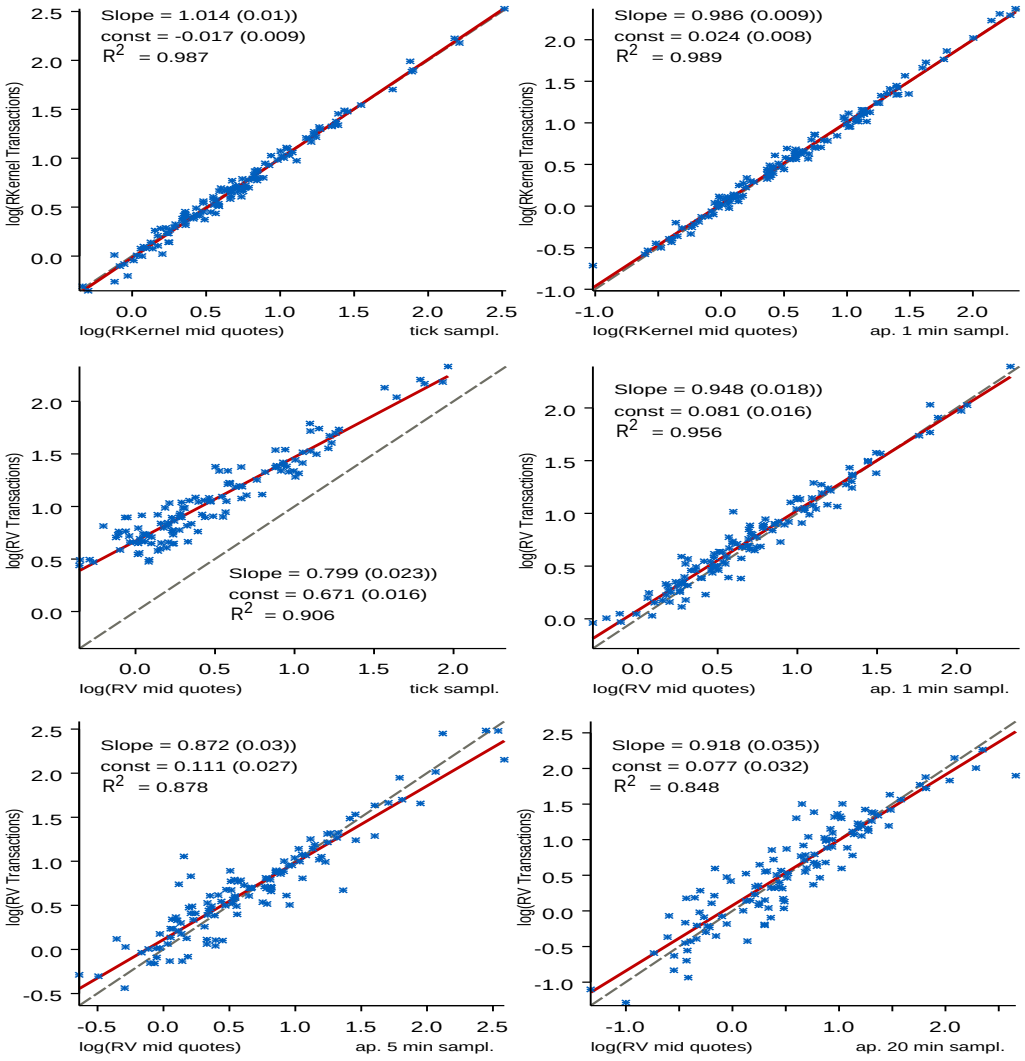

Figure 4: Scatterplots of estimates based on transaction prices plotted against the estimates based on mid-quote prices for Alcoa Inc.. Regression lines and regression statistics are included in the plots along with the 45° line.

Table 3 reports a measure for the disagreement between the estimates based on transaction prices and mid-quote prices. The statistics computed in the first row are the average Euclidian

distance from the pair of estimators to the 45° degree line. To be precise, let $V_{T,t}$ and $V_{Q,t}$ be estimators based on transaction date and quotation data, respectively, on day t, and let $\bar{V}_t$ be the average of the two. The distance from (V_T, V_Q) to the 45° degree line is given by

$$\sqrt{(V_{T,t} - \overline{V}_t)^2 + (V_{Q,t} - \overline{V}_t)^2} = \left|V_{T,t} - V_{Q,t}\right| / \sqrt{2},$$

and the first row of Table 3 reports the average of this distance computed over the 123 days in our sample.

Table 3: This Table present statistics that measure the disagreement between the daily estimates based on transaction prices and mid-quote prices.

	Realized Kernel		Simple Realised variance			
	tick	1 min	tick	1 min	5 min	20 min
Alcoa Inc (AA)						
Distance	0.089	0.105	1.119	0.170	0.312	0.406
Relative Distance	1.000	1.182	12.62	1.922	3.523	4.575
American International Group, Inc(AIG)						
Distance	0.020	0.038	0.458	0.061	0.088	0.132
Relative Distance	1.000	1.892	22.75	3.035	4.382	6.558
American Express (AXP)						
Distance	0.079	0.060	0.578	0.133	0.166	0.248
Relative Distance	1.000	0.755	7.277	1.669	2.095	3.117
Boeing Company (BA)						
Distance	0.047	0.051	0.564	0.106	0.121	0.242
Relative Distance	1.000	1.083	11.96	2.246	2.567	5.132
Bank of America Corporation (BAC)						
Distance	0.028	0.070	0.620	0.050	0.084	0.345
Relative Distance	1.000	2.509	22.21	1.775	3.004	12.35
Citigroup (C)						
Distance	0.033	0.052	0.722	0.080	0.139	0.250
Relative Distance	1.000	1.604	22.12	2.467	4.270	7.664

The distance is substantially smaller for the realised kernels than any of the realised variances, while our preferred estimator, the realised kernel based on tick-by-tick returns, has the least disagreement between estimates based on transaction data and those based on quote data. The relative distances are reported in the second row of Table 3, and we note that the disagreement between any of the realised variance estimators is more than twice that of the realised kernel.

Table 4: Summary statistics for realised kernel and realised variance estimators, applied to transaction prices or mid-quote prices at different sampling frequencies for Alcoa Inc.. The empirical correlations between the realised kernel based on tick-by-tick transaction prices and each of the estimators are given in column 4 and some empirical autocorrelations are given in columns 5-8.

	Mean (HAC)	Std.	$\rho([\widehat{Y}], K)$	acf(1)	acf(2)	acf(5)	acf(10)
Realised kernels based on transaction prices							
1 tick	2.349 (0.299)	1.865	1.000	0.53	0.31	-0.07	0.10
1 minute	2.270 (0.322)	2.060	0.955	0.47	0.25	-0.06	0.11
RV based on transaction prices							
1 tick	3.692 (0.349)	2.483	0.950	0.46	0.30	-0.13	0.07
1 minute	2.458 (0.263)	1.776	0.977	0.51	0.31	-0.11	0.06
5 minute	2.385 (0.293)	1.955	0.939	0.43	0.23	-0.08	0.05
20 minute	2.254 (0.280)	1.762	0.850	0.35	0.21	0.02	0.09
Realised kernels based on mid-quotes							
1 tick	2.349 (0.290)	1.836	0.998	0.52	0.31	-0.08	0.09
1 minute	2.218 (0.308)	1.970	0.937	0.44	0.24	-0.05	0.12
RV based on mid-quotes							
1 tick	1.754 (0.182)	1.252	0.923	0.46	0.28	-0.10	0.11
1 minute	2.317 (0.258)	1.632	0.977	0.52	0.34	-0.09	0.07
5 minute	2.516 (0.374)	2.395	0.971	0.48	0.28	-0.07	0.08
20 minute	2.268 (0.350)	2.308	0.868	0.38	0.19	0.01	0.05

Table 4 contains summary statistics for realised kernel and realised variance estimators for the Alcoa Inc. data over our 123 distinct days. The estimators are computed with transaction prices and mid-quote prices using different sampling frequencies. The sample average and standard deviation is given for each of the estimators and the fourth column has the empirical correlations between each of the estimators and the realised kernel based on tick-by-tick transaction prices. The Table confirms the high level of agreement between the realised kernels estimator based on transaction data and mid-quote data. They have the same sample mean and the sample correlation is nearly one. The time series standard deviation of the daily mid-quote based realised kernel is marginally lower than that for the transaction based realised kernel. The Table also shows the familiar upward bias of the tick-by-tick trade based RV and downward bias of the mid-quote version. Low frequency RV statistics have more variation than the tick-by-tick RK, while the RK statistic behaves quite like the 1-minute mid-quote RV.

Figure 5 contains histograms that illustrate the dispersion (across the 123 days in our sample) of various summary statistics. In a moment we will provide a detailed analysis of three other days, and we have marked the position of these days in each of the histograms. As is the case in most figures in this paper, the left panels correspond to transaction data and right panels to mid-quote data. The first row of panels present the log-difference between the realised kernel

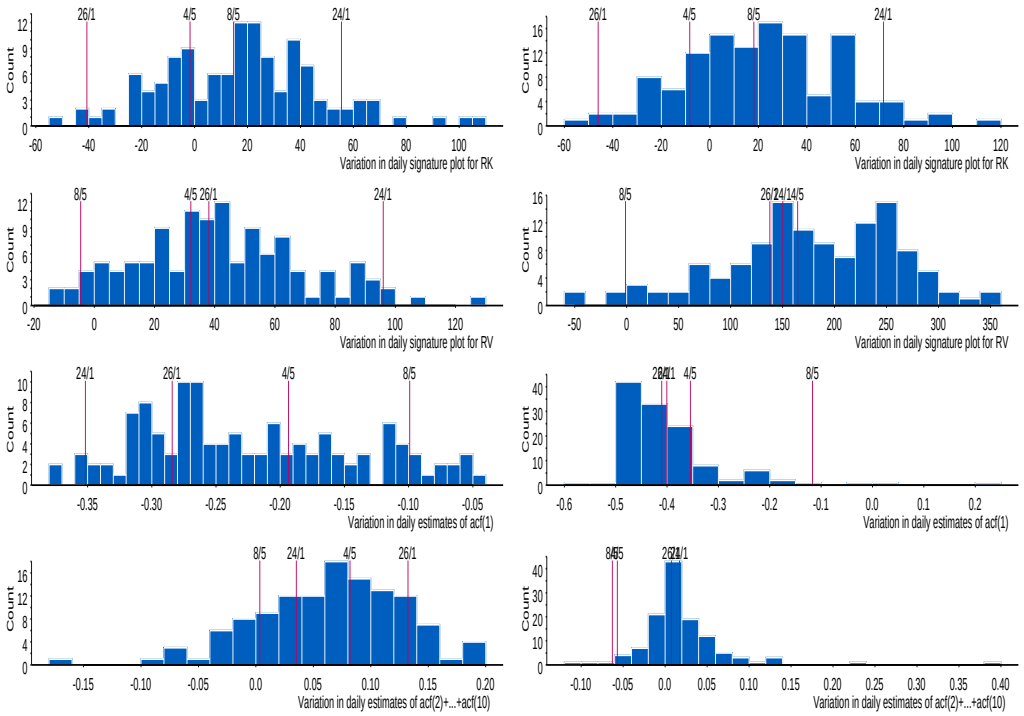

Figure 5: Histograms for various characteristics of the 102 days in our sample. Left panels are for transactions prices, right panels are for mid-quote prices. The two upper panels are histograms for the difference between the realised kernel based on 1-tick returns and that based on five-minute returns. The panels in the second row are the corresponding plots for the realised variance. Histograms of the first order autocorrelation are displayed in the panels in the third row. Finally the fourth and last row of panels are histograms for the sum of the 2nd to the 10th autocorrelation. We have identified the four days that we provide detailed results for in each of the histograms.

computed with tick-by-tick returns and the realised kernel based on five-minute returns. The day we analysed in greater details in the previous subsection, May 4th, is fairly close to the median in all of these dimensions. The three other days, May 8th, January 24th, and January 26th, are our examples of *challenging days.* January 24th and January 26th are placed in the two tails of the histogram related to the variation in the realised kernel. The three other dimensions we provide histograms for are: (2nd row) The log-difference between the realised variance computed with tick-by-tick returns and that computed with five minute returns; (3rd row) the distribution of the estimated first-order autocorrelation; and the 4th row contains histograms for the sum of the next nine autocorrelations (acf(2) through acf(10)).

Note the bias features of the realised variance that is shown in the second row of histograms. For transaction data the tick-by-tick realised variance tends to be larger than the realised variance sampled at lower frequencies, whereas the opposite is true for mid-quote data.

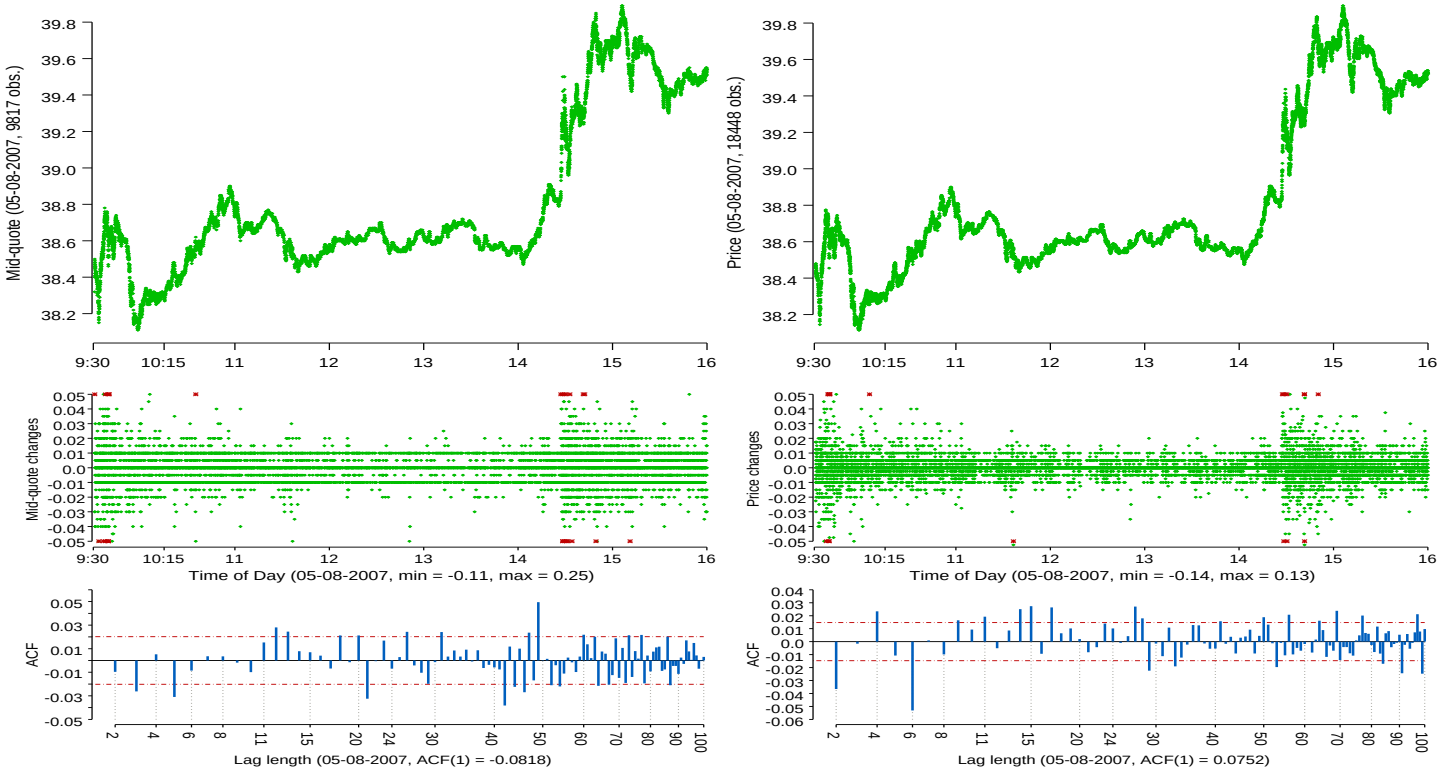

Figure 6: High-frequency prices and returns for Alcoa Inc. on May 8, 2007, and the first 100 autocorrelations for tick-by-tick returns. Left panels are for transaction prices and right panels are for mid-quote prices. Returns larger than 5 cents in absolute value are marked by red dots in the middle panels. The largest and smallest (most negative) returns are reported below the middle panels. Lower panels display the autocorrelations for tick-by-tick returns, starting with the second-order autocorrelation. The numerical value of the first-order autocorrelation is given below these plots. A log-scale is used for the x-axis such that the values for lower-order autocorrelations are easier to read of the plots.

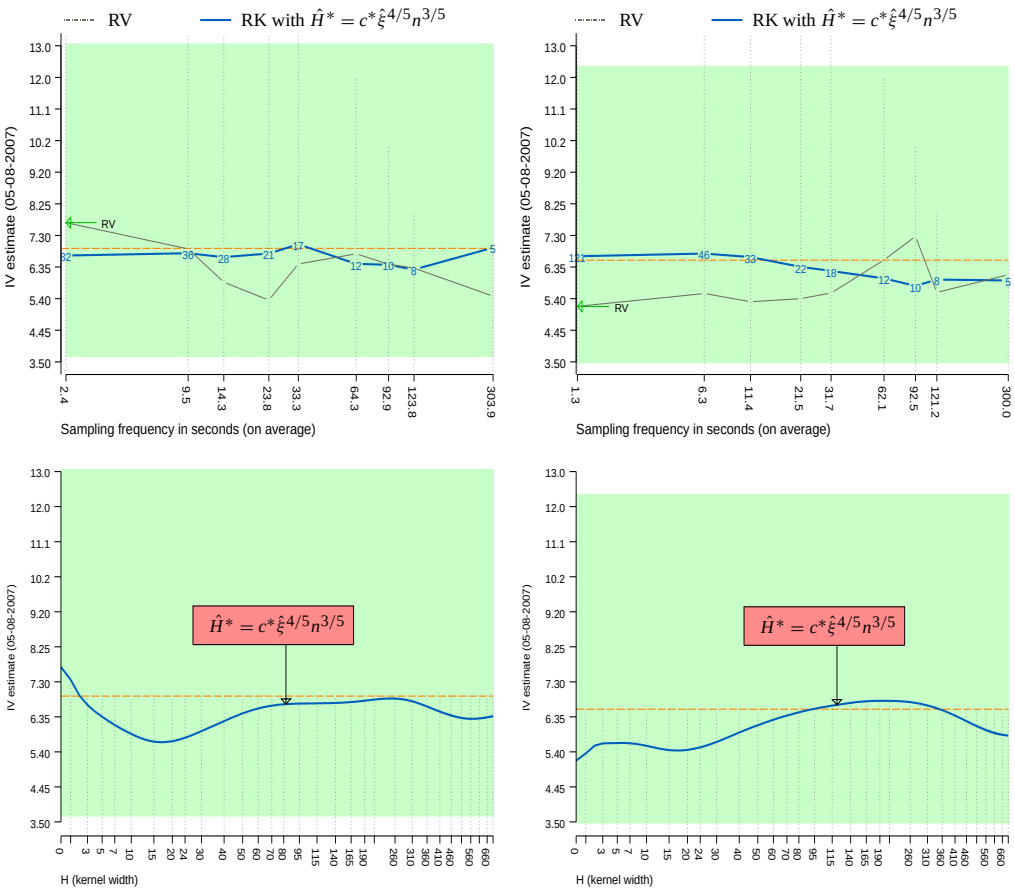

Figure 7: Signature plots for the realised kernel and realised variance for Alcoa Inc. on May 8, 2007. For details see Figure 3.

Next we turn to three potentially harder days which have features which are challenging for the realised kernel. These days were selected to reflect important empirical issues we have encountered when computing realised kernels across a variety of datasets.

4.4. A heteroskedastic day: May 8, 2007

We now look in detail at a rather different day, May 8th, 2007. Figure 6 suggests that this day has a lot of heteroskedasticity, with a spike in volatility at the end of the day. This day is also characterized by several large changes in the price. The transaction price changed by as much as 25 cents from one trade to the next and the mid-quote price by as much as 19 cents over a single quote update. Informally this is suggestive of jumps in the process. Although jumps can alter the optimal choice of H, they do not cause inconsistency in the realised kernel estimator.

The middle panels of Figure 6 visualise the different behaviour of the price throughout the day. The jump in volatility around 14:30 is quite clear from these plots.

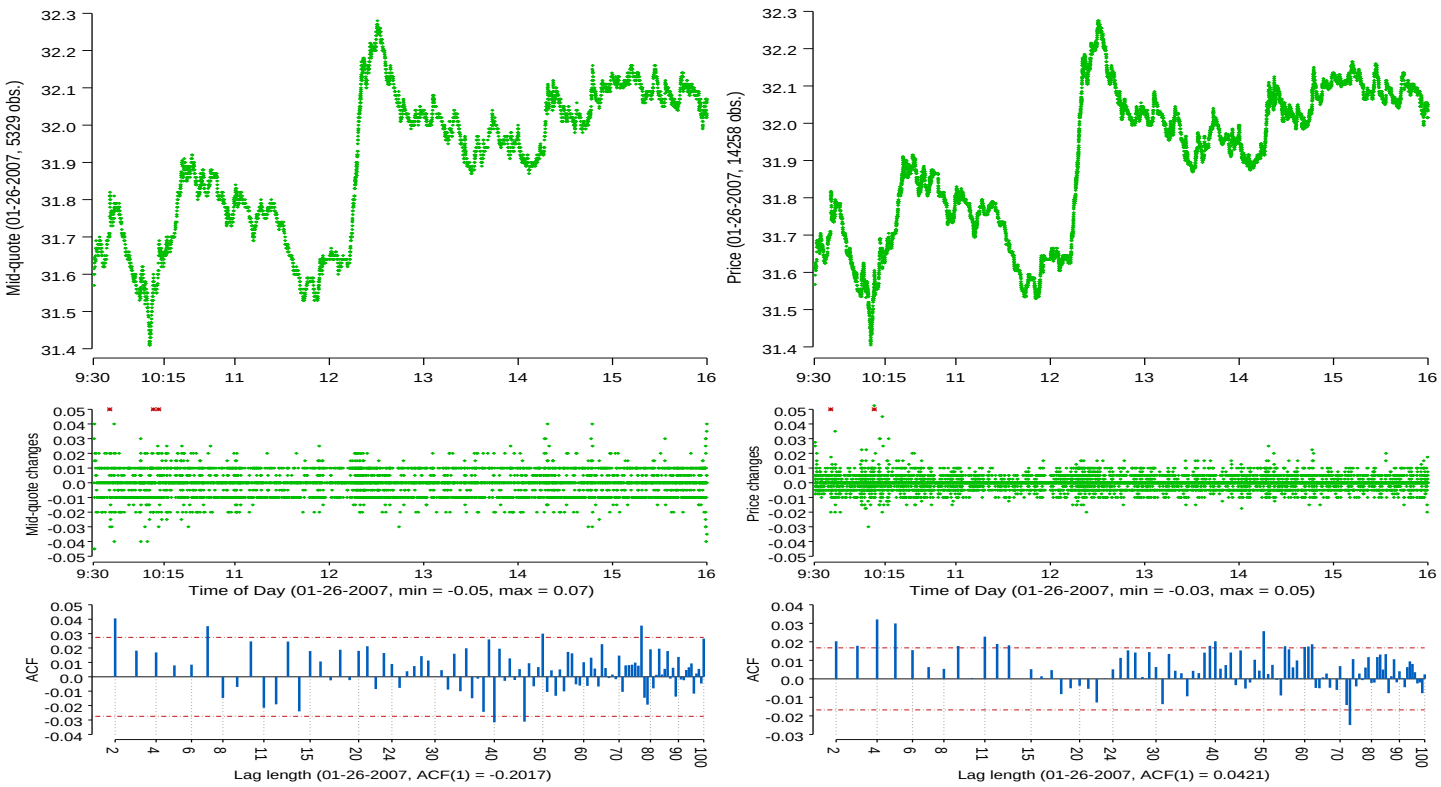

Figure 8: High-frequency prices and returns for Alcoa Inc. on January 26, 2007, and the first 100 autocorrelations for tick-by-tick returns. Left panels are for transaction prices and right panels are for mid-quote prices. Returns larger than 5 cents in absolute value are marked by red dots in the middle panels. The largest and smallest (most negative) returns are reported below the middle panels. Lower panels display the autocorrelations for tick-by-tick returns, starting with the second-order autocorrelation. The numerical value of the first-order autocorrelation is given below these plots. A log-scale is used for the x-axis such that the values for lower-order autocorrelations are easier to read of the plots.

In spite of the jump in volatility, and possibly jumps in the price process, Figure 7 offers little to be concerned about, in terms of the realised kernel estimator. Again the volatility signature plot is reasonably stable for both transaction prices and mid-quote prices and so one has quite some confidence in the estimate.

4.5. A "gradual jump": January 26, 2007

The high-frequency prices for January 26 is plotted in Figure 8. On this day the price increases by nearly 1.5% between 12:13 and 12:20. The interesting aspect of this price change is the gradual and almost linear manner by which the price increases in a large number of smaller increments. Such a pattern is highly unlikely to be produced by a semimartingale adapted to the natural filtration. The gradual jump produces rather disturbing volatility signature plots in Figure 9, that shows that the realised kernel is highly sensitive to the bandwidth parameter. This is certainly a challenging day.

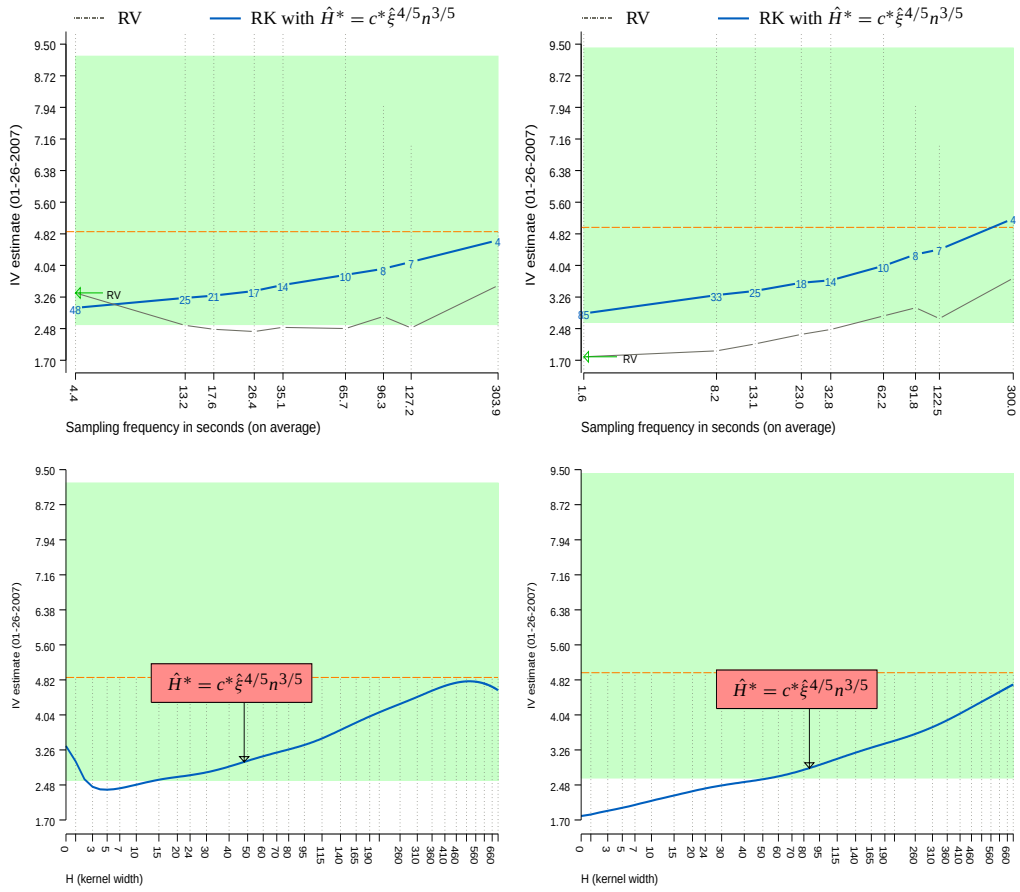

Figure 9: Signature plots for the realised kernel and realised variance for Alcoa Inc. on January 26, 2007. For details see Figure 3.

We zoom in on the gradual jump in Figure 10. The upper left panel has 96 upticks and 43 downticks. The lower plot shows that the volume of the transactions in the period that the price changes are not negligible, in fact, the largest volume trades on January 26 are in this period.

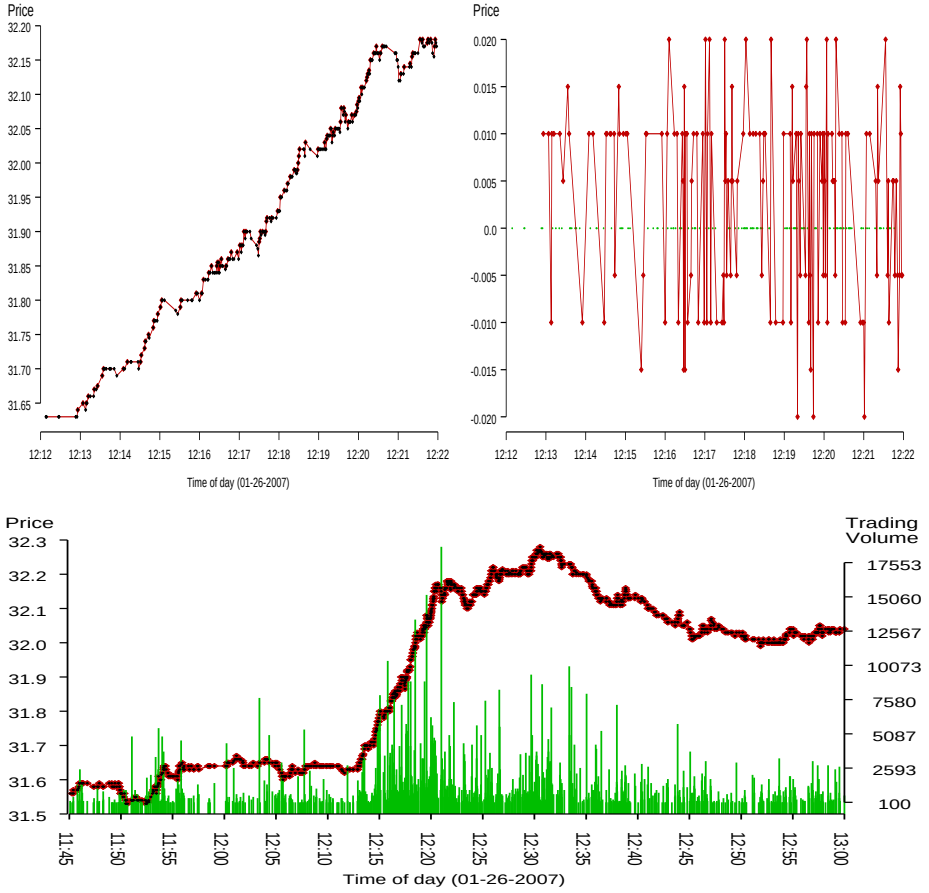

Figure 10: This Figure zooms in on the "gradual" jump on January 26, 2007. Prices and returns in the period from 2:12 pm to 12:22 pm are shown in the two upper panels. The lower panel shows the prices and volume (vertical bars) between 11:45 am and 1:00 pm.

One possible explanation of this is that there is one or a number of large funds wishing to increase their holding of Alcoa (perhaps based on private information) and as they buy the shares they consume the immediately available liquidity — they could not buy more at that price, the instantaneous liquidity may not exist, it can only be met by waiting for it to refill. If the liquidity had existed then the price may have shot up in a single move.

An explanation of such a scenario can be based on market microstructure theory (see e.g. the survey s by O'Hara (1995) or Hasbrouck (2007)). Dating back to Kyle (1985) and Admati and Pfleiderer (1988a, 1988b, 1989) the idea is to model the trading environment as comprising three kinds of traders: risk neutral insiders, random noise trades and risk neutral market makers. The noise trades are also known as liquidity traders, because they trade for reasons that are not

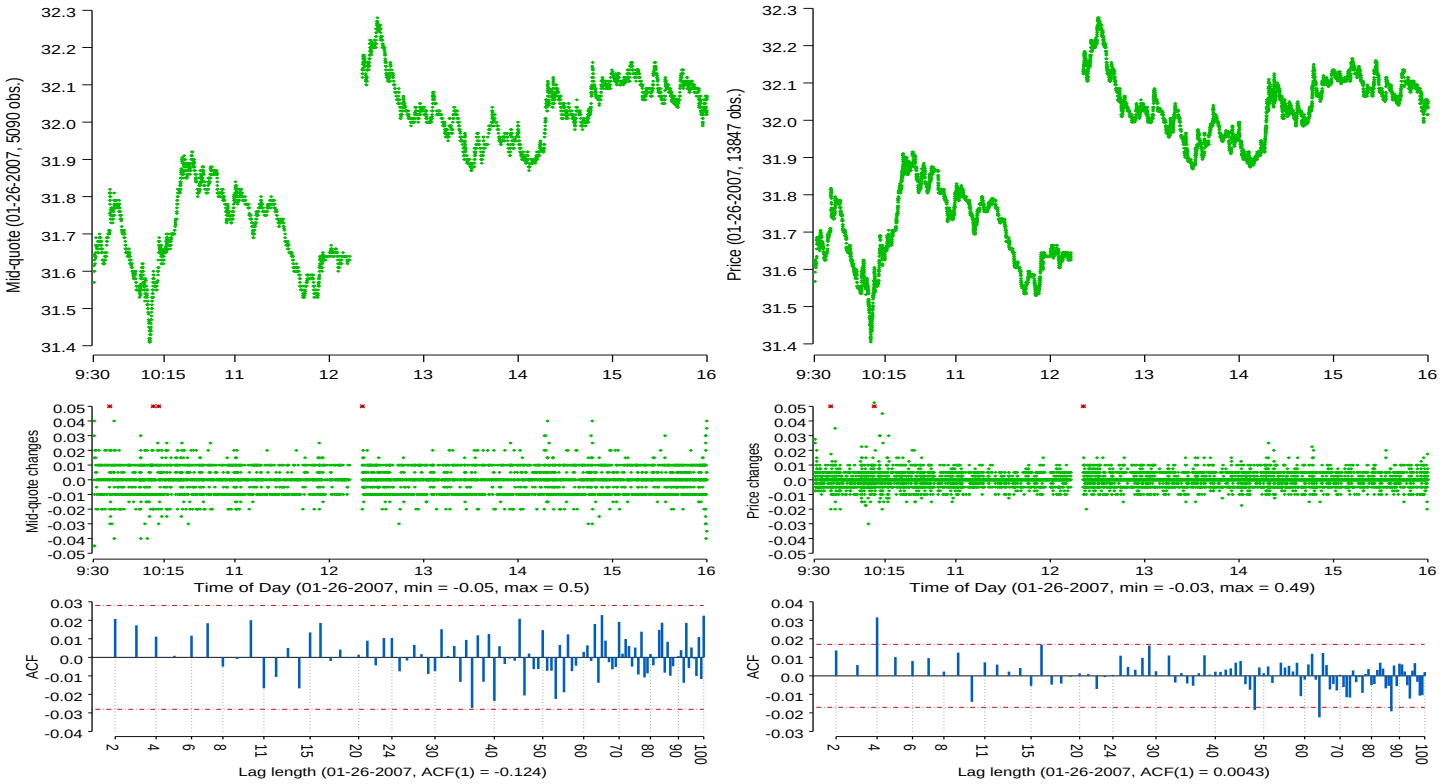

Figure 11: High-frequency prices and returns for Alcoa Inc. on January 26, 2007, and the first 100 autocorrelations for tick-by-tick returns, after prices between 12:13 and 12:21 are removed from the sample. The autocorrelations are very different after having removed observations between 12:13 and 12:21. Compare with Figure 8.

directly related to the expected value of the asset. As such they provide liquidity and it is their presence that explain what we encounter in Figure 10. An implication of the theory is that without these noise traders there would be no one willing to sell the asset on the way up to the new price level at 12:25. We would instead have seen a genuine jump.

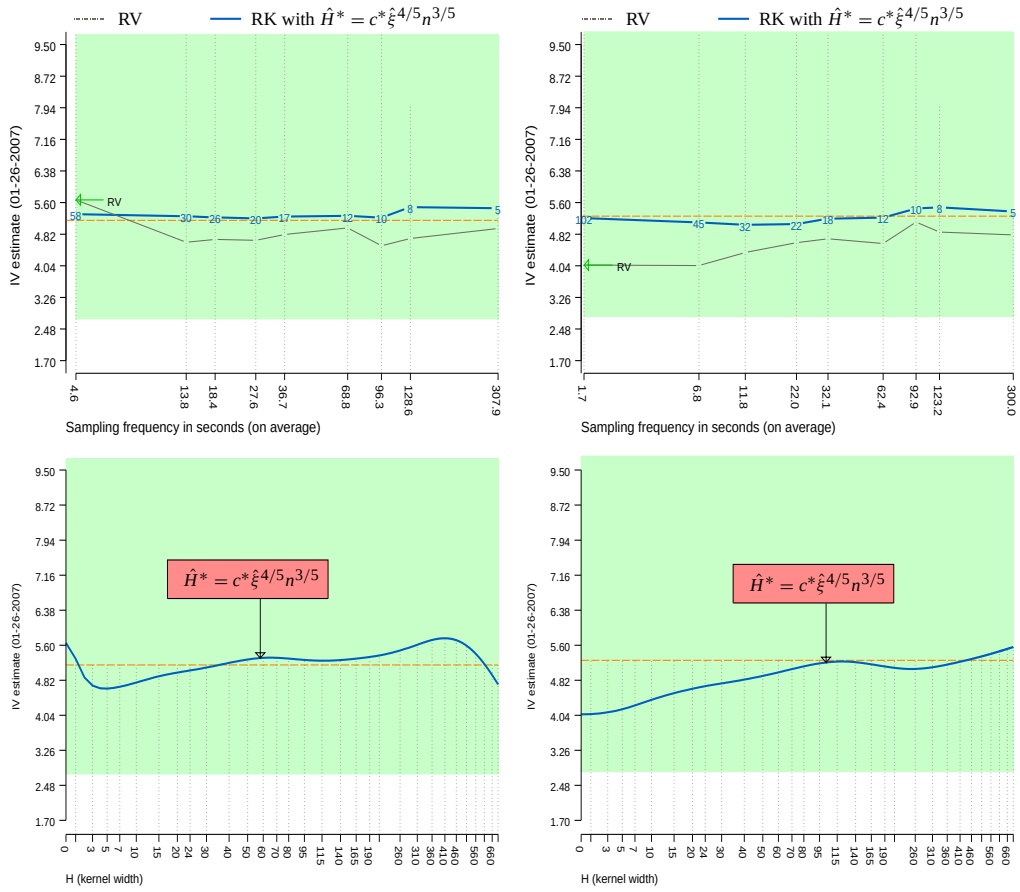

Figure 12: Signature plots for the realised kernel and realised variance for Alcoa Inc. on January 26, 2007, after deleting the prices between 12:13 pm and 12:21 pm. Compare with Figure 9 for details.

Mathematically we can think of a gradual jump in the following way. The efficient price jumps at time τ_j by ΔY_{τ_j} but $\Delta X_{\tau_j} \simeq 0$ which means that

$$\Delta Y_{\tau_j} \simeq -\Delta U_{\tau_j}.$$

Hence the noise process is now far from zero. As trade or quote time evolves the noise trends back to zero, revealing the impact of the jump on X, but this takes a considerable amount of new observations if the jump is quite big. This framework suggests a simple model

$$\begin{aligned} U_{\tau_j} &= V_{\tau_j} + \varepsilon_{\tau_j}, \\ V_{\tau_j} &= \rho V_{\tau_{j-1}} - \theta_{\tau_j}\Delta Y_{\tau_j}, \quad \rho \in [0,1), \end{aligned}$$

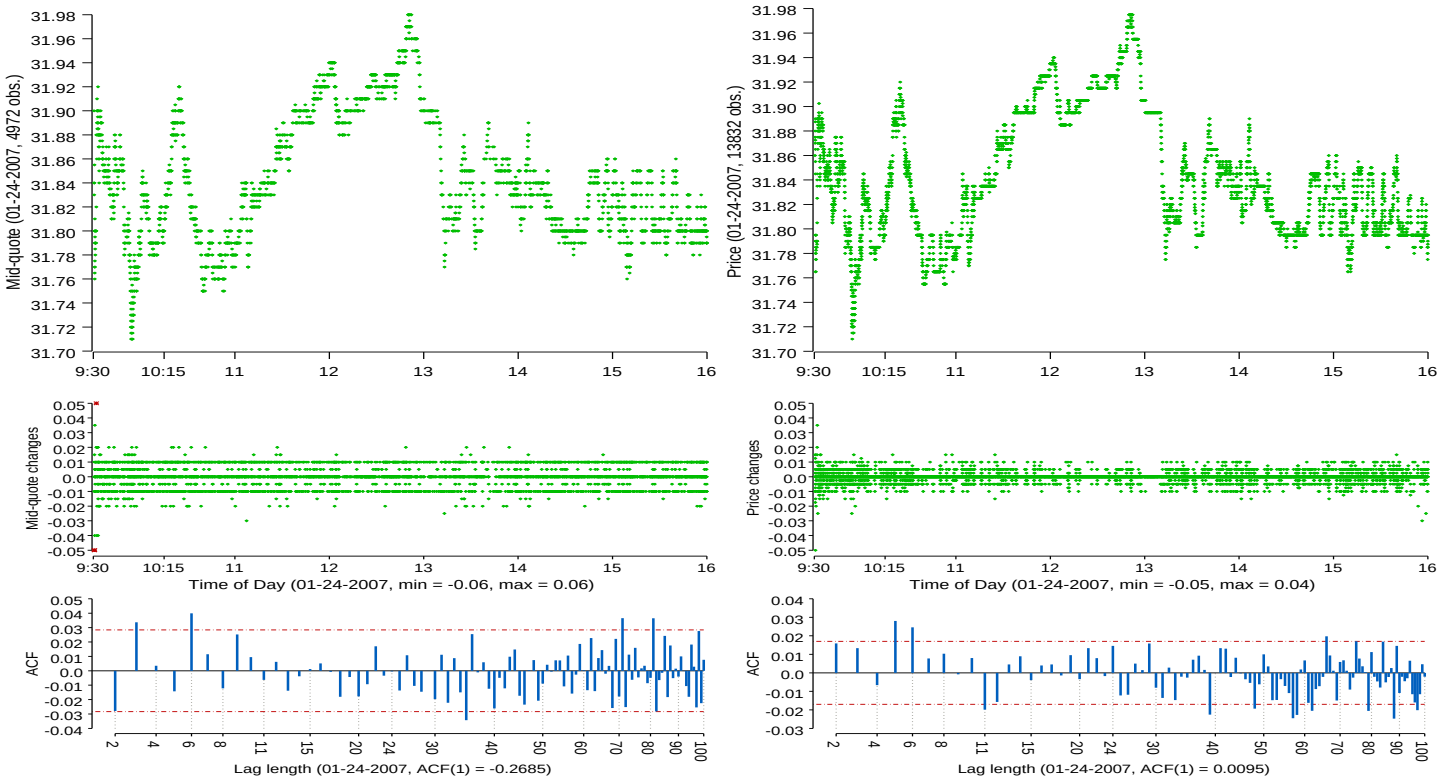

Figure 13: High-frequency prices and returns for Alcoa Inc. on January 24, 2007, and the first 100 autocorrelations for tick-by-tick returns. Left panels are for transaction prices and right panels are for mid-quote prices. Returns larger than 5 cents in absolute value are marked by red dots in the middle panels. The largest and smallest (most negative) returns are reported below the middle panels. Lower panels display the autocorrelations for tick-by-tick returns, starting with the second-order autocorrelation. The numerical value of the first-order autocorrelation is given below these plots. A log-scale is used for the x-axis such that the values for lower-order autocorrelations are easier to read of the plots.

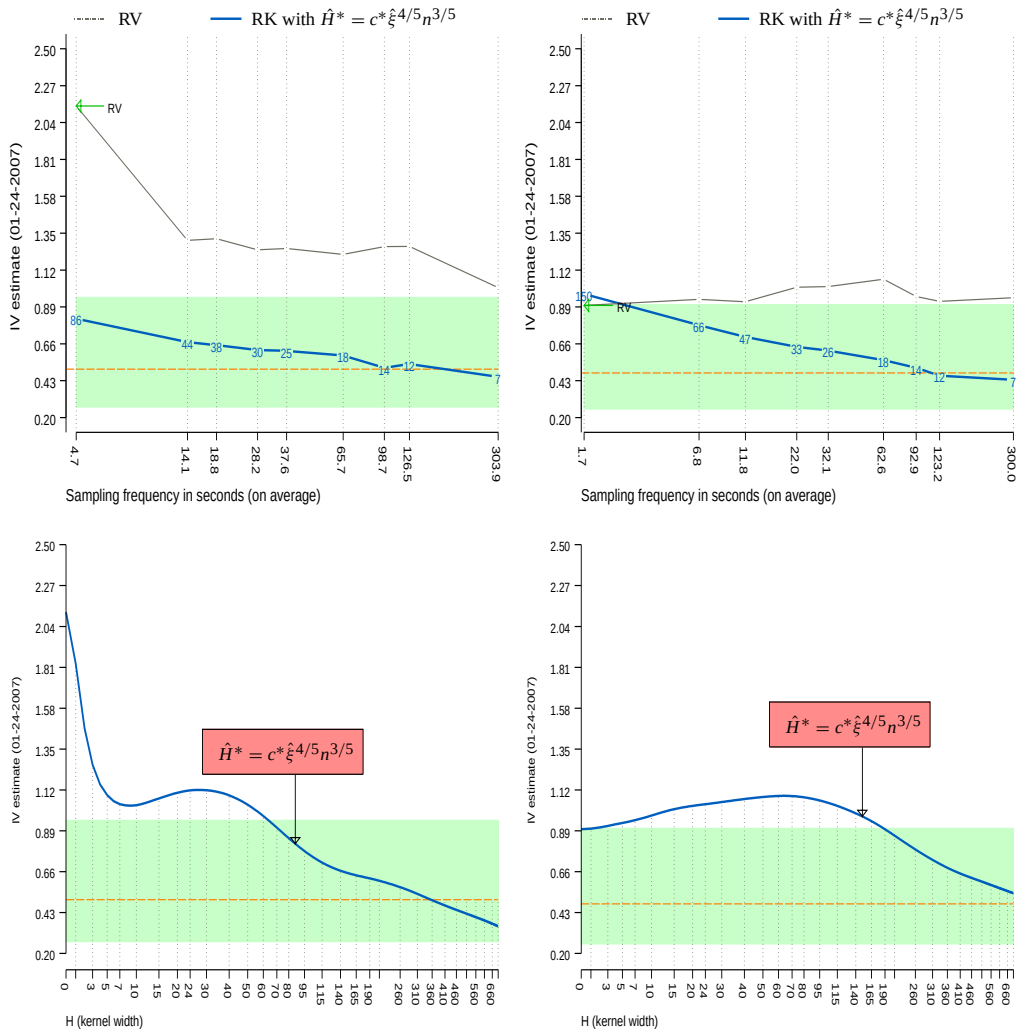

Figure 14: Signature plots for the realised kernel and realised variance for Alcoa Inc. on January 24, 2007. For details see Figure 3.

where $\varepsilon_{\tau_j} \in \mathcal{CS}$ and θ_{τ_j} is one for gradual jumps. Obviously this could induce very significant correlation between the noise and the price process. Of course not all jumps will have this characteristic. When public announcements are made, where the timing of the announcement is known a priori, then jumps tend to be absorbed immediately in the price process. In those cases $\theta_{\tau_j} = 0$. These tend to be the economically most important jumps, as they are difficult to diversify.

This line of thinking encouraged us to remove this gradual jump to replace it by a single jump. This is shown in Figure 11, while the corresponding results for the realised kernels are given in Figure 12. This seems to deliver very satisfactory results. Hence "gradual jumps" seem important in practice and challenging for this method. We do not currently have a method for

automatically detecting gradual jumps and removing them from the database.

4.6. A puzzling day: January 24, 2007

The feature we want to emphasize with this day is related to the spiky price changes. The upper panel of Figure 13 shows this jittery variation in the price, in particular towards the end of the day, where the price moves a lot within a narrow band. We believe this variation is true volatility rather than noise because the bid ask spread continues to be narrow in this period, about 2 cents most of the time.

January 24, 2007 is a day where the realised kernel is sensitive to the sampling frequency and choice of bandwidth parameters, H, as is evident from Figure 14. This may partly be attributed to pure chance, but we do not think that chance is the whole story here. Chance plays a role because the standard error of the realised kernel estimator depends on both the sampling frequency and bandwidth parameter. Rather the problem is that too large a H, or too low sampling frequency will overlook some of the volatility on this day – a problem that will be even more pronounce for the low-frequent realised variance. We will return to this issue in Figure 15.

Figure 14 also reveals a rather unusual volatility signature plot for the realised variance based on mid-quote prices. Usually the RV based on tick-by-tick returns is smaller than that based on moderate sampling frequencies, such as 20-minutes, but this is not the case here.

Figure 15 shows the prices that will be extracted at different sampling frequencies. The interesting aspect of this plots is that the realised variance, sampled at moderate and low frequencies, largely overlooks the intense volatility seen towards the end of the day.

Returns based on 20 minutes, say, will tend to be large in absolute value, during periods where the volatility is high. However, there is a chance that the price will stay within a relatively narrow band over a 20 minute period, despite the volatility being high during this period. This appears to the case toward the end of the trading day on January 24, 2007. The reason that we believe the rapid changes in the price is volatility rather than noise, is because the bid-ask spread is narrow in this period, so both bid and ask prices jointly move rapidly up and down during this period. Naturally, when prices are measured over 20 minutes intervals returns are small, yet volatility is high, the realised variance (based on 5-minute returns) will under-estimate the volatility. For the simple reason that the intraday returns do not reflect the actual volatility. This seems to be the case on this day as illustrated in the two lower panels in Figure 15. The two sparsely sampled RV cannot capture this variation in full, because the intense volatility cannot fully be unearthed by 20-minute intraday returns.

Because the realised kernel can be applied to tick-by-tick returns, it does not suffer from this problem to the same extent. Utilizing tick-by-tick data gives the realised kernel a microscopic ability to detect and measure volatility, that would otherwise be hidden at lower frequencies (due to chance). The "strength" of this "microscope" is controlled by the bandwidth parameter, and the realised kernel gradually looses it ability to detect volatility at the local level as H is increased. However, H must be chosen sufficiently large to alleviate the problems cause by noise.

On January 24, 2007, we believe that $K(X) \simeq 0.90$ is a better estimate of volatility than the subsampled realised variance based on 20 minute returns, whose point estimate is nearly half that of our preferred estimator.

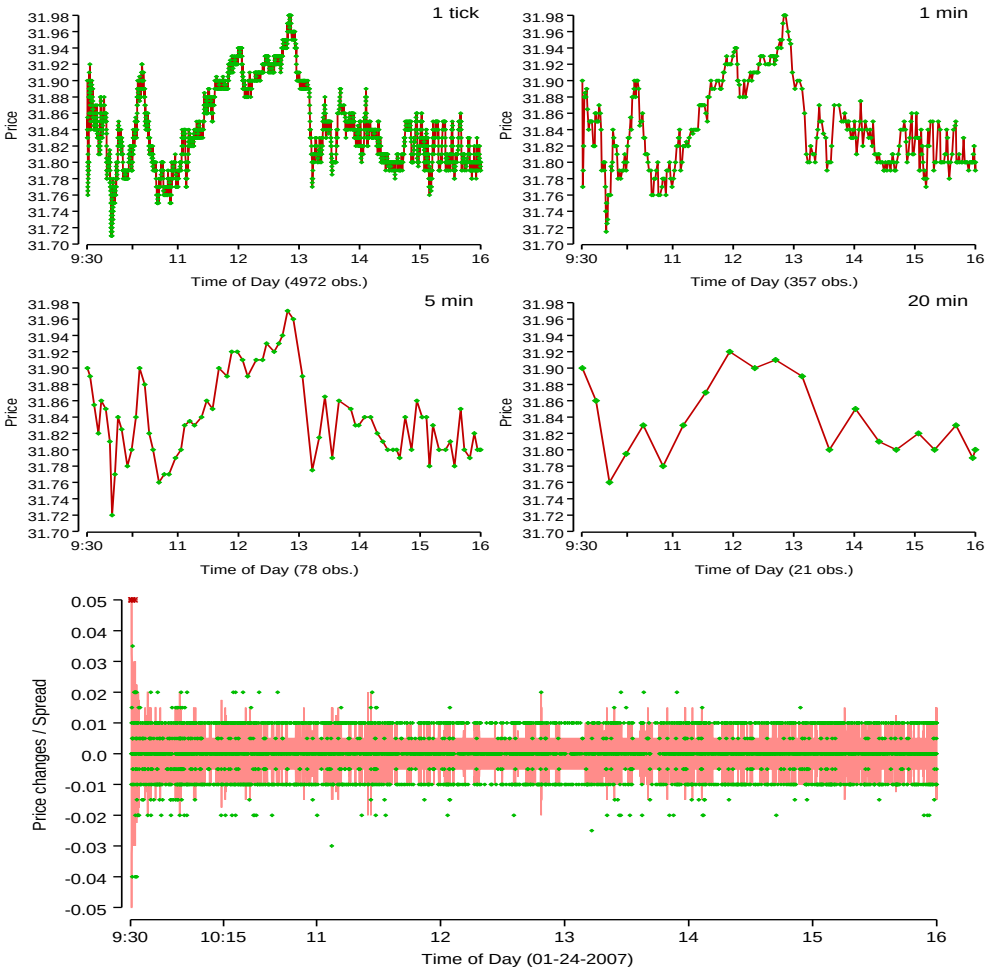

Figure 15: The four upper panels show the transaction prices for Alcoa Inc. on January 24, 2007 are extracted at different sampling frequencies. The lower panel presents the tick-by-tick return on transaction data (dots), and the spread as it varied throughout the day (vertical lines). An interesting aspect of these plots is that the realised variance, based on low sampling frequencies, misses the intense volatility be the end of the day.

5. CONCLUSIONS

In this paper we have tried to be precise about how to implement our preferred realised kernel on a wide range of data. Based on a non-negative form of the realised kernel, which uses a Parzen weight function, we implement it using an averaging of the data at the end conditions. The realised kernel is sensitive to its bandwidth choice. We detail how to choose this in practice.

A key feature of estimating volatility in the presence of noise is data cleaning. There is very little discussion of this in the literature and so we provide quite a sustained discussion of the

interaction between cleaning and the properties of realised kernels. This is important in practice, for in some application areas it is hard to extensively clean the data (e.g. quote data may not be available), while in other areas (such as when one has available trades and quotes from the TAQ database) extensive and rather accurate cleaning is possible.

We provide an analysis of the properties of the realised kernel applied simultaneously to trade and quote data. We would expect the estimation of $[Y]$ to deliver similar answers and they do, indicating the strength of these methods.

Finally, we identify an unsolved problem for realised kernels when they applied over relatively short periods. We call these "challenging days." They are characterized by lengthy strong trends being present in the data, which are not compatible with standard models of market microstructure noise.

ACKNOWLEDGEMENTS

This paper was presented at the Econometrics Journal invited session on Financial Econometrics at the Royal Economic Society's Annual Meeting. We thank Richard Smith for his invitation to give it. We also thanks Marius Ooms and Kevin Sheppard for helpful comments. The second and fourth author are also affiliated with CREATES, a research centre funded by the Danish National Research Foundation.

REFERENCES

Admati, A. R. and P. Pfleiderer (1988a). Selling and trading on information in financial markets. *American Economic Review 78*(2), 96–103.

Admati, A. R. and P. Pfleiderer (1988b). A theory of intra day patterns: Volume and price variability. *Review of Financial Studies 1*(1), 3–40.

Admati, A. R. and P. Pfleiderer (1989). Divide and conquer: A theory of intraday and day-of-the-week mean effects. *Review of Financial Studies 2*(2), 189–223.

Andersen, T. G., T. Bollerslev, and F. X. Diebold (2008). Parametric and nonparametric measurement of volatility. In Y. Aït-Sahalia and L. P. Hansen (Eds.), *Handbook of Financial Econometrics*. Amsterdam: North Holland. Forthcoming.

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

Andersen, T. G., T. Bollerslev, F. X. Diebold, and P. Labys (2001). The distribution of exchange rate volatility. *Journal of the American Statistical Association 96*, 42–55. Correction published in 2003, volume 98, page 501.

Andersen, T. G., T. Bollerslev, and N. Meddahi (2006). Market microstructure noise and realized volatility forecasting. Unpublished paper: Department of Economics, Duke University.

Andrews, D. W. K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica 59*, 817–858.

Bandi, F. M. and J. R. Russell (2008). Microstructure noise, realized variance, and optimal sampling. *Review of Economic Studies 75*, 339–369.

Barndorff-Nielsen, O. E., P. R. Hansen, A. Lunde, and N. Shephard (2006). Subsampling realised kernels. Unpublished paper: Nuffield College, Oxford.

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

Barndorff-Nielsen, O. E., P. R. Hansen, A. Lunde, and N. Shephard (2008b). Multivariate realised kernels: Consistent positive semi-definite estimators of the covariation of equity prices with noise and non-synchronous trading. Unpublished paper: Oxford-Man Institute of Quantitative Finance, University of Oxford.

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

Barndorff-Nielsen, O. E. and N. Shephard (2007). Variation, jumps and high frequency data in financial econometrics. In R. Blundell, T. Persson, and W. K. Newey (Eds.), *Advances in Economics and Econometrics. Theory and Applications, Ninth World Congress*, Econometric Society Monographs, pp. 328–372. Cambridge University Press.

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

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

Diebold, F. and G. Strasser (2007). On the correlation structure of microstructure noise in theory. and practice. Unpublished paper: University of Pennsylvania.

Falkenberry, T. N. (2001). High frequency data filtering. Technical report, Tick Data.

Hansen, P. R. and A. Lunde (2006). Realized variance and market microstructure noise (with discussion). *Journal of Business and Economic Statistics 24*, 127–218.

Hasbrouck, J. (2007). *Empirical Market Microstructure: Economic and Statistical Perspectives on the Dynamics of Trade in Securities Markets*. Oxford University Press.

Jacod, J., Y. Li, P. A. Mykland, M. Podolskij, and M. Vetter (2007). Microstructure noise in the continuous case: the pre-averaging approach. Unpublished paper: Department of Statistics, University of Chicago.

Kalnina, I. and O. Linton (2006). Estimating quadratic variation consistently in the presence of correlated measurement error. Unpublished paper: Department of Economics, LSE.

Kyle, A. S. (1985, November). Continuous auctions and insider trading. *Econometrica 53*(3), 1315–1335.

Li, Y. and P. Mykland (2007). Are volatility estimators robust to modelling assumptions? *Bernoulli 13*, 601–622.

Newey, W. K. and K. D. West (1987). A simple positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica 55*, 703–708.

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

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

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

Zhang, L., P. A. Mykland, and Y. Aït-Sahalia (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 and Economic Statistics 14*, 45–52.