Realized Covariance and Scrambling

Kevin Sheppard*
Department of Economics
University of Oxford

March 9, 2006

Abstract

Computing realized covariance from 5 minute returns, even for the most liquid stocks, produces estimates with an unmistakable bias towards zero. Using returns sampled more frequently than 20 minutes can lead to significant underestimation of covariance. Moreover, in some special cases, sampling even twice per day is too frequent. This paper revisits some stylized facts about covariance computed using high-frequency data and examines two models for their ability to explain common empirical regularities. Standard models where prices are contaminated with stochastically independent noise are unable to explain the behavior of realized covariance as the sampling frequency increases. Conditions for unbiasedness and consistency of realized covariance when returns are possibly scrambled are derived. The concept of scrambling is introduced to motivate an a general family of alternative specifications based on random censoring of returns. This class nests previously suggested corrections for realized covariance and points to a direction for creating unbiased and consistent estimators of realized covariance.

JEL Classification Codes: C32, G0, G1
Keywords: Realized Correlation, Realized Covariance, Realized Variance, Asynchronous Trading, Market Microstructure, Epps Effect, Scrambling

*Mailing address: Department of Economics, University of Oxford, Manor Road Building, Manor Road, Oxford OX1 3UQ, UK. email: kevin.sheppard@economics.ox.ac.uk. The author would like to thank Neil Shephard, Anders Rahbeck and Pauline Kennedy and seminar participants at the University of Oxford and CASS School of Business for discussion. All efforts have been made to ensure the paper is mistake free and any remaining errors are the sole responsibility of the author.

Widespread availability of asset price data at high frequencies and recent econometric advances have revolutionized the measurement of covariance. Realized measures exploit all available information to construct seemingly accurate model-free estimates of the covariance with clear advantages: they are valid for most arbitrage free price processes, are trivial to compute, and avoid needless and problematic assumptions about the dynamics of covariance. Realized measures are both intuitive (Merton 1980) and rigorous under weak assumptions on the price process (Andersen, Bollerslev, Diebold & Labys (2003) and Barndorff-Nielsen & Shephard (2004)).

A key insight of these results is that prices should be sampled as frequently as possible to maximize precision of realized measures. However, frequent sampling is only justified if prices are error free. Observed prices are contaminated by market microstructure noise through bid-ask bounce, price discretization, market closure, trading halts and asynchronous trading. Recent research into the effects of market microstructure noise have focused realized variance. Proposed adjustments to realized variance include filtering (Ebens (1999), Andersen, Bollerslev, Diebold & Ebens (2001) and Bandi & Russell (2005*a*)), subsampling (Zhang, Mykland & Aït-Sahalia 2004), correcting for overnight price changes (Hansen & Lunde 2004*b*), and using kernel estimators (Hansen & Lunde (2004*a*) and Barndorff-Nielsen, Hansen, Lunde & Shephard (2004)) to control or remove the bias. These papers have focused on models where observed prices are contaminated with independent (from the price process) additive noise. The effect of this noise is clear: sampling to frequently leads to a substantial positive bias in realized variance.

In contrast to the behavior of realized variance, realized covariances show a clear bias towards zero as the sampling frequency increases.[1] Epps (1979) originally documented the bias toward zero using returns on the big four automobile manufacturers, American Motors, Chrysler, Ford, and General Motors in 1971 and 1972. He documented monotonic increases in the correlation as the sampling frequency decreased from 10 minutes to 2 days, a phenomena subsequently known as the *Epps effect* in the market microstructure literature.

Differences between the scaling behavior of realized variance and realized covariance are primarily driven by price synchronization. Consider a single asset measured over two periods where returns in each period are independent. At the end of the first period, the price of the asset can either be updated to reflect the first period return or it can remain at its initial value. In either case, if the price is updated at the end of the first period or if only the final cumulative return is observed, the variance of returns can be estimated using the two returns. Suppose there is a second asset with the same properties. There are now four possible patterns of observation: both prices are updated in both periods, only one price is updated during the first period, or neither price is updated during the first period. When only one price is updated during the first period, there will be a clear bias in the computed covariance. The only way to be certain that the covariance is

[1] Throughout the paper we will refer to second moments across two assets as covariance and own second moments as variance.

unbiased is to *only* use the two-period return.

These problems are typically found in asset prices due to out-of-sync observations throughout the trading day. However, other asymmetric periods of inactivity, such as the settlement of the opening auction, delayed opening, late closing or trading halt are uniquely problematic for measuring covariance. For instance, trading on the NYSE officially begins at 9:30 and ends at 4:00 pm, yet the median difference between the first tick of first DJIA component to open and the first tick of the last DJIA component to open is over 10 minutes using data from 1993 to 1998. Any sampling scheme which samples more frequently than 10 minutes will have at least one out-of-sync return at the open on a typical day. However, when prices are considered in isolation, whether the first tick occurs at 9:30, 9:40 or 10:00, or the stock doesn't trade for an entire day, the first non-zero return will contain the cumulative effect of the variance during the closed period.

To explain these findings, this paper first examines the implications of a simple model where a vector of prices is additively contaminated by stochastically independent noise. Under this data generating process, realized covariance is shown to be an unbiased but inconsistent estimator of the covariance as the sampling frequency increases. The intuition behind this results is simple: independent (from everything) noise has no effect on average, but, as the sampling frequency increases, the amount of noise increases without bound, affecting the variability of the estimator. There will be an optimal sampling window trading noise induced variance at very high-frequency against having too few observations at lower frequencies. In a more general framework, Bandi & Russell (2005*b*) have considered this problem to derive this frequency. However, the model cannot generate commonly found bias in realized covariance and using a optimal sampling frequency based on a unrealistic data generating process is of questionable value.

However, many of the empirical regularities found in the Dow Jones 30 can be replicated using a delayed news model. A special case of this model has recently been explored in the context of fixed windows for estimating the variance of a stock across multiple exchanges (Martens 2004). This paper introduces the concept of *scrambling* to describe the link between the price generating process and the sampling process. Scrambling is nearly self-descriptive; prices are scrambled if the order of observation is only weakly related to the order of price generation. This allows for standard scenarios where prices are simply observed out-of-sync due to non-trading and also includes processes where observed increments are not synchronized even when both trade. Two other concepts, *ordered* prices and *descrambled* prices are introduced to clarify standard cases of perfectly synchronized returns and ex-post synchronized returns, respectively.

We examine the properties of realized covariance as the time between samples decreases and find that realized covariances can converge to zero even for assets with perfectly correlated returns. Additionally, this model produces cross-correlations, leaves realized variance estimates essentially unaltered, and does not produce autocorrelated returns. We consider two asymptotic experiments,

one where we let the number of samples diverge holding the probability of price updates remain constant and one where we let the probability of new prices decrease as the number of samples diverges. In the first case, realized variance is asymptotically unbiased while realized covariance remains biased but has a nonzero limit as long as the quadratic covariation is nonzero. In the second case, realized variance remains unbiased and converged to zero irrespective of the quadratic covariation of the price process.

Returning to the data, a simple independent transaction model provides a fairly good approximation to the observed returns. However, the covariance of some assets, those with the highest daily correlation typically found in the same sector, exhibits scaling issues beyond those implied by the model.

Section 2 describes the data used in this study and presents a set of empirical regularities. Section 3 shows that pure noise contamination cannot explain the bias in found in realized covariance. Section 4 describes an no-news model and examines its ability to explain these findings. Section 5 considers unbiased and consistent estimators and revisits the data in light of these finding and section 6 concludes.

2 Data and Empirical Regularities

The data used in this paper consist of prices of the Dow Jones Industrial Average constituents over the period from January 4, 1993 to May 29, 1998, a total of 1365 trading days. Prices were extracted from mid-quotes and were corrected for dividends and splits. All 30 stocks were listed on the NYSE and only quotes from this exchange were used. Prices were further filtered from the official opening quote until 16:10 and only include valid entries. Additionally, obvious outliers were removed.[2] Price grids were constructed using last price interpolation. One and two-day returns were computed using closing prices.

A second data set, consisting of the remaining constituents of Epps (1979), Chrysler (later Daimler Chrysler), Ford and General Motors is used to illustrate some interesting aspects of realized correlation measurement. Returns on these three assets were available from January 4, 1993 until December 31, 2001 (2262 trading days) and were constructed in the same manner as the DJIA stocks.[3]

Table 1 contains ticker symbols, firm names, and quote frequency summary statistics for the 30 Dow Jones Industrial Average stocks. The average number of quotes per day ranged from a low of 250 (UK) to a high of 1077 (MO). However, many of these quotes only alter depth and do not

[2]Quotes from the TAQ database often contain 2 types of errors: 0's and moving a decimal place (i.e. 32.25 becomes 3.22 or 322.50. Removal of these were the only corrections made to the data.

[3]There were 6 days during the sample that at least one of these three did not trade. As the focus of this paper is on the measurement rather than the modeling of the dynamics of the correlation, these have been omitted from all 3 series.

contain new prices for either the bid or ask. The table also contains the number of *informative* quotes per day which only include those where either the bid or the ask price (or both) changed from the previous quote. Approximately 1 in 3 quotes are informative, although the ratio varies from 20% to nearly half. The table also contains the percentage of return windows which contain informative quotes when the window length is 1, 5, 10 or 30 minutes. On average, one-quarter of 1-minute windows contain informative quotes. By five minutes, over half of the windows contain informative quotes while 73% of the 10-minute windows contain informative quotes. When using 30 minute windows, over 85% contain informative quotes. However, the average is somewhat misleading: bias in covariance is driven by the least frequently observed price. For instance, when sampling every 30 minutes, one-quarter of all returns will be zero for Walmart and Union Carbide. If price revisions were independent, then roughly 1 in 3 of the windows with a quote revision in one will correspond to no new information for the other (and a 0 return). In the actual price data for WMT and UK, 29% of the 30-minute windows where one had an informative quote was not matched by a revision in the other.

Realized covariance between assets i and j on day t, based on m samples per day is defined as

$$RC_{ijt}^{(m)} = \sum_{n=1}^{m} r_{itn} r_{jtn} = \sum_{n=1}^{m} (p_{itn} - p_{itn-1})(p_{jtn} - p_{jtn-1}) \tag{1}$$

where p_{it0} and p_{jt0} are defined to be closing prices on the previous day. Realized covariance was computed using 1 (m=400), 5 (80), 10 (40), and 30 (14) minute returns while daily covariance was computed using 1 and 2 day close-to-close returns. To facilitate comparisons across different sampling frequencies, *pseudo*-realized correlations are employed. The term *pseudo* indicates that while the covariances are constructed using a variable window length, variances used to standardized the realized covariances were always computed from 5-minute returns; pseudo realized correlations are approximately scale free and changes in the pseudo-correlations are uniquely attributable to changes in realized covariance.

Table 2 contains scaling information for both the average correlation, constructed using the average realized covariance divided by the square-root of the product of the average 5-minute realized variances, and the maximum correlation of each of the 30 stocks. Realized covariance computed from one-minute returns show substantial changes when compared to daily correlations, differing on average (across all 435 correlations) by .11 (50%). By five minutes, the average downward bias has decreased to 15% and correlations computed from 10-minutes are essentially unbiased. However, average correlations are misleading as the largest correlations were increasing throughout the range of sampling frequencies in most assets.

Three asset pairs, CHV-XON, GM-TRV and JNJ-MRK had daily correlation in excess of 50%. Correlations measured using ten minute returns among these three were still biased downward

30% when compared to daily correlations. Two pairs are in the same industry while GM and Travelers share common exposure through GM's large financial arm GMAC. We will examine the issue of closely related firms in detail when we consider the scaling behavior of the automobile manufacturers' covariance. Figure 1 contains a plot of the pseudo-correlations against the log of time. Realized covariances were computed on a grid of 15 seconds from 15 seconds to 3.25 hours (half-day). The pictured correlations are quantiles of distribution of realized correlations computed at the 0%, 1%, 25%, the median, 75%, 99% and 100% quantiles. All but the upper two of these quantiles appear to have flattened by 20 minutes. However, the top two quantiles, and particularly the max (XON-CHV for the lowest frequency returns), are still increasing over the entire range.

m-sample realized variance is computed in an analogous manner

$$RV_{it}^{(m)} = \sum_{n=1}^{m} r_{in}^2 = \sum_{n=1}^{m} (p_{itn} - p_{itn-1})^2. \tag{2}$$

However, in stark contrast to realized covariances, realized variances evidence no systematic scaling bias. Table 3 contains the (annualized) volatility computed using 1, 5, 10 and 30 minute windows as well as 1 and 2 day returns. All series show little systematic bias as the sampling frequency changes and differ by less than 15% across the various windows. Figure 2 contains the quantiles of the thirty realized variance series. Each series was constructed using returns sampled from 15 seconds to 3.25 hours (1/2 day) and were standardized by the 5-minute realized variance (m=80). Compared to 5-minute RV, the variances appear to be cross-sectionally median unbiased and are symmetric in their dispersion, although there is possibly a slight decrease for the highest sampling frequencies. These results indicate there is a fundamental difference in the scaling behavior of realized variance and realized covariance.

Revisiting the behavior of realized covariance among same industry assets, we also examine the returns of the big three automobile manufacturers. These three have numerous sources of shared risk: changes in the macroeconomic climate, labor contracting, interest rates, etc. Figure 3 contains the realized variance and pseudo-realized correlation signature plots using returns computed from 1 seconds to 3.25 hours. The top panel, containing the correlation signature plot, is striking. Measured covariances are monotonically increasing from 30 seconds until the end of the range. As was the case with the DJIA stocks, the volatility signature plot is relatively flat, although GM shows evidences some downward bias as the sampling frequency increases. Table 4 contains the quote, variance and correlation summary statistic for these three stocks. They are more active than a typical DJIA stock, although much of this is attributable to the longer sample which covers the period when penny-tick sizes were introduced, corresponding to a marked increase in quote activity. Further, while over 90% of the 30 minute windows contained informative quotes, all series demonstrated significant bias even when sampled this infrequently.

To understand the nature of the bias of realized covariance (and realized variance) estimators, it is simple to decompose the difference the m sample realized covariance and the covariance computed suing daily returns. Let r_{it} denote the daily return on the asset i. Using m uniformly spaced samples, $r_{it} = \sum_{n=1}^{m} r_{itn}$, and the cross-product of returns is

$$r_{it}r_{jt} = \sum_{n=1}^{m} r_{itn} \sum_{o=1}^{m} r_{jto} = \sum_{n=1}^{m} r_{itn}r_{jtn} + \sum_{o=1}^{m} \sum_{q=1,q\neq o}^{m} r_{ito}r_{itq} = RC_{ij}^{(m)} + \sum_{o=1}^{m} \sum_{q=1,q\neq o}^{m} r_{ito}r_{itq} \tag{3}$$

Clearly the realized covariance is embedded in the cross product of daily returns. However, the cross product also include m^2-m terms which capture the relationship between the leads and lags of r_{itn} on the high frequency returns of r_{jt}. If the covariance measured using daily returns is different than that measured using m returns, the difference *must* be captures through these leads and lags. Figure 6 contains the cross-correlations for 4 pairs of assets, three from the DJIA, UK-WMT, BA-GE, and XON-CHV and F-GM from the auto manufacturers. The m-sample cross-correlation between asset i and lags of asset j at lag n was computed using 1-minute returns:

$$\rho_n^{i|j(m)} = \frac{\sum_{q=n+1}^{mT} r_{iq}r_{jq-n}}{\sqrt{\sum_{q=n+1}^{mT} r_{iq}^2 \sum_{q=n+1}^{mT} r_{jq-n}^2}}. \tag{4}$$

All cross-correlograms have the same behavior for first few lags, although the magnitude of the effect varies.[4] After 5 to 15 minutes, the cross-correlations typically become insignificant, although they are positive too often to be random. However, for XON-CHV and F-GM, the cross-correlations are large and almost always positive. Moreover, there are asymmetries in the relationships. CHV has more significant positive relationships to lagged XON than the opposite, while GM leads F more than F leads GM. While we do not present auto-correlograms of any assets, they are remarkably flat. This can be inferred by examining the scaling of realized variances where little change was observed.

Five traits are common among the 33 assets studied in this paper:

- Bias in realized covariance constructed from high frequency returns
- Little or no bias in realized variance constructed from high frequency returns [5]
- Numerous positive cross-correlations with other assets when sampled at higher frequencies

[4]The correlation between a lagged return of asset i and a contemporaneous return of asset j is sometimes referred to as a cross-auto-correlation. This name is ambiguous and misleading, and we will use cross-correlation to refer to this quantity in the same sense as an auto-correlation refers to the correlation between lags and contemporaneous returns of the same asset.

[5]Because mid-quote returns were used rather than trades, and are bid-ask bounce free, realized variance do not exhibit positive bias documented in Hansen & Lunde (2004*c*), *inter alia*.

- No autocorrelation
- Intra-industry pairs exhibit the strongest bias with increasing correlation across a day or more

Two different noise models will be examined for their ability to capture these five regularities. The first, an additive noise model, has been successful in understanding the bias in realized variance computed from frequently sampled trades. The second, a no-news model specified through a multiplicative error, considers the case where high frequency returns are censored and aggregated into future returns.

3 Additive Noise

Pure noise models, where observed prices are contaminated by stochastically independent errors, have been successful in understanding the behavior of realized variance when computed using frequently sampled returns (Hansen & Lunde (2004*c*), Zhang et al. (2004) and Barndorff-Nielsen et al. (2004)). In this framework, realized variance converges to the variance of the error times the number of samples as the number of samples grows large.

The price process is assumed to be mean zero random walk with random covariance.

Assumption 1 (PP) *A K by 1 vector price process,*

$$\mathbf{p}_t^* = \int_0^t \mathbf{\Omega}_s d\mathbf{W}_s$$

where $\mathbf{\Omega}_s\mathbf{\Omega}_s = \mathbf{\Sigma}_s$, $\mathbf{W}_t$ is a K dimension Brownian motion and $\mathbf{\Sigma}_s$ is uniformly positive definite, independent of $\mathbf{W}$ and Lipschitz element-by-element (a.s.).

Without loss of generality, we restrict out attention to the interval $t \in [0, 1]$. Observed prices are assumed to be contaminated with vector noise process which is stochastically independent of the price process and uncorrelated.

Assumption 2 (AN) *Observed prices, $\mathbf{p}_t$ are measured with an additive error, $\mathbf{p}_t = \mathbf{p}_t^* + \mathbf{u}_t$. The noise process $\mathbf{u}$ satisfies the following properties:*

i. $E[\mathbf{u}] = \mathbf{0}$
ii. $\mathbf{u} \perp\!\!\!\perp \mathbf{p}$
iii. $u_i \perp\!\!\!\perp u_j$, $i \neq j$, $i, j = \{1, 2, \dots K\}$
iv. $u_{is} \perp\!\!\!\perp u_{it}$, $i \neq j$, $i, j = \{1, 2, \dots K\}$ $\forall s \neq t$
v. $E[u_{it}^2] = \upsilon_i < \infty$

Assumptions 1 and 2, when reduced to a scalar process, are equivalent to those of Hansen & Lunde (2004*c*). Prices are assumed to be sampled uniformly over $[0,1]$ to generate m returns. We are specifically interested in the behavior of the realized covariance estimator between two elements of $\mathbf{p}$, i and j. The m-sample realized covariance is defined to be

$$RC_{ij}^{(m)} = \sum_{n=1}^{m} r_{in} r_{jn} \tag{5}$$

where $r_{in} = p_{in/m} - p_{i(n-1)/m}$ is the return on the interval $[\frac{n-1}{m}, \frac{n}{m}]$. Defining $\epsilon_{in} = u_{in/m} - u_{i(n-1)/m}$, realized covariance can be rewritten in terms of the true return process and the errors

$$RC_{ij}^{(m)} = \sum_{n=1}^{m} (r_{in}^* + e_{in})(r_{jn}^* + e_{jn}) = \sum_{n=1}^{m} r_{in}^* r^* jn + \sum_{n=1}^{m} e_{in} r_{jn}^* + \sum_{n=1}^{m} e_{jn} r^* in + \sum_{n=1}^{m} e_{in} e_{jn}. \tag{6}$$

The first term is the standard realized covariance estimator, the sum of the product of high-frequency returns, while the remaining terms have unknown effects. Proposition 1 analyzed the behavior of the realized covariance estimator under a pure noise process.

Proposition 1 *Under assumptions PP and AN and conditioning on* $\{\mathbf{\Sigma}_t\}$*,*

$$E[RC_{ij}^{(m)}] = \int_0^1 \sigma_{ijs} ds$$

and

$$Var[RC_{ij}^{(m)}] = \frac{\int_0^1 \left\{ \sigma_{iis}\sigma_{jjs} + \sigma_{ijs}^2 \right\} ds}{m} + 2\omega_j^2 \int_0^1 \sigma_{iis} ds + 2\omega_i^2 \int_0^1 \sigma_{jjs} ds + 6m\omega_i^2 \omega_j^2$$

Thus $RC_{ij}^{(m)}$ is an unbiased estimator on the integrated covariance but has a variance that is increasing in the number of samples. In the case that $\omega_i = \omega_j = 0$, the reduces to the standard case (Barndorff-Nielsen & Shephard 2004). This results is substantively different that the case for realized variance where the estimator is divergent. If prices were contaminated by an additive noise process, we would expect realized covariances to become increasingly unstable when computed using prices sampled frequently. However, figure 1 paints a different picture. Using the highest sampling frequencies, realized covariances become less disperse with obvious bias.

A pure additive noise model is also incapable of matching the other important empirical regularity; cross-correlations will be zero.

Proposition 2 *Under assumptions PP and AN,*

i. $E[r_{in}r_{jn+h}]/(E[r_{in}^2]E[r_{jn}^2])^{(1/2)} = 0, \forall h \neq 0$

The pure noise model also cannot generate any of the pattern evident in the data. However, if the assumption of stochastically independent noise was relaxed, this model may be able to capture some or all of the commonly observed properties. Examining (6), there are two opportunities for bias to be generated: in the covariance of the error and the return or in the covariance between the errors. Generating bias toward zero using only the covariance between the errors would require a negative covariance that depends on window length. However, this would bias covariance for all sampling frequencies and isn't supported by the data. Introducing bias through the covariance of the latent returns and the error terms would require an essentially degenerate behavior and is not logically consistent when more than two stocks were considered.

4 Multiplicative Noise

As evidenced in the DJIA stocks, many windows contain no new price information when prices are frequently sampled and it is rarer still that two assets have simultaneous price updates. Frictions generated by a lack of new price information behave very differently when considered cross-sectionally. Lo & MacKinley (1990) have considered the case where stocks trade with different intensities and the effects on the efficient markets hypothesis. Under their asynchronous trading model, in each period, a random shock determines whether prices are updated to reflect the efficient price or if they remain at the previous closing price.

When prices take previous values and sampled prices do not correspond to the same point in time, prices are said to be scrambled. Let $(t_{im})_{m\geq 0}$ be a set of stopping times that correspond to the observation nodes of p_{it}. These do not have to be regularly spaced or predictable. Let $(\tau_{in})_{n\geq 0}$ be a simple point process associated with asset i referred to as the measurement nodes.[6]

Definition 1 (Scrambling) *Prices are scrambled with respect to a set of observation nodes if there exists m such that $\hat{\tau}_i \neq \tilde{\tau}_j$ for some $i, j \in \{1, \ldots, K\}$ where $\hat{\tau}_i = \max\{\tau_{in} : \tau_{in} \leq \tau_m\}$ and $\tilde{\tau}_j = \max\{\tau_{jn} : \tau_{jn} \leq \tau_m\}$. Returns are scrambled if constructed from scrambled prices.*

Scrambling implies a few properties of the observed returns:

- The price of at least one asset at some point in time must be a previous price of that asset.
- The price of another asset sampled at the same point must have correspond to a price at a different point in time.

[6] A simple point process is a point process with weakly increasing values. If a simple point process if adapted to the filtration, then it is a set of stopping times. While there is no restriction that the measurement nodes cannot be stopping times (which allows them to be conditioned on), it is important to consider cases where the observed price is never know to be the efficient price.

Scrambling does not require the sampling times to correspond to the synchronization times. Scrambled returns can include last price interpolated returns and can also include trades or quotes occurring at a then stale price. This corresponds to an important empirical finding where the length of the cross-correlation is much larger than a pure synchronization story.
For example, suppose asset i was very liquid and the price observed at any time was the efficient price while asset j was an illiquid asset that typically requires 10 minutes for indicative prices to reflect the efficient price. Sampling from these prices would generate scrambled prices as the price at any observation node would correspond to the price at that point in time for asset i and the 10-minute stale price for asset j. Random scrambling, where either asset leads at any observation node is another possibility.

Conversely, the definition of ordered returns is

Definition 2 (Ordered) *Prices are ordered if* $\forall m$, $\hat{\tau}_i = \tilde{\tau}_j \; \forall i, j \in \{1, \ldots, K\}$ *where* $\hat{\tau}_i = \max\{\tau_{in} : \tau_{in} \leq \tau_m\}$ *and* $\tilde{\tau}_j = \max\{\tau_{jn} : \tau_{jn} \leq \tau_m\}$. *Returns are ordered if constructed from ordered prices.*

Ordering implies a few properties of the observed returns:

- The standard setup of sampling without error at any point in time corresponds to ordered prices and produces ordered returns.
- Ordered prices can include stale prices as long as all prices were synchronous.
- Prices can still be ordered even if the price process (occasionally) generates out-of-sync prices, because ordering is a function of both the price generation process and the sampling scheme.

In the standard setup (Andersen et al. (2003) and Barndorff-Nielsen & Shephard (2004)), returns are always assumed to be ordered.

Rather than require synchronization with the current efficient price, one could imagine a scenario where the current price reflects an efficient price some time between the last efficient price and the current efficient price, inclusive. Consider a single asset and suppose that initial price was known with certainty at the beginning of the sample period ($p_0 = p_0^*$). Thus, at the first sampling node, $p_{1/m} = p_{\tau_1}, \exists \tau_1 \in [0, 1/m]$. At the second sampling node, $p_{2/m} = p_{\tau_2}, \exists \tau_2 \in [\tau_1, 2/m]$, and so forth. The set $\{1/m, 2/m, \ldots, 1\}$ are known as the observation nodes while the set $\tau_1, \tau_2, \ldots \tau_m$ is known as the measurement nodes. Assuming that the observation and measurement nodes correspond to the same points in $[0, 1]$ (but τ_j is not necessarily equal to j/m), the n^{th} observed return, r_n, corresponding to $p_{n/m} - p_{(n-1)/m}$ can be expressed in terms of the efficient returns, $r_q^* = p_{q/m}^* - p_{(q-1)/m}^*$

$$r_n = p_{\tau_n} - p_{\tau_{n-1}} \tag{7}$$

$$= \sum_{q=1}^{n} x_{qn} \left(p^*_{\frac{q}{m}} - p^*_{\frac{q-1}{m}} \right) \tag{8}$$

$$= \sum_{q=1}^{n} x_{qn} r^*_q \tag{9}$$

where x_{qn} are variables (possibly random) which take the value 1 if $\frac{q}{m} \in (\tau_{n-1}, \tau_n]$. Observed returns capture all of the returns between the most recent measurement node and the previous measurement node. However, unlike other models, price changes do not necessarily reflect the current efficient price. If two nodes are the same ($\tau_{n-1} = \tau_n$), the observed return will be 0.

The x_{qn} variables have some useful properties which will be exploited in examining the properties of realized estimators when returns may be scrambled. Specifically $x_{qn}x_{on} = 0$ for any $q \neq o$. Intuitively, since $\{\tau\}$ is an increasing sequence, a return can only be observed once (or possibly not at all). Thus, if $x_{qn} = 1$, so that the efficient return r^*_q was observed in r_n, it cannot be observed in any other return. If observed returns are related to the latent prices in this manner, the m by 1 vector of observed returns can be expressed compactly in terms of the latent returns

$$\mathbf{r}^{(m)} = \mathbf{r}^{*(m)}\mathbf{X}^{(m)} \tag{10}$$

where $\mathbf{r}^{(m)} = [r_1 \ \dots \ r_m]$ and the matrix $\mathbf{X}^{(m)}$ is shorthand for

$$\mathbf{X}^{(m)} = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1m-1} & x_{1m} \\ 0 & x_{22} & x_{23} & \dots & x_{2m-1} & x_{2m} \\ 0 & 0 & x_{33} & \dots & x_{3m-1} & x_{3m} \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 & x_{mm} \end{bmatrix}. \tag{11}$$

This formulation for observed returns is generic and is applicable as long as the present prices reflect some previous or contemporaneous efficient price. For instance, in the standard setups (Andersen et al. (2003) and Barndorff-Nielsen & Shephard (2004)), $\mathbf{X}^{(m)} = \mathbf{I}_m$, the identity matrix and every measured price corresponds to the efficient price at that interval.

Using the above expression for observed returns, the m-sample (assuming regular spaces) realized variance can be expressed

$$RV^{(m)} = \mathbf{r}^{*(m)}\mathbf{X}^{(m)}\mathbf{X}^{(m)\prime}\mathbf{r}^{*(m)\prime} \tag{12}$$

Similarly, defining $\mathbf{r}_i^{(m)}$ to be the observed returns from the i$^{\text{th}}$ asset, with $\mathbf{X}_i^{(m)}$ defined accordingly, the m-sample realized covariance between assets i and j can be expressed

$$RC_{ij}^{(m)} = \mathbf{r}_i^{*(m)} \mathbf{X}_i^{(m)} \mathbf{X}_j^{(m)\prime} \mathbf{r}_j^{*(m)\prime} \tag{13}$$

Again, if both $\mathbf{X}$ matrices are the identity matrix, this expression collapses to the usual realized covariance estimator, $RC^{(m)} = \sum_{n=1}^{m} r_{in}^* r_{jn}^*$ computed from the efficient prices. With only weak assumption on the structure of the $\mathbf{X}$ matrices, it is possible to derive some useful properties of realized estimators.

Assumption 3 (DX) $\mathbf{X}$*, an* m *by* m *deterministic matrix, satisfies*

i. $x_{kl} = 1$ *or* $x_{kl} = 0$

ii. $\sum_{k=1}^{m} x_{kl} \leq 1$

iii. $x_{mm} = 1$

iv. $\|\mathbf{X}^{(m)}\|_1$ *is* $o(m^{\lambda})$ *for some* $\lambda \in [0, 1)$ *where* $\|\cdot\|_1$ *denotes the maximum absolute column sum norm.*

Proposition 3 *Under assumption PP and DX i-iii, if* $p_{i0} = p_{i0}^*$

$$E[RV_i^{(m)}] = \int_0^1 \sigma_{iis} ds \tag{14}$$

Additionally, if $p_{j0} = p_{j0}^*$ *and* $tr(\mathbf{X}_i^{(m)} \mathbf{X}_j^{(m)\prime}) = m$,

$$E[RC_{ij}^{(m)}] = \int_0^1 \sigma_{ijs} ds \tag{15}$$

where $tr(\cdot)$ *is the trace operator.*

Realized variance is unbiased as long as the last price is observed. However, unbiasedness of realized covariance requires a further condition on the trace of $\mathbf{X}_i^{(m)\prime} \mathbf{X}_i^{(m)}$. If this condition is met, the product of these matrices will have a unit diagonal, and every cross-product of the two returns will contribute to realized variance or covariance. If some returns never appear in the same observed return, then realized covariance will generally be biased. Specifically, in the case that the integrated covariance is positive over any interval, realized covariance will be biased towards zero.

Consider a simple example. Suppose measurement nodes always correspond with observation nodes and a 4-sample realized covariance estimator is computed for two assets. Further, suppose

the efficient price of asset i is observed every period while the efficient price of asset j is only observed in even periods.

$$\mathbf{X}_i = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \text{ and } \mathbf{X}_j = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{16}$$

$$\mathbf{X}_i\mathbf{X}_j' = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \tag{17}$$

$$RC_{ij} = \mathbf{r}_i\mathbf{X}_i\mathbf{X}_j'\mathbf{r}_i' = r_{i1}r_{j2} + r_{i2}r_{j2} + r_{i4}r_{j3} + r_{i4}r_{j3} \tag{18}$$

and taking expectation conditional on the covariance process,

$$E[RC_{ij}] = \int_{1/4}^{1/2} \sigma_{ijs}\mathrm{d}s + \int_{3/4}^{1} \sigma_{ijs}\mathrm{d}s \tag{19}$$

which will generally not be equal to the integrated covariance.

A simple general condition is available on the structure of the $\mathbf{X}$ matrices to ensure the variance of the realized measures goes to zero.

Proposition 4 *Under assumption PP and DX for* $\mathbf{X}_i$*, if* $p_{i0} = p_{i0}^*$

$$V[RV_i^{(m)}] \to 0 \tag{20}$$

Additionally, if $p_{j0} = p_{j0}^*$ *and DX holds for* $\mathbf{X}_j$

$$V[RC_{ij}^{(m)}] \to 0 \tag{21}$$

The assumption that the column sum norm grows slower than the sample size ensures that the maximum number of efficient returns contained in any observed return is small relative to the number of samples. As long as this is true, the variance will vanish from either estimator. For realized covariance, these conditions are only sufficient and there are cases where the variance could tend to zero even if every efficient return is represented in some observed return. Consider a degenerate case where prices of asset i are updated continuously and those of asset j are only updated in the final period. The realized covariance estimator consist only of the covariance between

the returns at the last observation node and would be have a variance that converged to zero as m diverged.

Combining these results leads to a set of conditions for a consistent estimator.

Proposition 5 *Under assumption PP and DX for* $\mathbf{X}_i$*, if* $p_{i0} = p_{i0}^*$

$$RV_i^{(m)} \xrightarrow{p} \int_0^1 \sigma_{iis} ds \tag{22}$$

Additionally, if $p_{j0} = p_{j0}^*$, $\lim_{m\to\infty} m^{-1} tr(\mathbf{X}_i^{(m)} \mathbf{X}_j^{(m)\prime}) = 1$ *and DX holds for* $\mathbf{X}_j$

$$RC_{ij}^{(m)} \xrightarrow{p} \int_0^1 \sigma_{ijs} ds \tag{23}$$

The conditions for consistency of realized variance are the same as those for the variance to go to zero because RV is always unbiased as long as the first and last prices are recorded. Realized covariance requires that the number of efficient returns appearing observed prices tend to the sample size for large sampled and that no observed return contain too many efficient returns.

While the cases of deterministic returns are interesting in as much as they nest models previously examined, that are hardly realistic and the structure of the relations ship between observed returns and efficient returns does not require this. Further there is never an assumption that any observed return be known to be computed using the efficient price at the same point in time. Fortunately, in the case of random $\mathbf{X}$ matrices, the these propositions can be readily extended to cases where the measurement nodes are random as long as they are independent of the integrated variance. The structure of the $\mathbf{X}$ matrices ensures that realizations will consist of 1's and 0's. Thus, $\mathbf{X}$ can be considered as special Bernoulli matrices.

The properties of the cross products of $\mathbf{X}$ are particularly interesting. Examining the elements of $\mathbf{X}_i^{(m)} \mathbf{X}_i^{(m)\prime}$, the n$^{\text{th}}$ diagonal element is the probability that the n$^{\text{th}}$ efficient return is observed in the sample. However, the structure of $\mathbf{X}_i^{(m)} \mathbf{X}_j^{(m)\prime}$ is more interesting. In this case, the n$^{\text{th}}$ diagonal element is the probability the n$^{\text{th}}$ efficient returns from asset i and asset j are measured in the same return. If prices are always ordered, this is clearly one. However, under scrambling this can range from 0 to 1. Elements above the diagonal in the qs position are the probability that efficient return q from asset i appears in the same observed return with efficient return s from asset j, while below diagonal elements are the opposite. This leads to a new assumption and some results in the case of stochastic observation matrices.

Assumption 4 (SX) $\mathbf{X}$*, an* m *by* m *stochastic matrix, satisfies*

i. $x_{kl} = 1$ *or* $x_{kl} = 0$

ii. $\sum_{k=1}^{m} x_{kl} \leq 1$

iii. $Pr(x_{mm} = 1) = 1$

iv. $\|\mathbf{X}^{(m)}\|_1$ *is* $o_p(m^{\lambda})$ *for some* $\lambda \in [0, 1)$ *where* $\|\cdot\|_1$ *denotes the maximum absolute column sum norm.*

Proposition 6 *Suppose* $Pr(x_{imm} = 1) = 1$*. Under assumptions PP and SX i-iii, if* $p_{i0} = p_{i0}^*$*,*

$$E[RV_i^{(m)}] = \int_0^1 \sigma_{iis} ds \tag{24}$$

If additionally, SX iv holds,

$$RV_i^{(m)} \xrightarrow{p} \int_0^1 \sigma_{iis} ds \tag{25}$$

The assumption that $Pr(x_{imm} = 1) = 1$ is made for simplicity and to assure that the estimator is unbiased in any sample. Consistency could ensured under a weaker condition that $\lim_{m\to\infty} m^{-1} tr(\mathbf{X}_i^{(m)} \mathbf{X}_i^{(m)\prime}) \xrightarrow{p} 1$ which would imply that most returns (all but $o(m)$) contribute to realized variance. A realized covariance is consistent under similar conditions.

Proposition 7 *Under assumption PP and SX i-iii for both* $\mathbf{X}_i$ *and* $\mathbf{X}_j$*, if* $p_{i0} = p_{i0}^*$*,* $p_{i0} = p_{i0}^*$ *and* $E[tr(\mathbf{X}_i^{(m)} \mathbf{X}_j^{(m)\prime}] = m$

$$E[RC_{ij}^{(m)}] = \int_0^1 \sigma_{ijs} ds \tag{26}$$

If additionally, $\lim_{m\to\infty} m^{-1} tr(\mathbf{X}_i^{(m)} \mathbf{X}_j^{(m)\prime}) \xrightarrow{p} 1$ *and SX iv holds for both*

$$RC_{ij}^{(m)} \xrightarrow{p} \int_0^1 \sigma_{ijs} ds \tag{27}$$

As in the non-stochastic case, unbiasedness and consistency of realized variance puts additional requirements on the behavior of the measurement nodes. This theorem also points out the major problem with realized covariance. In general, if the measurement nodes are not perfectly dependent (with positive dependence), the realized covariance estimator will not be unbiased. Consider a simple example where the probability of observing an efficient asset price at an observation node is $1 - \pi_i$ for asset i and $1 - \pi_j$ for asset j, and the conditional on observing the price, observed prices always reflect the current efficient price. Asset prices can only observed over two periods, are always observed at the second period, and the probability of observing one is independent of the other. This is a realized covariance version of the model of Lo & MacKinley (1990). In this case, the elements of the $\mathbf{X}$ matrices take the value one with the following probabilities

$$Pr(\mathbf{X}_i = 1) = \begin{bmatrix} 1-\pi_i & \pi_i \\ 0 & 1 \end{bmatrix} \text{ and } Pr(\mathbf{X}_j = 1) = \begin{bmatrix} 1-\pi_j & \pi_j \\ 0 & 1 \end{bmatrix} \tag{28}$$

and

$$E[\mathbf{X}_i\mathbf{X}_j'] = \begin{bmatrix} \pi_i\pi_j + (1-\pi_i)(1-\pi_j) & \pi_i \\ \pi_j & 1 \end{bmatrix} \tag{29}$$

If $E[tr(\mathbf{X}_i\mathbf{X}_j')] = 2$ then $\pi_i = \pi_j/(2\pi_j - 1)$ which implies $\pi_i = 0$ or $\pi_i = 1$ corresponding to the case of never or always observing the efficient price, respectively. In the limit as m grows large, the diagonal elements of $E[\mathbf{X}_i\mathbf{X}_j']$ converge to

$$\frac{(1-\pi_i)(1-\pi_j)}{1-\pi_i\pi_j} \tag{30}$$

and realized covariance converges to

$$\int_0^1 \frac{(1-\pi_i)(1-\pi_j)}{1-\pi_i\pi_j}\sigma_{ijs}\mathrm{d}s = \frac{(1-\pi_i)(1-\pi_j)}{1-\pi_i\pi_j}\int_0^1 \sigma_{ijs}\mathrm{d}s \tag{31}$$

Thus, realized covariance is just a constant scaling of the integrated covariance and if a consistent estimators of π_i and π_2 are available, the bias could be estimated and a bias free estimator could be constructed. It's worth noting that the biased estimators also have variance that tends to zero since the column sums are $o_p(\epsilon)$ for any $\epsilon > 0$ and observed returns contains only finite runs of efficient returns with arbitrarily high probability.

However, if the data were generated from a model consistent with this specification, realized covariance would not systematically decrease as the sampling frequency increased (figure 1). A very simple simulation exercise demonstrates this. A bivariate brownian motion was simulated with daily variances of 1 and a correlation of 0.5. Returns were the efficient price with probability 50%, otherwise the previous price. 1000 simulations were performed. Figure 7 contains the median and 5% and 95% of the realized covariance computed from the simulated data. All three lines are converging to approximately $.16 = 0.5(1-0.5)(1-0.5)/(1-0.5^2)$, indicating that process has a non zero limit.

What if the probability of observing an observation was not constant but depended on the number of samples? Consider the case where $m\pi_i$ is $O(m^\lambda)$ for $\lambda \in (0,1)$.[7] In this case, $\pi_i = c_i m^{\lambda_i - 1}$ and $\pi_j = c_i m^{\lambda_j - 1}$ where c_i and c_j are less than one. Realized covariance will converge to 0 and the variance of realized covariance will also converge to zero since $\|\mathbf{X}_i\|_1$ and $\mathbf{X}_i\|_1$ are $o_p(m^{\lambda+\epsilon})$ for some

[7]The case where $\lambda = 0$, when intra-period returns are never observed, is special because the only return for any assets is the entire sample, and realized variance and realized covariance are both unbiased but clearly inconsistent.

$\epsilon > 0$. This isn't particularly surprising. The frequency of observation is becoming increasingly rare but returns are still observed arbitrarily often. Using data from the same simulation described above, but censoring according to $\pi_i = \pi_j = m^{-1/2}$, figure 8 shows that the realized correlations to tend to zero as the sampling frequency increases.

The interesting aspect of this specification is that realized variance is *still* consistent! Because the condition for the variance is be zero met, as long as the last observation is observed, realized variance will be unbiased with variance that goes to zero. Figure 9 contains the median and 5% and 95% quantiles of the realized variances. The median is essentially unbiased and very close to it's uncensored counterpart. In this setup, RV will be consistent and asymptotically normal but the rate of convergence will be different. This is easily observed as a simple modification of the assumptions of Barndorff-Nielsen & Shephard (2004) to account for the relatively rare measurement nodes.

5 Unbiased and Consistent Estimators

The ultimate goal of covariance estimation using high frequency data is to provide precise measures of the integrated covariance over some period, usually a day. The structure of this problem points to a method to construct unbiased estimators. From the definition of realized covariance,

$$RC_{ij}^{(m)} = \mathbf{r}_i^{*(m)} \mathbf{X}_i^{(m)} \mathbf{X}_j^{(m)\prime} \mathbf{r}_j^{*(m)\prime} \tag{32}$$

Consider a modified estimator of the form

$$RC_{ij}^{(m)} = \mathbf{r}_i^{*(m)} \mathbf{X}_i^{(m)} \mathbf{Q}_{ij} \mathbf{X}_j^{(m)\prime} \mathbf{r}_j^{*(m)\prime} \tag{33}$$

where $\mathbf{Q}_{ij}$ is a matrix which depends on the assumed process governing the measurement nodes. In the classic case, $\mathbf{Q}_{ij}$ is trivial, $\mathbf{I}_m$. However, cleverly choosing $\mathbf{Q}_{ij}$ can produce an unbiased and/or consistent estimator. For instance, one unbiased estimator can be constructed using descrambled returns, assuming the measurement nodes are stopping times rather than just realizations of a simple point process.

Definition 3 (Descrambled) *Suppose that prices when sampled according to (t_m) are scrambled and that there exists a non-empty set of stopping times $(\tilde{t}_q) \subset (t_m)$ such that prices sampled at $(\tilde{t}_q)$ are ordered. Prices sampled according to $(\tilde{t}_q)$ and returns constructed from these prices are said to be descrambled.*

A consistent estimator based on descrambled returns has been proposed by Hayashi & Yoshida (2005). Their estimator computes returns only when prices are known to be ordered and requires

that the measurement nodes be stopping times in addition to simple point processes. Consider the price of two returns observed to construct 4 returns. The prices are assumed to be known to be synchronized when ever observed. If asset i is observed at $t = 1, 3, 4$ while asset j is observed at $t = 2, 3, 4$, the $\mathbf{X}$ matrices can be described

$$\mathbf{X}_i = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \text{ and } \mathbf{X}_j = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{34}$$

A matrix $\mathbf{Q}_{ij}$ can be defined

$$\mathbf{Q}_{ij} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}' = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{35}$$

which will produce an unbiased estimator, noting that

$$\mathbf{X}_i \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = \mathbf{X}_j \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{36}$$

As long as the maximum column sum of the transformed $\mathbf{X}$ is finite, this estimator will be consistent as the transformed returns are ordered even though the original returns were not. However, the consistent estimator of Hayashi & Yoshida (2005) has some issues when the number of assets is large. If prices are only sampled when all assets are synchronized, the number of nosed will generally be very small when the number of stock is large. Alternatively, using only pairs to choose the descrambled returns can produce a non-positive definite covariance estimate, an undesirable property which renders it unsuitable for many applications.

Consistent estimators under pure censoring, where the probability of observing a synchronized return for asset i is $1-\pi_i$ and for observing a synchronized return for asset j is $1-\pi_j$. As previously noted, realized covariance converges to

$$\frac{(1-\pi_i)(1-\pi_j)}{1-\pi_i\pi_j} \int_0^1 \sigma_{ijs} \mathrm{d}s \tag{37}$$

and a consistent estimator can be constructed using $\mathbf{Q}_{ij} = \frac{(1-\pi_i)(1-\pi_j)}{1-\pi_i\pi_j}\mathbf{I}_m$. This framework can be easily extended to diurnal observation probabilities. Suppose that the probability of observing a

return for asset i is not $1 - p_{it}$ where p_{it} is continuous with the probability of observing asset j similarly defined. In this case, realized covariance converged to

$$\int_0^1 \psi_t \sigma_{ijs} \mathrm{d}s \tag{38}$$

where $\psi_t = \frac{(1-\pi_{it})(1-\pi_{jt})}{1-\pi_{it}\pi_{jt}}$. $\mathbf{Q}_{ij}$ can be defined as $diag(\psi_t)^{-1}$ where t correspond to the observation nodes. Consistency can be checked using proposition 7 on a transformed $\tilde{\mathbf{X}}_i = \mathbf{X}_i \mathbf{Q}_{ij}^{1/2}$. Trivially, as long as neither π_{it} or π_{jt} equal one (the case of not observing the process), $\|\tilde{\mathbf{X}}_i\|_1$ will be $o(m^\epsilon)$ for any $\epsilon > 0$ and by construction the normalized trace converges to 1.

in general, unbiased and consistent estimators will depend on the specific assumptions underlying the measurement node process, but can be constructed as long as the price process can be regularly observed. Only in cases where at least one of the prices cannot be observed for a finite amount of time can no consistent estimator be constructed.

6 Conclusion

Market microstructure noise affects both realized covariance and realized variance. This paper shows that nature of the effect is very different. Realized covariance is not biased in the presence of pure additive noise but shows a massive bias if returns are scrambled. However, sufficient conditions for an estimator of the integrated covariance to be consistent are generally stricter than those for an estimator of the integrated variance to be consistent. This difference reduces to one simple fact: observed prices of a single series, even if reflecting a past efficient price, are *always* synchronized. Moreover, realistically the observed prices to two assets will likely never be perfectly synchronized.

Examining the behavior of both realized covariance and realized variance constructed using various sampling frequencies shows radically different patterns. Using mid quote prices from the DJIA and from the big three US automobile manufacturers, realized covariance show large changes as the sampling frequency increases while realized variances evidence little change. High-frequency returns are not autocorrelated but typically have many significant cross-correlations.

Two models, one with a simple additive error and one which allows for scrambled returns, were examined for their abilities to match these empirical facts. Models with simple independent additive noise are incapable of matching any of the patterns evidenced in either data set. However, a simple model which allows for returns to occur out of order can generate large biases in realized correlation. Sufficient conditions were examined for realized covariance to be consistently estimated and a general procedure for producing unbiased and/or consistent estimators has been described. Revisiting the data in light the results of this paper, a simple random censoring model appears to perform very well in terms of reducing the obvious scaling bias in the unadjusted realized covariance.

However, there is evidence that this model may not be appropriate for assets in the same industry group such as the big 3 automobile manufacturers or two oil producers.

This paper has left a number of important questions unanswered. First, can a generic estimator using high frequency returns be constructed that is consistent under a wide variety of conditions. For instance, suppose that the returns were known to be subject to random censoring but with unknown probabilities. Under what conditions on the censoring process could a consistent estimator be constructed. What happens if returns are both scrambled and subject to market microstructure noise. This paper has shown that the usual realized covariance estimator is inconsistent under these circumstances, but can a new estimator, possibly using a kernel, be used to produce consistent estimates. Finally, the largest issues for any microstructure noise contaminated estimator remains. Can a consistent estimator be constructed if the scrambling process and the instantaneous covariance are not independent. We leave these as issues for further research.

Appendix

Proof of Proposition 1: The m sample realized covariance can be written

$$RC_{ij}^{(m)} = \sum_{n=1}^{m} r_{in}^* r_{jn}^* + \sum_{n=1}^{m} r_{in}^* e_{jn} + \sum_{n=1}^{m} r_{in}^* e_{in} + \sum_{n=1}^{m} e_{in} e_{jn} \tag{1}$$

Taking expectations, noting that $E[r_{in}^* r_{jn}^*] = \int_{\frac{n-1}{m}}^{\frac{n}{m}} \sigma_{ijs} \mathrm{d}s$ and that e_n is independent of r_n,

$$E[RC_{ij}^{(m)}] = \sum_{n=1}^{m} \int_{\frac{n-1}{m}}^{\frac{n}{m}} \sigma_{ijs} \mathrm{d}s + \sum_{n=1}^{m} E[r_{in}^*]E[e_{jn}] + \sum_{n=1}^{m} E[r_{in}^*]E[e_{in}] + \sum_{n=1}^{m} E[e_{in}]E[e_{jn}] = \int_0^1 \sigma_{ijs} \mathrm{d}s \tag{2}$$

To compute the variance of realized covariance, note that independence implies

$$Var(RC_{ij}^{(m)}) = Var(\sum_{n=1}^{m} r_{in}^* r_{jn}^*) + Var(\sum_{n=1}^{m} r_{in}^* e_{jn}) + Var(\sum_{n=1}^{m} r_{in}^* e_{in}) + Var(\sum_{n=1}^{m} e_{in} e_{jn}) \tag{3}$$

$$Var(\sum_{n=1}^{m} r_{in}^* r_{jn}^*) = \sum_{n=1}^{m} Var(r_{in}^* r_{jn}^*) \qquad \text{by independence} \tag{4}$$

$$= \sum_{n=1}^{m} \frac{\int_{\frac{n}{m}}^{\frac{n-1}{m}} \left\{ \sigma_{iis} \sigma_{jjs} + \sigma_{ijs}^2 \right\} \mathrm{d}s}{m} \qquad \text{assumption PP} \tag{5}$$

$$= \frac{\int_0^1 \left\{ \sigma_{iis} \sigma_{jjs} + \sigma_{ijs}^2 \right\} \mathrm{d}s}{m} \tag{6}$$

$$Var(\sum_{n=1}^{m} r_{in}^* e_{jn}) = Var(\sum_{n=1}^{m} r_{in}^* (u_{jn} - u_{jn-1})) \tag{7}$$

$$= \sum_{n=1}^{m} Var(r_{in}^* (u_{jn} - u_{jn-1})) \qquad \text{by independence or } r_{in}, r_{in+h} \tag{8}$$

$$= \sum_{n=1}^{m} Var(r_{in}^*) Var(u_{jn} - u_{jn-1}) \qquad \text{by independence or } r_{in}, e_{jn} \tag{9}$$

$$= 2\omega_j^2 \sum_{n=1}^{m} \int_{\frac{n}{m}}^{\frac{n-1}{m}} \sigma_{iis} \mathrm{d}s \qquad \text{by normality of } r_{in} \text{ and Assumption 2} \tag{10}$$

$$= 2\omega_j^2 \int_0^1 \sigma_{iis} \mathrm{d}s \tag{11}$$

$$Var(\sum_{n=1}^{m} r_{jn}^* e_{in}) = 2\omega_i^2 \int_0^1 \sigma_{jjs} \mathrm{d}s \qquad \text{by symmetry} \tag{12}$$

$$Var(\sum_{n=1}^{m} e_{in}e_{jn}) = E(\sum_{n=1}^{m}(e_{in}e_{jn})^2) + 2E(\sum_{n=2}^{m} e_{in}e_{in-1}e_{jn}e_{jn-1}) + 0 \qquad \text{by definition of } e_n \tag{13}$$

$$= \sum_{n=1}^{m}(2\omega_i^2)(2\omega_j^2) + 2\sum_{n=1}^{m}\omega_i^2\omega_j^2 \tag{14}$$

$$= 6m\omega_i^2\omega_j^2 \tag{15}$$

Combining provides the desired result.

Proof of Proposition 2: $E[(r_{in}r_{jn+h}] = E[(r^*_{in} + e_{in})(r^*_{jn} + e_{jn})]$. By assumption 1, $E[r^*_{in}r^*_{jn+h}] = 0$, and by assumption 2 $E[r^*_{in}e_{jn+h}] = 0$, $E[r^*_{jn+h}e_{in}] = 0$, and $E[e_{in}e_{jn+h}] = 0$. Additionally, assumption 1 and assumption 2.i. provide that both returns and shocks are mean zero.

Proof of Proposition 3: Realized volatility under scrambling can be written

$$RV^{(m)} = \mathbf{r}^{*(m)}\mathbf{X}^{(m)}\mathbf{X}^{(m)\prime}\mathbf{r}^{*(m)\prime} \tag{16}$$

which is equivalent to

$$RV^{(m)} = (\mathbf{r}^{*(m)} \otimes \mathbf{r}^{*(m)})\text{vec}(\mathbf{X}^{(m)}\mathbf{X}^{(m)\prime}) \tag{17}$$

Taking expectations, nothing that $E[r^{*}_{in}{}^2] = \int_{(n-1)/m}^{n/m}\sigma_{iis}\mathrm{d}s$ and $E[r^*_q r^*_s] = 0$ for $q \neq s$.

$$E[RV^{(m)}] = \left[\int_{1/m}^{0}\sigma_{iis}\mathrm{d}s \;\; \mathbf{0}_m \;\; \int_{(1/m}^{2/m}\sigma_{iis}\mathrm{d}s \;\; \ldots \mathbf{0}_m \int_{(m-1)/m}^{1}\sigma_{iis}\mathrm{d}s\right]\text{vec}(\mathbf{X}^{(m)}\mathbf{X}^{(m)\prime}) \tag{18}$$

where $\mathbf{0}_m$ is a 1 by m vector of zeros. Finally, since $\mathbf{X}^{(m)}$ is a matrix of 1's and 0's with at most one 1 per row, $\mathbf{X}^{(m)}\mathbf{X}^{(m)\prime}$ must have either 1 or 0 in each diagonal place. However, by $DXiii.$, $x_{mm} = 1$, so all returns are represented in some return. Thus, every diagonal element is one, so

$$E[RV^{(m)}] = \sum_{n=1}^{m}\int_{(n-1)/m}^{n/m}\sigma_{iis}\mathrm{d}s = \int_0^1\sigma_{iis}\mathrm{d}s \tag{19}$$

The proof for realized covariance is identical except that the trace condition, combined with a value of 1 or zero is sufficient to guarantee each diagonal element is 1.

Proof of Proposition 4: If $\|\mathbf{X}^{(m)}\|_1$ is $o(m^\lambda)$, then

$$Var(\sum_{n=1}^{m} r_n^2) \leq Var(\sum_{n=1}^{m-K}(\sum_{q=1}^{K} r^*_{n+q})^2) \leq o(m^\lambda)Var(\sum_{n=1}^{m} r^{*}_{n}{}^2) = 2o(m^\lambda)(\int_0^1 \frac{\sigma_s^4\mathrm{d}s}{m} + o(\frac{1}{m})) \to 0 \tag{20}$$

Similarly, if $\|\mathbf{X}_i^{(m)}\|$ is $o(m^\lambda)$ and $\|\mathbf{X}_j^{(m)}\|_1$ is $o(m^\lambda)$, then

$$Var(\sum_{n=1}^{m} r_{in}r_{jn}) \leq Var(\sum_{n=1}^{m-K}(\sum_{q=1}^{K} r^*_{in+q}r^*_{jn+q})) \tag{21}$$

$$\leq o(m^\lambda)Var(\sum_{n=1}^{m} r^*_{in}r^*_{jn}) \tag{22}$$

$$= 2o(m^\lambda)(\int_0^1 \frac{\{\sigma_{iis}\sigma_{jjs} + \sigma^2_{ijs}\}\mathrm{d}s}{m} + o(\frac{1}{m})) \to 0 \tag{23}$$

where K is $o(m^\lambda)$.

Note: Equality would follow if K overlapping returns were used in the construction of the realized covariance.

Proof of Proposition 5:

Realized variance ($RV^{(m)}$) is trivially unbiased as long as DX is met since the last observation is always recorded by assumption. By proposition 4, the variance goes to 0 and by Chebyshev's inequality, it must converge in probability.

Noting that the diagonal elements of $\mathbf{X}_i^{(m)}\mathbf{X}_j^{(m)\prime}$ are less than (or equal to) 1, if $m^{-1}tr(\mathbf{X}_i^{(m)}\mathbf{X}_j^{(m)\prime}) \to 1$, then $\exists m$ such that $|m^{-1}tr(\mathbf{X}_i^{(m)}\mathbf{X}_j^{(m)\prime}) - 1| < \epsilon$ for any $\epsilon > 0$. Letting x_{ijm} by a diagonal element, then

$$1 - \frac{\epsilon}{m} \leq x_{ijm} \leq 1 \tag{24}$$

$$\left(1 - \frac{\epsilon}{m}\right) \int_0^1 \sigma_{ijs} \mathrm{d}s \leq E[RC_{ij}^{(m)}] \leq \int_0^1 \sigma_{ijs} \mathrm{d}s \tag{25}$$

so $E[RC_{ij}^{(m)}] = \int_0^1 \sigma_{ijs} \mathrm{d}s + o(\frac{1}{m})$. From proposition 4, $Var(RC_{ij}^{(m)}) \to 0$ and by Chebyshev's inequality, $RC_{ij}^{(m)} \xrightarrow{p} \sigma_{ijs}$.

Proof of Proposition 6:

$$RV^{(m)} = (\mathbf{r}^{*(m)} \otimes \mathbf{r}^{*(m)})\mathrm{vec}(\mathbf{X}^{(m)}\mathbf{X}^{(m)\prime}) \tag{26}$$

Since $\mathbf{X}^{(m)}$ is independent from $\mathbf{p}_t$ and $\mathbf{r}_t$,

$$E(RV^{(m)}) = E(\mathbf{r}^{*(m)} \otimes \mathbf{r}^{*(m)})E(\mathrm{vec}(\mathbf{X}^{(m)}\mathbf{X}^{(m)\prime})) \tag{27}$$

$$E[RV^{(m)}] = \left[\int_{1/m}^0 \sigma_{iis}\mathrm{d}s \;\; \mathbf{0}_m \;\; \int_{(1/m}^{2/m} \sigma_{iis}\mathrm{d}s \;\; \ldots \mathbf{0}_m \int_{(m-1)/m}^1 \sigma_{iis}\mathrm{d}s\right] E(\mathrm{vec}(\mathbf{X}^{(m)}\mathbf{X}^{(m)\prime})) \tag{28}$$

Since $Pr(x_{mm} = 1) = 1$, $E(\mathbf{X}^{(m)}\mathbf{X}^{(m)\prime})$ has a unit diagonal with probability 1, and

$$E[RV^{(m)}] = \left[\int_{1/m}^0 \sigma_{iis}\mathrm{d}s \;\; \mathbf{0}_m \;\; \int_{(1/m}^{2/m} \sigma_{iis}\mathrm{d}s \;\; \ldots \mathbf{0}_m \int_{(m-1)/m}^1 \sigma_{iis}\mathrm{d}s\right] E(\mathrm{vec}(\mathbf{X}^{(m)}\mathbf{X}^{(m)\prime})) \tag{29}$$

$$= \sum_{n=1}^m \int_{(n-1)/m}^{n/m} \sigma_{iis}\mathrm{d}s \tag{30}$$

$$= \int_0^1 \sigma_{iis}\mathrm{d}s \tag{31}$$

Proof of Proposition 7:

$$RC_{ij}^{(m)} = (\mathbf{r}_i^{*(m)} \otimes \mathbf{r}_j^{*(m)})\mathrm{vec}(\mathbf{X}_i^{(m)}\mathbf{X}_j^{(m)\prime}) \tag{32}$$

Since $\mathbf{X}_i^{(m)}$ and $\mathbf{X}_j^{(m)}$ independent from $\mathbf{p}_t$ and $\mathbf{r}_t$,

$$E(RC_{ij}^{(m)}) = E(\mathbf{r}_i^{*(m)} \otimes \mathbf{r}_j^{*(m)})E(\mathrm{vec}(\mathbf{X}_i^{(m)}\mathbf{X}_j^{(m)\prime})) \tag{33}$$

$$E[RC_{ij}^{(m)}] = \left[\int_{1/m}^0 \sigma_{ijs}\mathrm{d}s \;\; \mathbf{0}_m \;\; \int_{(1/m}^{2/m} \sigma_{ijs}\mathrm{d}s \;\; \ldots \mathbf{0}_m \int_{(m-1)/m}^1 \sigma_{ijs}\mathrm{d}s\right] E(\mathrm{vec}(\mathbf{X}_i^{(m)}\mathbf{X}_j^{(m)\prime})) \tag{34}$$

Since the diagonal elements of $E(\mathrm{vec}(\mathbf{X}_i^{(m)}\mathbf{X}_j^{(m)\prime}))$ are less than or equal to 1 by construction (they are probabilities), by assumption this matrix has a unit diagonal with probability 1, and

$$E[RC^{(m)}] = \left[\int_{1/m}^{0} \sigma_{ijs} \mathrm{d}s \;\; \mathbf{0}_m \int_{(1/m}^{2/m} \sigma_{ijs} \mathrm{d}s \;\; \ldots \mathbf{0}_m \int_{(m-1)/m}^{1} \sigma_{ijs} \mathrm{d}s \right] E(\mathrm{vec}(\mathbf{X}_i^{(m)} \mathbf{X}_j^{(m)\prime})) \tag{35}$$

$$= \sum_{n=1}^{m} \int_{(n-1)/m}^{n/m} \sigma_{ijs} \mathrm{d}s \tag{36}$$

$$= \int_{0}^{1} \sigma_{ijs} \mathrm{d}s \tag{37}$$

References

Andersen, T., Bollerslev, T., Diebold, F. X. & Labys, P. (2003), 'Modeling and forecasting realized volatility', *Econometrica* **71**(1), 3–29.

Andersen, T. G., Bollerslev, T., Diebold, F. X. & Ebens, H. (2001), 'The distribution of stock return volatility', *Journal of Financial Economics* **61**, 43–76.

Bandi, F. & Russell, J. (2005*a*), Microstructure noise, realized variance, and optimal sampling. University of Chicago.

Bandi, F. & Russell, J. (2005*b*), Realized covariation, realized beta, and microstructure noise. University of Chicago.

Barndorff-Nielsen, O. E. & Shephard, N. (2004), 'Econometric analysis of realised covariation: high frequency based covariance, regression and correlation in financial economics', *Econometrica* **73**(4), 885–925.

Barndorff-Nielsen, O., Hansen, P. R., Lunde, A. & Shephard, N. (2004), Regular and modified kernel-based estimators of integrated variance: The case with independent noise. Stanford University.

Ebens, H. (1999), Realized stock volatility. Johns Hopkins University, Working Paper 420.

Epps, T. W. (1979), 'Comovements in stock prices in the very short run', *Journal of the American Statistical Society* **74**, 291–296.

Hansen, P. R. & Lunde, A. (2004*a*), Realized variance and market microstructure noise. Stanford University.

Hansen, P. R. & Lunde, A. (2004*b*), A realized variance for the whole day based on intermittent high-frequency data. Stanford University.

Hansen, P. R. & Lunde, A. (2004*c*), An unbiased measure of realized variance. Stanford University.

Hayashi, T. & Yoshida, N. (2005), 'On covariance estimation of non-synchronously observed diffusion processes', *Bernoulli* **11**(2), 359–379.

Lo, A. & MacKinley, A. C. (1990), 'Econometric analysis', *Journal of Financial Economics* **19**, 3–29.

Martens, M. (2004), Estimating unbiased and precise realized covariances. Erasmus University Rotterdam.

Merton, R. C. (1980), 'On estimating the expected return on the market: An exploratory investigation', *Journal of Financial Economics* **8**(4), 323–361.

Zhang, L., Mykland, P. & Aït-Sahalia, Y. (2004), A tale of two time scales: Determining integrated volatility with noisy high-frequency data. Forthcoming *Journal of the American Statistical Association*.

DJIA Summary Statistics

Ticker	Firm Name	Quotes Per Day	Informative Quotes Per Day	% of intervals with change 1 min	5 min	10 min	30 min
AA	Alcoa Inc.	313	135	0.254	0.639	0.797	0.893
ALD	Allied Signal Inc.	320	117	0.219	0.579	0.743	0.870
AXP	American Express Co.	443	146	0.233	0.526	0.661	0.805
BA	Boeing Co.	431	150	0.268	0.615	0.751	0.855
TRV	Travelers	475	116	0.220	0.552	0.701	0.837
CAT	Caterpillar Inc.	373	152	0.282	0.665	0.812	0.894
CHV	Chevron	387	139	0.252	0.600	0.746	0.859
DD	DuPont	628	191	0.307	0.656	0.788	0.879
DIS	Walt Disney Co.	565	164	0.278	0.612	0.750	0.864
EK	Eastman Kodak	479	149	0.257	0.589	0.728	0.844
GE	General Electric Co.	785	213	0.330	0.678	0.806	0.886
GM	General Motors Corp.	378	131	0.244	0.589	0.741	0.863
GT	Goodyear	279	94	0.182	0.521	0.694	0.848
HWP	Hewlett-Packard Inc.	794	249	0.390	0.773	0.886	0.915
IBM	Intl. Bus. Machines	554	272	0.421	0.782	0.881	0.910
IP	Intl. Paper	447	135	0.241	0.597	0.752	0.868
JNJ	Johnson & Johnson	579	155	0.269	0.598	0.737	0.854
JPM	J.P. Morgan & Co.	397	170	0.289	0.658	0.803	0.893
KO	Coca-Cola Co.	544	131	0.222	0.517	0.664	0.821
MCD	McDonald's Corp.	307	97	0.184	0.475	0.624	0.789
MMM	3M Co.	331	137	0.254	0.620	0.771	0.874
MO	Altria Group Inc.	1077	218	0.328	0.676	0.806	0.890
MRK	Merck & Co. Inc.	455	174	0.277	0.571	0.698	0.822
PG	Procter & Gamble Co.	672	231	0.348	0.689	0.811	0.889
S	Sears	446	134	0.242	0.596	0.745	0.864
T	AT&T	330	115	0.216	0.527	0.674	0.823
UK	Union Carbide	250	77	0.150	0.422	0.568	0.733
UTX	United Tech. Corp.	311	126	0.230	0.586	0.744	0.863
WMT	Wal-Mart Stores Inc.	361	90	0.160	0.394	0.523	0.701
XON	Exxon	531	163	0.272	0.595	0.731	0.849

Table 1: Summary Statistics: This table contains the average number of quotes per day for each stock. Informative quotes are those where either the bid price or the ask price changed from the previous quote. The last four columns show the percentage of intervals which contain an informative quote when sampling using 1, 5, 10 and 30 minute windows. Quotes were measured from 9:30 until 16:10.

Correlation Scaling

	Average Correlation						Maximum Correlation					
	1 min	5 min	10 min	30 min	1 day	2 day	1 min	5 min	10 min	30 min	1 day	2 day
AA	0.083	0.165	0.200	0.188	0.216	0.236	0.114	0.225	0.258	0.237	0.392	0.505
ALD	0.112	0.206	0.244	0.230	0.233	0.260	0.161	0.290	0.331	0.302	0.344	0.415
AXP	0.106	0.196	0.228	0.195	0.216	0.237	0.148	0.271	0.306	0.281	0.446	0.476
BA	0.101	0.189	0.221	0.193	0.205	0.209	0.155	0.267	0.293	0.267	0.309	0.370
C	0.101	0.173	0.207	0.181	0.186	0.195	0.179	0.300	0.382	0.414	0.564	0.572
CAT	0.102	0.195	0.237	0.222	0.228	0.256	0.140	0.266	0.317	0.292	0.340	0.413
CHV	0.109	0.211	0.240	0.213	0.202	0.199	0.176	0.353	0.410	0.440	0.588	0.555
DD	0.122	0.229	0.263	0.228	0.239	0.255	0.178	0.328	0.362	0.307	0.314	0.348
DIS	0.110	0.209	0.243	0.210	0.205	0.212	0.164	0.296	0.336	0.293	0.305	0.291
EK	0.083	0.156	0.180	0.163	0.148	0.165	0.112	0.209	0.236	0.208	0.204	0.233
GE	0.152	0.281	0.314	0.274	0.285	0.293	0.210	0.373	0.406	0.362	0.438	0.451
GM	0.107	0.200	0.240	0.216	0.233	0.251	0.179	0.300	0.382	0.414	0.564	0.572
GT	0.090	0.165	0.201	0.194	0.202	0.233	0.121	0.218	0.257	0.248	0.312	0.374
HWP	0.108	0.206	0.238	0.203	0.193	0.191	0.166	0.302	0.352	0.331	0.432	0.434
IBM	0.111	0.214	0.245	0.209	0.207	0.211	0.166	0.308	0.352	0.331	0.432	0.434
IP	0.102	0.183	0.216	0.194	0.207	0.225	0.138	0.253	0.294	0.251	0.392	0.505
JNJ	0.127	0.233	0.266	0.231	0.214	0.217	0.186	0.335	0.386	0.397	0.559	0.584
JPM	0.120	0.232	0.269	0.252	0.278	0.282	0.166	0.325	0.368	0.345	0.446	0.476
KO	0.135	0.249	0.278	0.249	0.257	0.253	0.210	0.371	0.401	0.361	0.465	0.466
MCD	0.100	0.184	0.217	0.189	0.194	0.193	0.137	0.256	0.292	0.259	0.332	0.347
MMM	0.108	0.211	0.243	0.220	0.204	0.207	0.154	0.295	0.326	0.293	0.314	0.328
MO	0.096	0.177	0.205	0.184	0.179	0.173	0.137	0.253	0.286	0.248	0.277	0.294
MRK	0.122	0.224	0.254	0.217	0.228	0.221	0.181	0.329	0.386	0.397	0.559	0.584
PG	0.134	0.252	0.285	0.250	0.240	0.226	0.203	0.373	0.406	0.362	0.465	0.466
S	0.105	0.197	0.228	0.202	0.235	0.251	0.143	0.266	0.294	0.259	0.348	0.379
T	0.112	0.208	0.244	0.213	0.200	0.192	0.157	0.292	0.327	0.280	0.300	0.283
UK	0.079	0.142	0.174	0.166	0.170	0.178	0.098	0.177	0.208	0.213	0.283	0.348
UTX	0.103	0.190	0.224	0.213	0.226	0.267	0.157	0.274	0.307	0.299	0.344	0.415
WMT	0.102	0.184	0.212	0.192	0.201	0.189	0.143	0.251	0.282	0.260	0.348	0.379
XON	0.127	0.248	0.268	0.225	0.210	0.190	0.198	0.371	0.410	0.440	0.588	0.555

Table 2: Correlation Scaling: This table contains the average correlation for each of the Dow Jones constituents as the sampling window increases from 1 minute to 30 minutes. All correlations were computed using variances computed with 5-minute returns. One-day and two-day returns were computed using close-to-close returns, overlapped for the two-day. The average correlation is clearly climbing until 10 minutes. The right panel contains the maximum correlation for each of the stocks. 26 of the 30 stocks had increases maximum when going from 10 minute returns to 2 day.

Variance Scaling (Annualized)

	1 min	5 min	10 min	30 min	1 day	2 day
AA	0.237	0.244	0.247	0.244	0.258	0.265
ALD	0.262	0.268	0.271	0.262	0.247	0.243
AXP	0.278	0.278	0.274	0.259	0.260	0.258
BA	0.248	0.255	0.257	0.246	0.249	0.247
C	0.313	0.312	0.311	0.297	0.307	0.318
CAT	0.256	0.269	0.274	0.264	0.276	0.279
CHV	0.211	0.211	0.210	0.204	0.208	0.204
DD	0.242	0.246	0.245	0.236	0.235	0.234
DIS	0.246	0.250	0.247	0.238	0.240	0.238
EK	0.273	0.275	0.275	0.274	0.288	0.294
GE	0.211	0.217	0.213	0.201	0.200	0.200
GM	0.243	0.250	0.252	0.249	0.265	0.267
GT	0.244	0.243	0.242	0.237	0.236	0.232
HWP	0.316	0.331	0.333	0.324	0.346	0.342
IBM	0.270	0.281	0.281	0.276	0.305	0.304
IP	0.254	0.253	0.250	0.242	0.251	0.250
JNJ	0.250	0.251	0.249	0.237	0.239	0.240
JPM	0.202	0.208	0.208	0.203	0.223	0.221
KO	0.214	0.214	0.211	0.206	0.214	0.214
MCD	0.239	0.236	0.231	0.219	0.216	0.219
MMM	0.204	0.211	0.211	0.207	0.205	0.196
MO	0.285	0.286	0.286	0.279	0.288	0.287
MRK	0.242	0.248	0.245	0.235	0.260	0.258
PG	0.230	0.237	0.236	0.224	0.219	0.213
S	0.258	0.264	0.265	0.260	0.288	0.292
T	0.222	0.225	0.225	0.219	0.238	0.242
UK	0.290	0.286	0.286	0.276	0.269	0.269
UTX	0.207	0.218	0.222	0.221	0.210	0.214
WMT	0.304	0.296	0.288	0.273	0.273	0.273
XON	0.198	0.201	0.196	0.188	0.191	0.182

Table 3: Volatility Scaling: Annualized Volatility from prices sampled using 1, 5, 10 and 30 minute returns and 1 and 2 day (overlapping) returns. For intra-daily frequency, average variance was computed as $\overline{RV}^{(m)} = T^{-1}\sum_{t=1}^{T} RV_t^{(m)}$. The reported numbers are $\sqrt{252\overline{RV}^{(m)}}$.

Big 3 Auto. Manu. Summary Statistics

	Quotes Per Day	Informative Quotes Per Day	% of intervals with change 1 min	5 min	10 min	30 min
C	632	212	0.328	0.677	0.804	0.905
F	722	261	0.365	0.679	0.791	0.895
GM	852	314	0.414	0.732	0.843	0.930

Variance Scaling (Annualized)

	1 min	5 min	10 min	30 min	1 day	2 day
C	0.355	0.356	0.355	0.351	0.340	0.348
F	0.328	0.331	0.328	0.323	0.330	0.308
GM	0.277	0.296	0.300	0.301	0.317	0.312

Correlation Scaling

	1 min	5 min	10 min	30 min	1 day	2 day
C-F	0.182	0.255	0.292	0.324	0.511	0.533
C-GM	0.163	0.243	0.283	0.314	0.493	0.502
F-GM	0.186	0.290	0.340	0.402	0.576	0.596

Table 4: Summary statistic for the big 3 auto makers. The top panel contains the number of quotes and number of quotes with a price change per day. It also contains the percentage of high frequency returns which contain an informative quote. The middle panel contains (annualized) volatility when computed using returns ranging from 1-minute to 2 days. There is little systematic bias and all volatilities lie in a 15% range. The bottom panel contains the *pseudo*-correlations (realized covariance divided by 5-minute realized variance) of the three pais. They are all monotonically increasing, and have significant bias even when sampled using 30-minute returns.

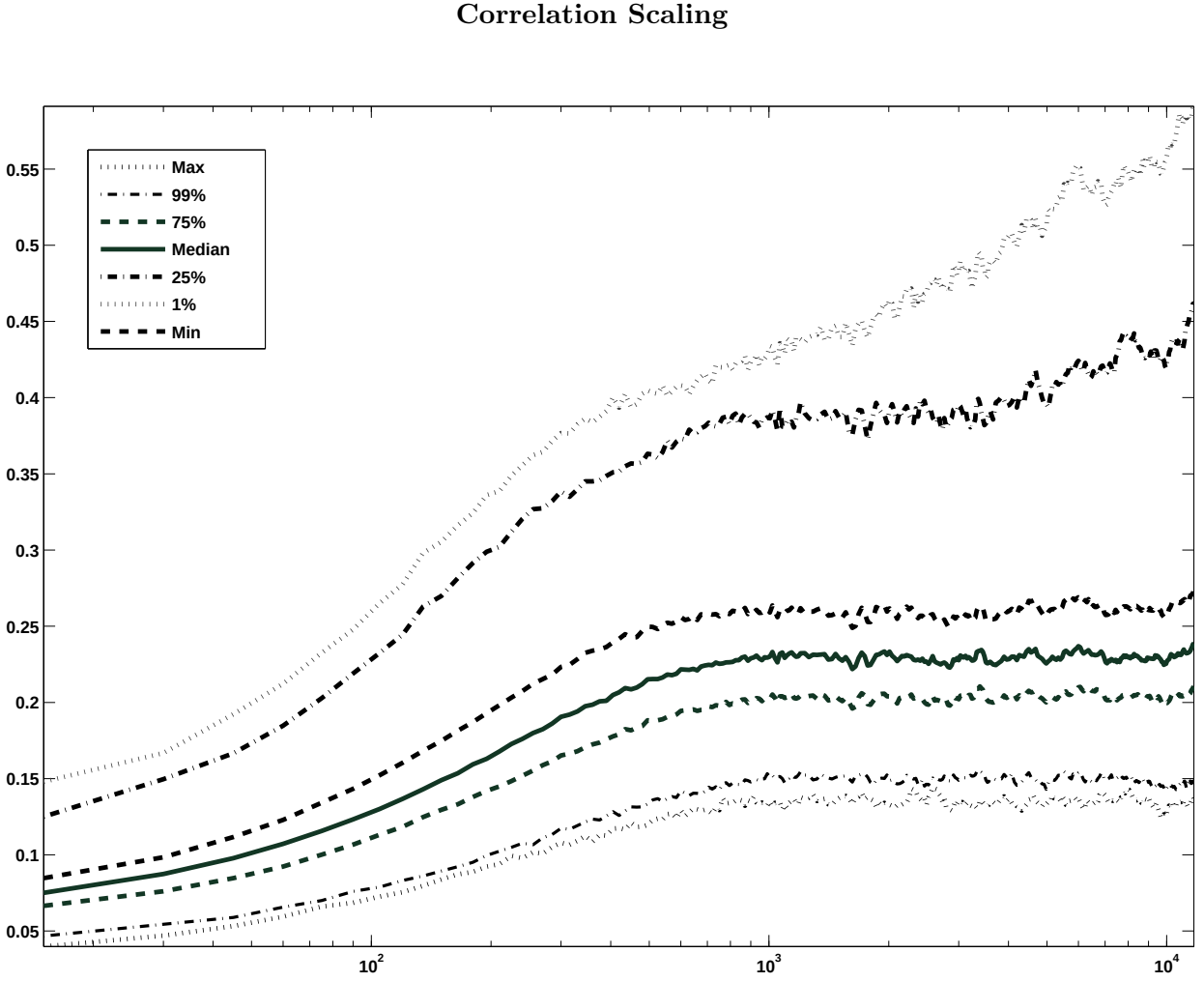

Figure 1: Correlation Scaling: Quantiles of correlation computed from 15 seconds to 3.25 hours (1/2 day). For each asset pair of the DJIA, realized covariance was computed using window lengths ranging from 15 seconds to 3.25 hours (1/2 day). The realized covariances were then transformed in to correlation using the 5-minute realized variances to facilitate comparisons across different window lengths. There is substantial bias when using returns sampled more frequently than 1000 seconds (18 minutes).

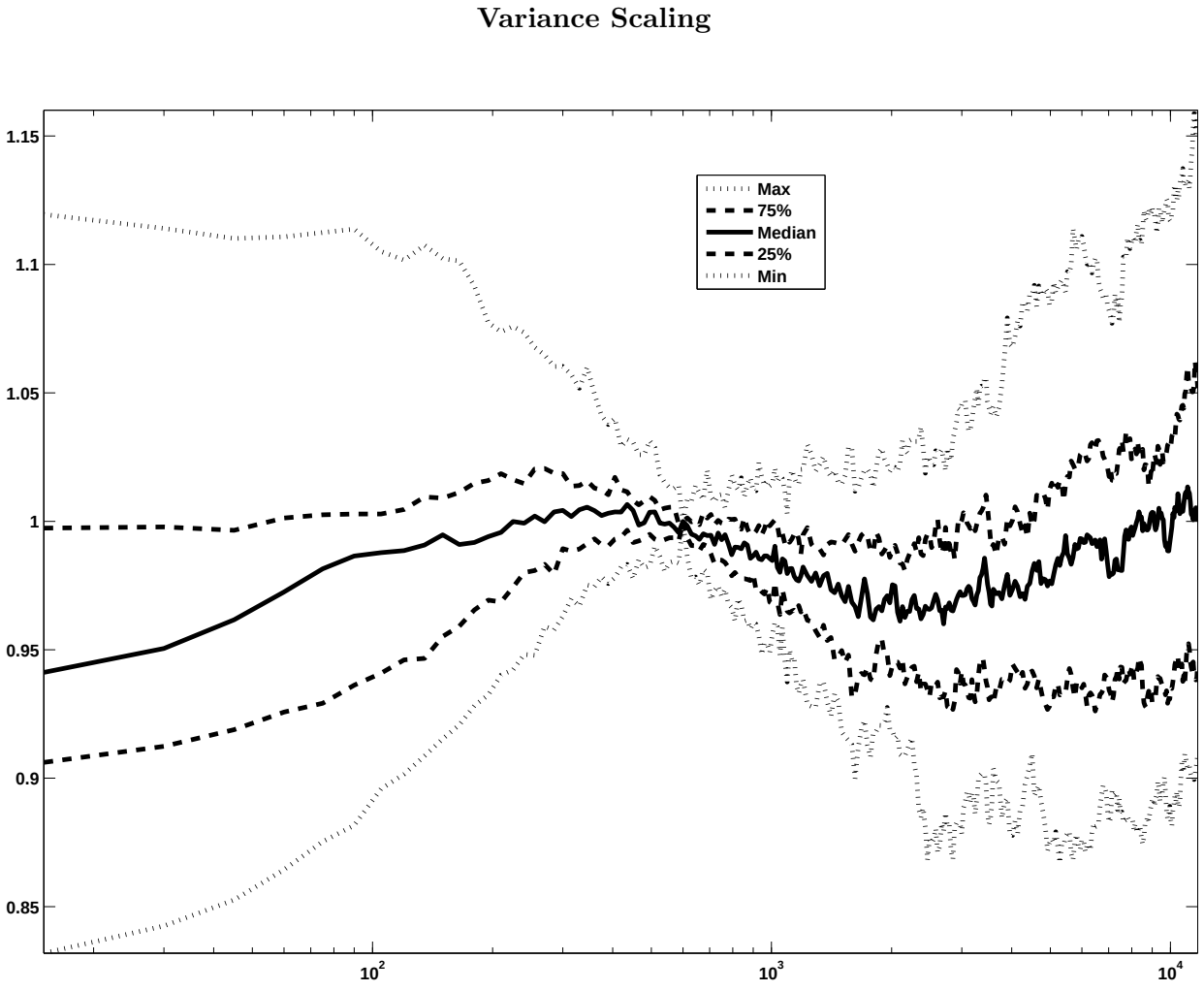

Figure 2: Variance Scaling: Quantiles of the standardized variances computed using returns ranging from 15 seconds to 3.25 hours (1/2 day). Variances were computed for each of the 30 DJIA stocks using each windows length. Variances at each window length were then divided by the 5-minute realized variance. The symmetry and lack of any systemic bias contrasts starkly with the quantiles of realized correlation.

Variance and Correlation Scaling for the Auto Manufacturers

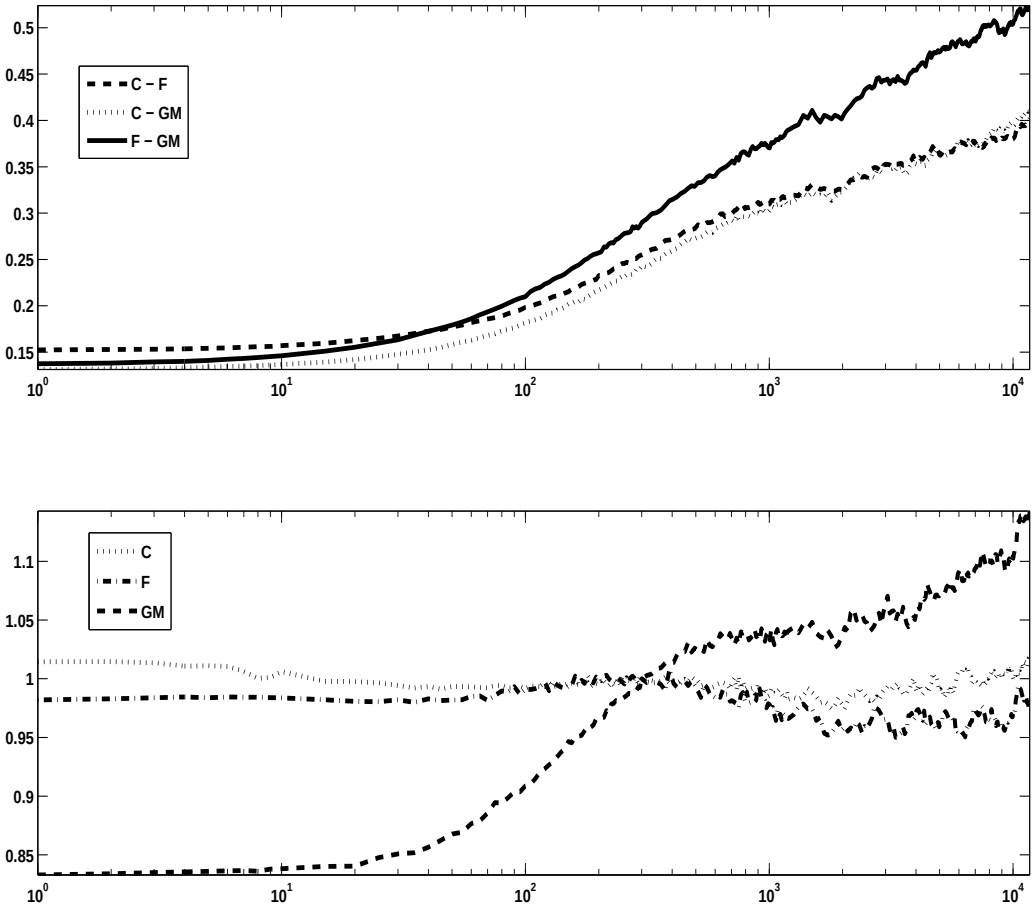

Figure 3: Variance and Correlation Scaling (Automobile Manufacturers): The top panel plots realized correlation, where the variance for each sample window was computed using 5-minute realized covariance. Log scaling of each covariance is linear for the range of sampling windows, from 1 second to 3.25 hours (1/2 trading day). The bottom figure shows the realized variance computed using returns from 1 second to 3.25 hours standardized by the 5-minute realized variance. The Realized variances range over a 15% while the realized covariances change by 300%. The time scales in both figures is logarithmic.

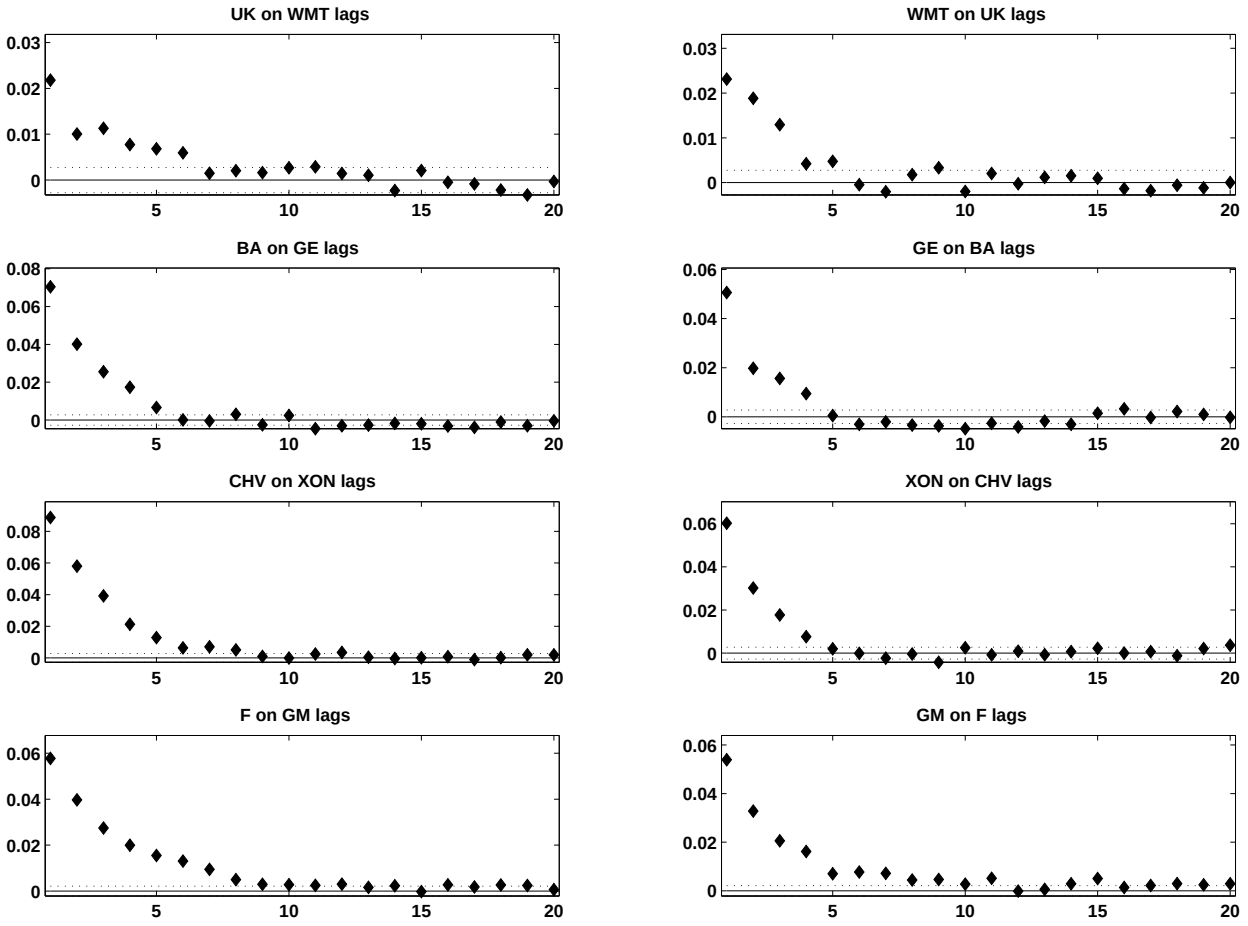

Figure 4: The cross correlations were constructed using 1 minute returns, and measure the correlation between the contemporaneous returns of one asset and the lagged returns of the other. All series evidence positive cross-correlations, although the intra-industry pairs of XON-CHV and F-GM exhibit more time dependence. Specifically, none of the 20 cross correlations for either F-GM pairing is negative while most are statistically different from zero.

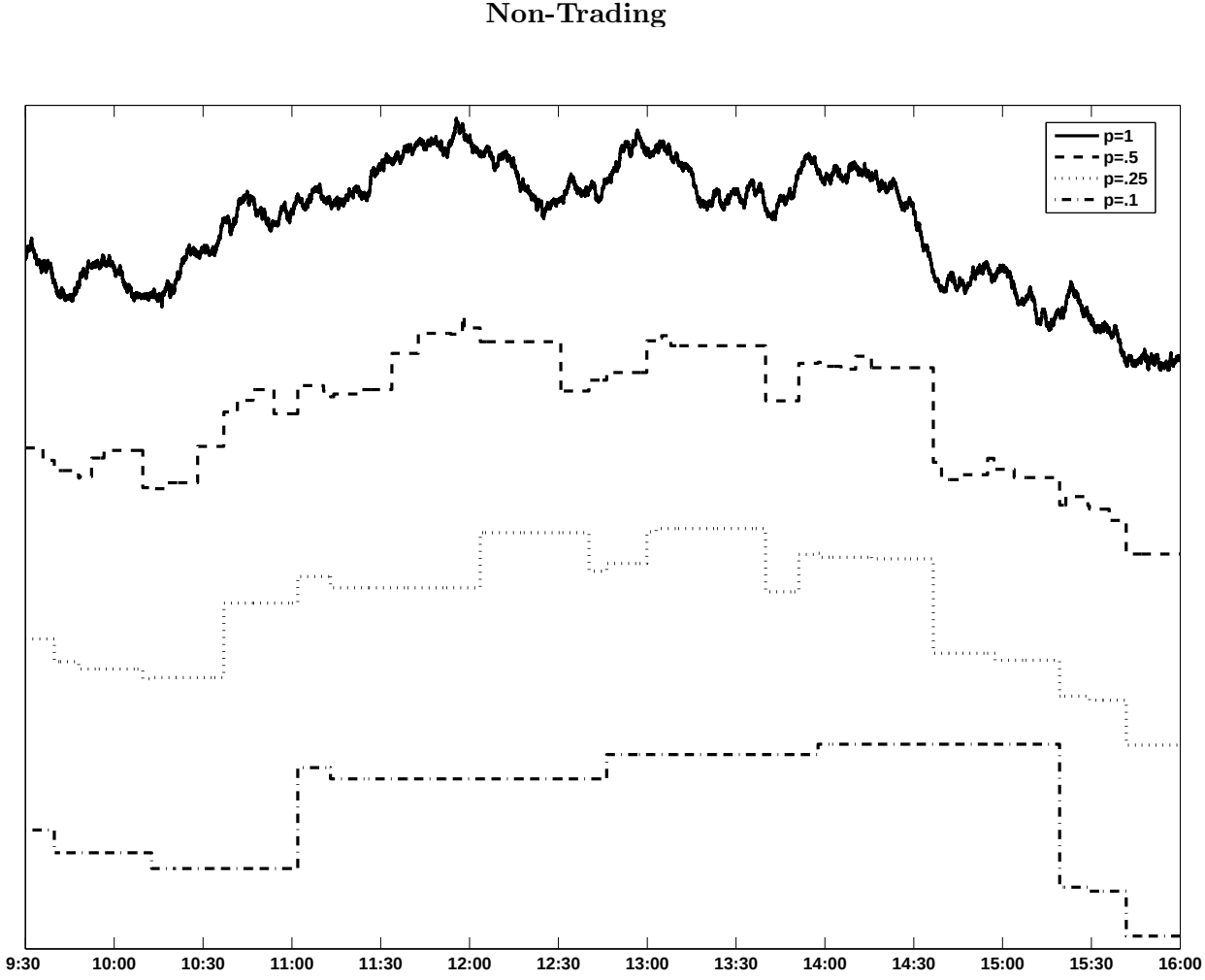

Figure 5: These four series show the evolution o prices as the probability of no trade using a 5-minute windows decreases from 1 through 5. and .25 to .1. All series were constructed using the same random data. The variance of the daily return was set to be 1.

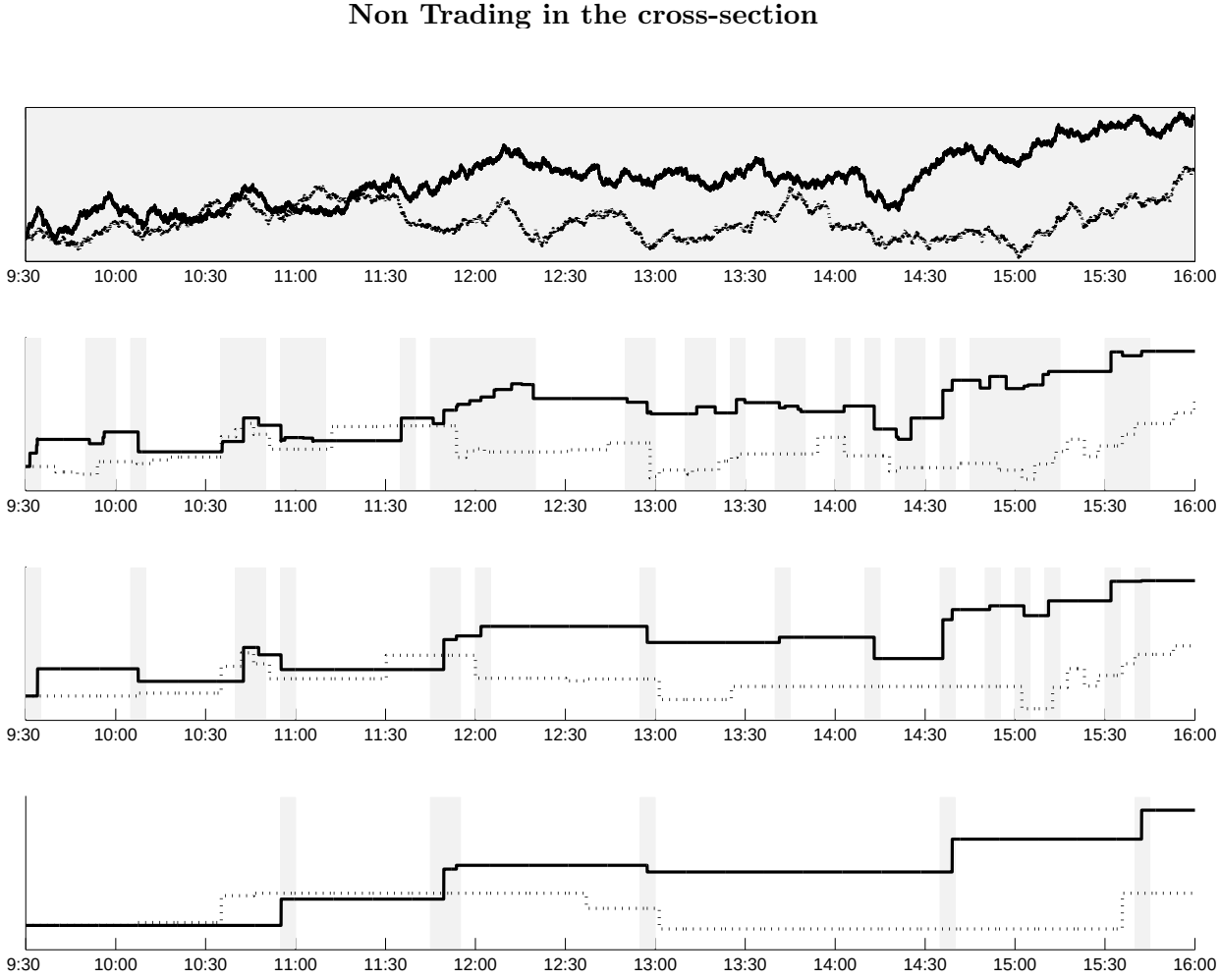

Figure 6: The four figures consider show the behavior of prices as the probability of observing a trade in a 5-minute windows decreases from 1 (top panel) thorough .5 and .25 (the sample average of DJIA stocks) and finally .1. Grey shaded areas 5-minute periods where both assets had a return. In the top panel, all windows contain trades, while in the bottom, only 6 of 78 periods contain new prices of both stocks. The variance of open to close return for each series was set to 1 with a correlation of .5.

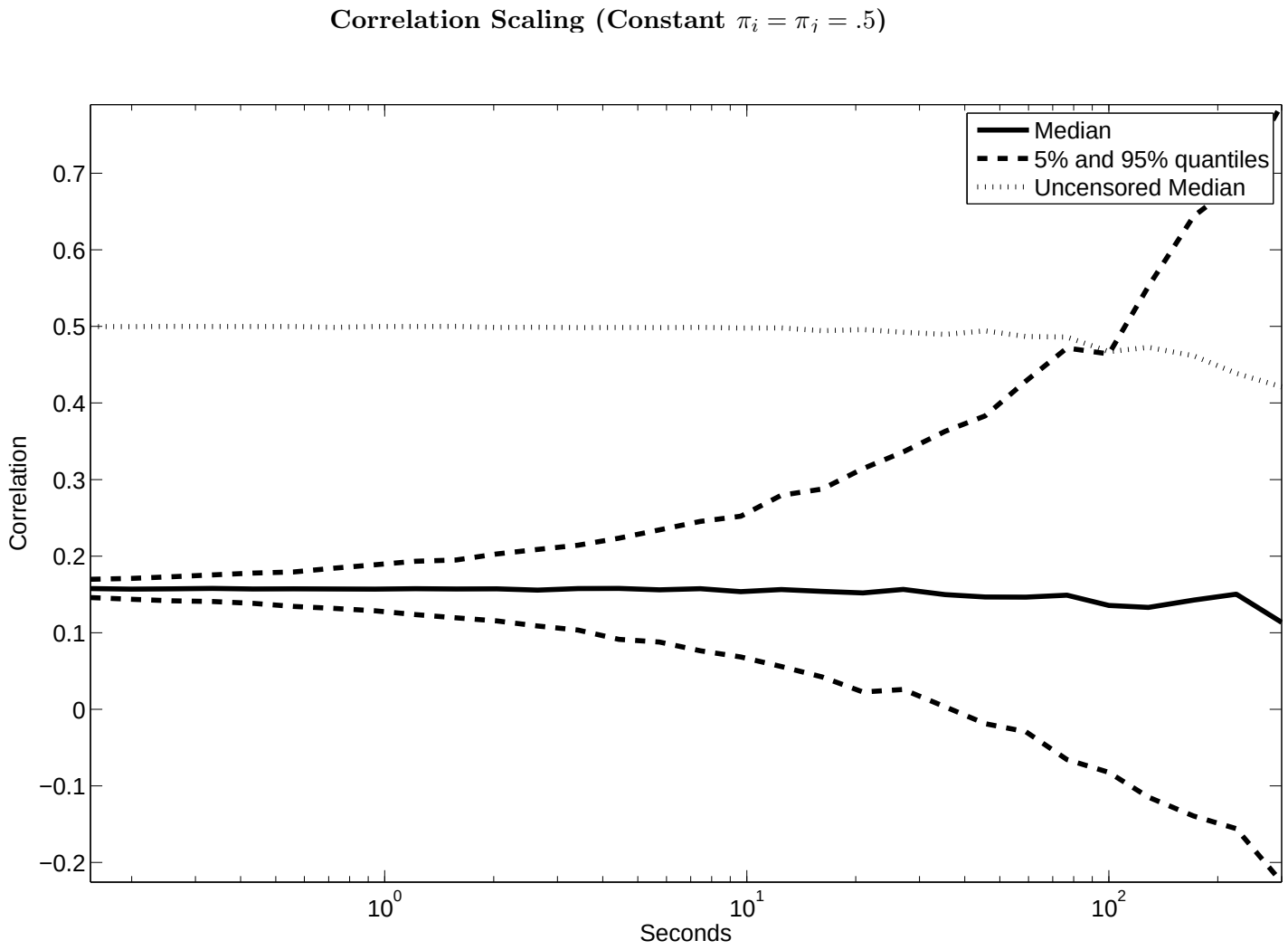

Figure 7: Realized correlation measured at various sampling frequencies from 1 second to 1/2 hour. Prices were simulated from a pair of correlated Brownian motions ($\sigma_{iis} = \sigma_{jjs} = \frac{1}{m}$ and $\sigma_{ij} = \frac{.5}{m}$). The probability that the observed price corresponds to the efficient price at any sample was 0.5. The median correlation is biased for any sampling frequency by $\frac{(1-0.5)(1-0.5)}{1-0.5^2}$, however the bias is not increasing in the number of samples.

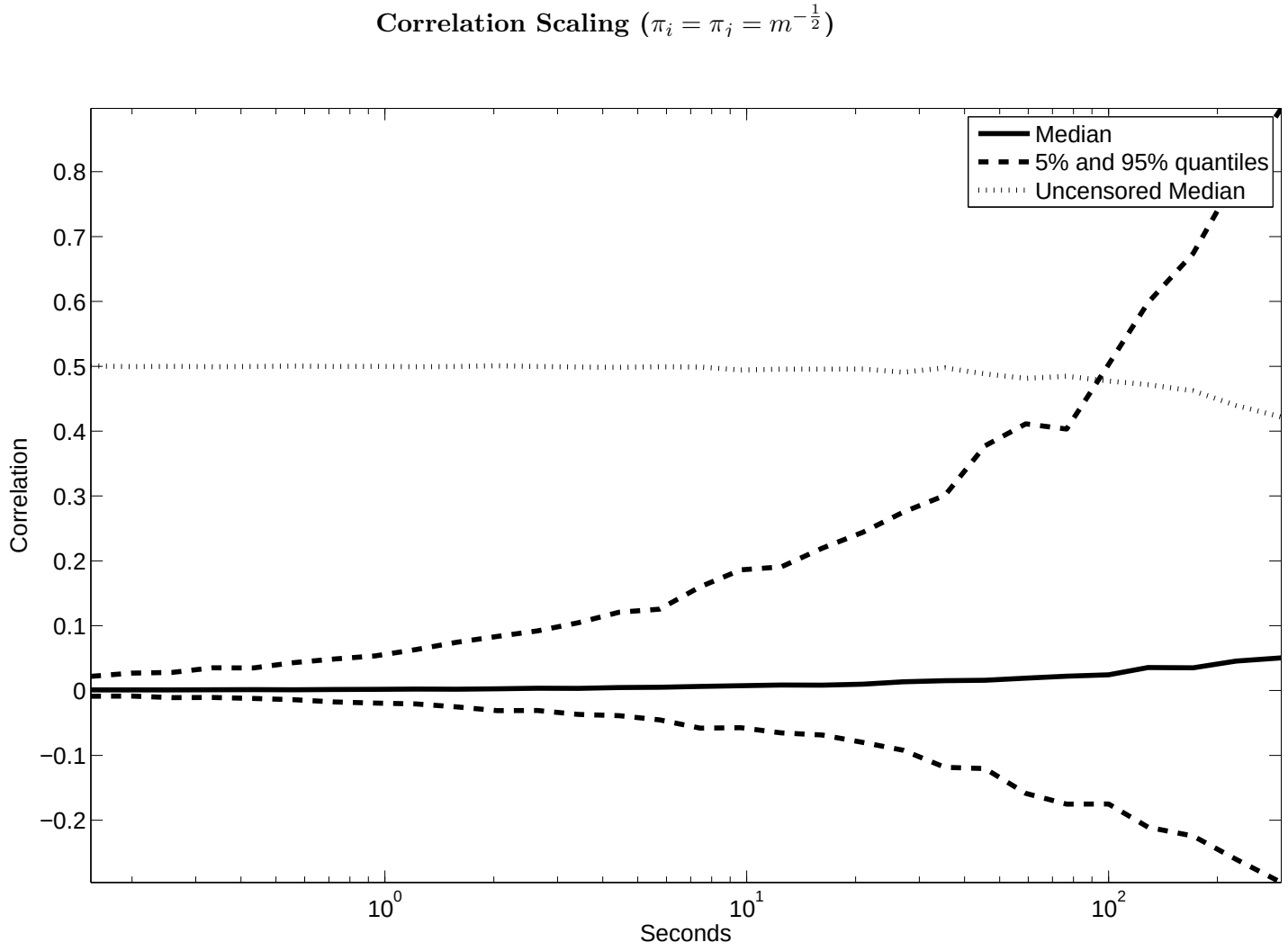

Figure 8: Realized correlation measured at various sampling frequencies from 1 second to 1/2 hour. Prices were simulated from a pair of correlated Brownian motions ($\sigma_{iis} = \sigma_{jjs} = \frac{1}{m}$ and $\sigma_{ij} = \frac{.5}{m}$). The probability that the observed price corresponds to the efficient price at any sample was $m^{-\frac{1}{2}}$ and the last efficient price was always correctly recorded. The median correlation is biased for any sampling frequency by $\frac{(1-0.5)(1-0.5)}{1-0.5^2}$, and the bias is clearly increasing in the number of samples.

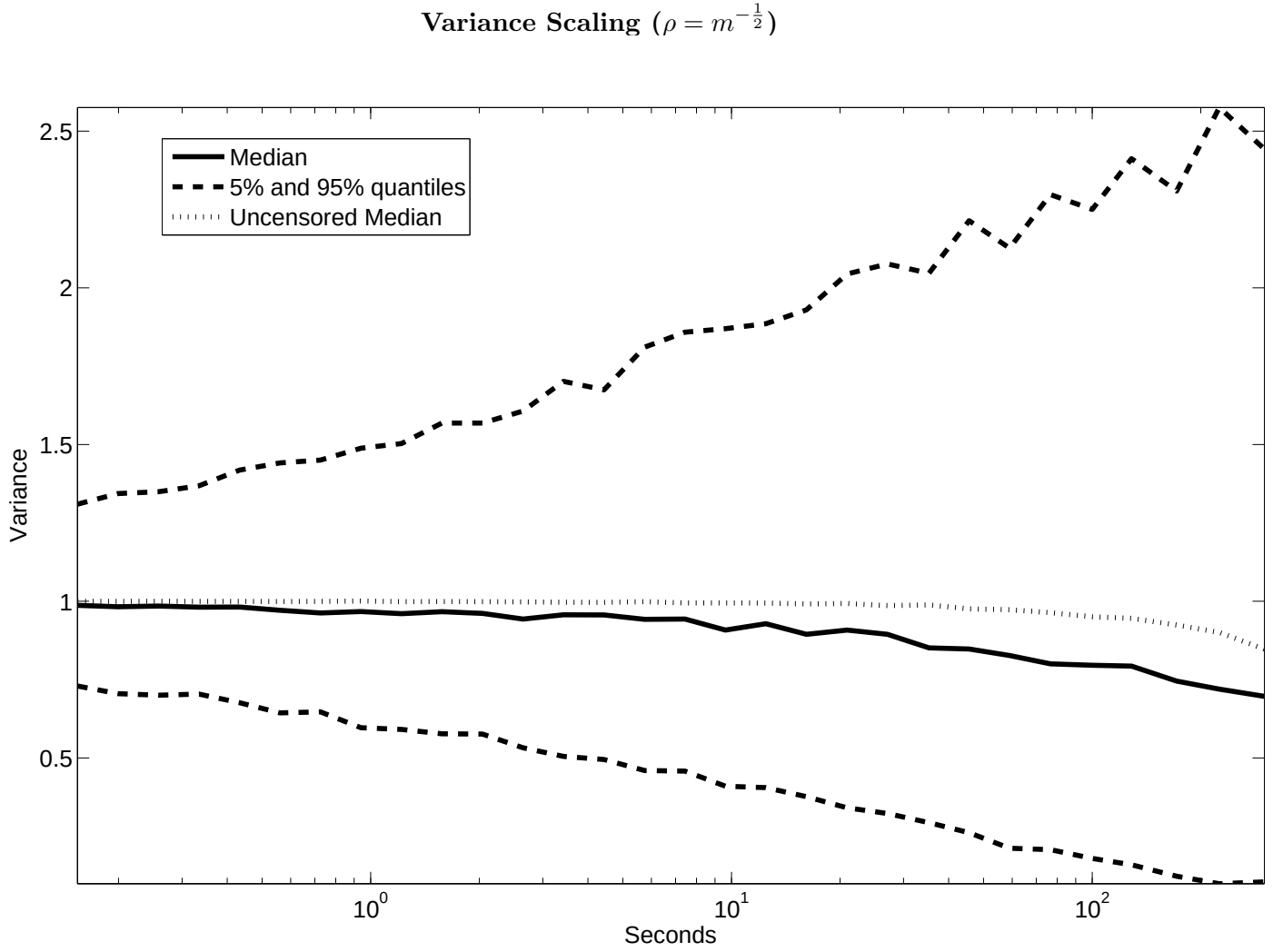

Figure 9: Realized variance measured at various sampling frequencies from 1 second to 1/2 hour. Prices were simulated from a pair of correlated Brownian motions ($\sigma_{iis} = \sigma_{jjs} = \frac{1}{m}$ and $\sigma_{ij} = \frac{.5}{m}$). The probability that the observed price corresponds to the efficient price at any sample was $m^{-\frac{1}{2}}$ and the last efficient price was always correctly recorded. The median variance is slightly biased for any sampling frequency due to right skew in the distribution of realized variance. It is mean unbiased at any sampling frequency.

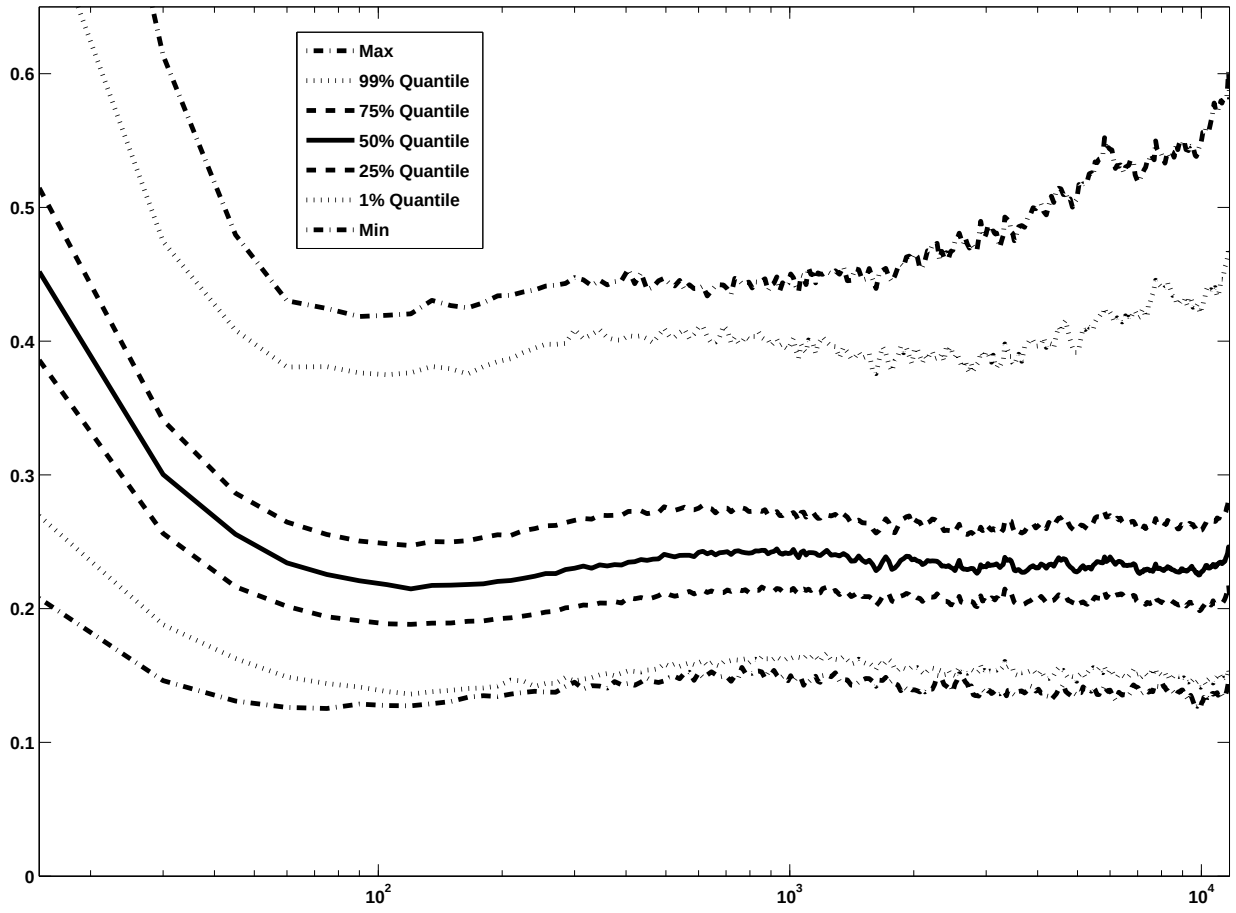

Figure 10: Distribution of realized (pseudo) correlations when debiased assuming a constant (throughout the day and across days)censoring rate. Returns were sampled from 1 second to 20 minutes. $\hat{\pi}_i$ and $\hat{\pi}_j$ were computed form the frequency of intervals with an informative quote (either the bid price, the ask price or both must change). For a large range of sampling frequencies, the distribution is fairly unchanged. The two potential issues with this model come from (a) the large upturn when sampled too frequently and (b) the constant increase for the pseudo correlations for the top 1% of correlation pairs. The large upturn when sample too frequently is due to the overnight covariance which was aligned in most samples. The continued increase for certain (usually intra-industry) pairs is an unresolved mystery.