Robust Estimation of Realized Correlation

Yiyao Luo[a]

[a] *University of North Carolina at Chapel Hill* *

November 11, 2022

Abstract

We introduce a new correlation estimator that combines the Quadrant correlation estimator with subsampling. Unlike the well-known Pearson and Kendall correlation estimators, the subsampled Quadrant estimator is consistent under time-varying volatilities and more precise than the Quadrant without subsampling. When we estimate correlations of high-frequency financial data, the subsampled Quadrant compares favorably to alternative estimators. The subsampled Quadrant estimator is robust to the microstructure noise, rounding error, asynchronous trading, and jumps found in financial data. On implementing the new estimator with actual high-frequency stock transaction data, we illustrate distinct patterns of correlation estimates along different sampling frequencies. Also, we propose a new approach that embraces the subsampled Quadrant to estimate intra-day market betas and find new insights about the intra-day variance in market betas for 22 stocks.

Keywords: Correlation, Subsampling, Robustness, Consistency, Epps effect, High-frequency data, Microstructure, Jump

*Address: University of North Carolina, Department of Economics, 107 Gardner Hall Chapel Hill, NC 27599-3305

1 Introduction

The correlation coefficient measures the linear association between two series of data. Normally, when you intend to fit your data with a multivariate distribution model, estimating the correlation parameter is necessary. Even in the simplest linear regression model, the linear correlation between explanatory and dependent variables serves essentially as a premise. In modern finance, the correlation between asset returns is indispensable. The covariation based on correlation and volatility is the key statistic in risk management, portfolio selection, hedging and pricing derivatives, etc. Many recent financial innovations have been built to utilize the correlation structure of two or more assets. Additionally, in empirical finance, a market beta is simply the correlation between the asset and market returns multiplied by the relative volatility of asset returns to market volatility. High-frequency financial data are widely available and facilitate accurate estimation of variances, covariances, and correlations over a short period, such as a day or an hour. The new estimator is useful for this purpose because it is robust to the prevalent time variation in volatility.

The correlation is typically estimated by the well-known sample correlation, which is based on estimates of the covariance and two variances. This estimator is also known as the Pearson correlation estimator. In the context of high-frequency data, the realized correlation is computed from realized measures of covariances and variances, such as the realized 2×2 covariance matrix, see Andersen et al. (2003), Barndorff-Nielsen and Shephard (2004*a*), and Zhang et al. (2005). A key characteristic of high-frequency financial data is the presence of market microstructure noise: prices are observed with errors due to the bid-ask bounces, discretization of prices (defined by the tick size), asynchronous tradings. Thus, there is a discrepancy between the latent efficient price and the observed price, which can severely erode the accuracy of non-robust estimators and distort inference. One example was documented by Epps (1979), who showed that the sample correlation declines as prices are sampled at high frequencies, cause by a lack of synchronicity in the observed prices. Time-varying volatility within the trading day, see Andersen et al. (2019), is another feature that distorts the standard (Pearson) realized correlation, and so do jumps in the prices. The latter can be circumvented by using a truncation estimator, see Raymaekers and Rousseeuw (2021), but correlation estimation in the presence of time-varying volatilities requires a new alternative estimator. We explore many classical robust correlation estimators based on ranks and the frequency of sign concordance. Among these estimators, the quadrant estimator emerges as the only robust estimator. Unfortunately, the quadrant estimator is very inefficient.

The Quadrant probability estimator measures the probability that a pair of bivariate random variables both are positive or negative, so-called sign concordance. Rank concordance is defined on two pairs of bivariate random variables. Compared to the first pair, the Kendall probability estimator measures the chance that the second selected pair is higher or less on both sides. Under the premise that variables are normally distributed or mixed-normally distributed, we can map the probabilities of sign concordance as well as rank concordance to the correlation coefficient of the bivariate variables through a closed-form identity. Not only the simplicity of computation attracts us, but also the robustness of Quadrant and Kendall estimators leads us to consider them as alternatives to Pearson. The robustness here is defined by their bounded influence functions that capture the effect of adding a tiny fraction of outliers into the sample. The limitation of Pearson while facing microstructure issues and jumps can be reasoned by its unbounded influence function.

Nonetheless, neither the Quadrant nor the Kendall is as precise as the Pearson in the ideal situation where observations are normally distributed without contamination. Kendall is more efficient than Quadrant if we overlook the boosting cost of computation. Yet, Kendall embraces a similar defect as Pearson when the bivariate observations are drawn from models with time-changing variances. After all, we favor the Quadrant estimator based on its simplicity, robustness, and invariance of time-varying variances. To improve the accuracy of the Quadrant in a finite high-frequency financial sample, we come up with a subsampled Quadrant estimator. The subsampled Quadrant is an extension of the Quadrant estimator using the subsampling approach proposed by Zhang et al. (2005). The subsampling approach helps the Quadrant estimator gain more information within the finite sample, as shown in the elevation of efficiency and comparison of precision in the later simulation study.

The main contribution of this paper is a new robust correlation estimator that combines the resiliency of the Quadrant estimator with subsampling to enhance efficiency greatly. We show that the subsampled Quadrant estimator is consistent under time-varying volatilities and derive its asymptotic distribution under some assumptions. Then we disclose the robustness of the subsampled Quadrant in a series of simulation studies in which a variety of sources of microstructure contaminations are included. Empirically, we apply those estimators of interest to the real data from NYSE Trade and Quote (TAQ) database and illustrate the reverse Epps effect on the Quadrant and subsampled Quadrant estimators. Besides, we utilize the subsampled Quadrant estimator and others to estimate the intra-day market betas and to explore the diverse changing patterns of betas.

1.1 Literature

Our paper contributes to the literature on high-frequency financial econometrics and nonparametric correlation estimators. For robust correlation estimation in the high-frequency financial framework, much effort has been put into unraveling and alleviating the Epps effect of Pearson. The main features of high-frequency financial data that contribute to the Epps effect include lead-lag, discretization, and asynchrony (see, e.g., Renò (2003), Precup and Iori (2007), and Münnix et al. (2011)). Tóth and Kertész (2007) characterized the effect in an analytical expression as a function of the rate parameter from Poisson sampling and further extended it by decomposing the correlation at a certain time scale as a function of the correlation at smaller scales in Tóth and Kertész (2009). Mastromatteo et al. (2011) developed the expression by separating the effect of asynchrony and that of lead-lag. The Epps effect under three sampling schemes is compared in Chang et al. (2021), where they find correlations emerge faster under event time than calendar time while correlations emerge linearly under volume time.

Another strand of literature focuses on correcting covariance and variance estimators, which can be adapted to help Pearson. The studies can be classified by the microstructure issues they are tackling. For microstructure noise, researchers provided solutions such as sparse sampling (Andersen et al. (2001) and Bandi and Russell (2008)), multi-scale estimators by Zhang et al. (2005), pre-averaging estimators (Podolskij and Vetter (2009), Jacod et al. (2009), Christensen et al. (2010) and Christensen et al. (2013)), quasi-maximum likelihood estimator from Aït-Sahalia et al. (2010), and measures based on the Kalman filter (Corsi et al. (2015)) and the EM algorithm (Shephard and Xiu (2017)). The canonical estimator is introduced by Hayashi and Yoshida (2005) to address the asynchrony in observed price processes, which is further extended to be unbiased in the presence of noise (Griffin and Oomen (2011)) and lagged correlations (Voev and Lunde (2007)). Zhang (2011) extended two-scale RV to integrated covariation estimation in the simultaneous presence of noise and asynchronous trading. Barndorff-Nielsen et al. (2011) concurrently proposed a multivariate realized kernel covariance estimator. Additional work in this stream of research includes Malliavin and Mancino (2002), Martens (2004), Bandi and Russell (2005), and Boudt et al. (2011). Studies on estimation with rounded prices can be traced back to Delattre and Jacod (1997). More recent seminal work includes Rosenbaum (2009), Li and Mykland (2015), and Li et al. (2018). Last but not least, the presence of jumps in asset prices also introduces bias in covariance measurements. The bipower variation and covariation estimators are proposed in Barndorff-Nielsen and Shephard (2004*b*) and Barndorff-Nielsen and Shephard (2007). Other

seminal work includes thresholds covariances in Mancini and Gobbi (2012), the outlyingness weighted covariances in Boudt et al. (2011), and disentangled covariances in Boudt et al. (2012*b*).

Researchers put forward various alternatives to Pearson in the literature on nonparametric correlation estimators. Croux and Dehon (2010) reviewed nonparametric correlation measures, and derived the influence functions and gross-error sensitivities of Quadrant, Kendall, and Spearman's correlation measures, showing that these nonparametric measures are robust to the existence of outliers in the data. Further comparative analysis between Kendall's and Spearman's correlation measures reveals that the former measure is superior to the latter in terms of mean squared errors for contaminated normal distribution and that a mixture of two measures should be applied for unbiasedness, see Xu et al. (2013). Another comprehensive overview of robust correlation measures supported that correlation measures constructed via robust principal variables are more accurate than the Pearson estimator in the framework of contaminated normal distribution, see Shevlyakov and Smirnov (2011). Boudt et al. (2012*a*) studied the breakdown point and influence function of the Gaussian rank correlation measure and stated that it is less affected by a small number of outliers but as efficient as Pearson's measure. Raymaekers and Rousseeuw (2021) generalized the Gaussian rank correlation measure and proposed a class of correlation measures built on robust g-product moments, in which the Quadrant and Spearman measures are also included. Last but not least, a robust correlation measure can be attained by a robust covariance estimator such as the minimum covariance determinant method (see Rousseeuw (1984)) and used as a benchmark estimator in comparative studies, see Croux and Dehon (2010), Boudt et al. (2012*a*), and Raymaekers and Rousseeuw (2021).

Consequently, robust correlation measures are appealing for intra-day covariance estimation, especially when jumps and asynchronous tradings contaminate high-frequency prices. However, the literature on econometrics for high-frequency data has not paid much attention to those alternative measures. Boudt et al. (2012*b*) proposed to estimate the covariance of prices with jumps via disentangling it into robust realized volatility and Gaussian rank correlation measure. Vander Elst and Veredas (2016) developed a broader comparison, including Quadrant, Kendall, and Spearman correlation measures.

1.2 Organization of Paper

This paper is organized as follows. Section 2 reviews the benchmark correlation estimators and introduces the subsampled Quadrant estimator. In section 3, we present the properties of the estimators, including efficiency, consistency, and robustness. Section 4 reports the results of a series of simulation

studies based on the Levy and Heston model adding prevailing microstructure issues and jumps. The empirical illustrations are presented in section 5. We extend the correlation estimation of bivariate variables to the higher dimensional correlation matrices in section 6. Section 7 concludes.

2 Correlation Estimators

Before introducing the new estimator, we review the classical, Pearson, Quadrant, and Kendall estimators and discuss a few additional estimators. We evaluate the relative merits of these estimators, especially in a context with noisy high-frequency financial data. Under ideal conditions, the Pearson estimator is the most efficient; however, in the context of high-frequency data, the Pearson estimator is inconsistent under realistic assumptions. The Quadrant and Kendall estimators are robust estimators that bypass the need to estimate variances and covariance. The Kendall estimator dominates the Quadrant estimator in precision, with some exceptions, but is more computationally demanding. The new estimator combines the simplicity of the Quadrant estimator with methods that mitigate the effects of market microstructure noise.

2.1 Classical Correlation Estimators: Pearson, Quadrant, and Kendall

Let (x_i, y_i), $i = 1, \ldots, n$ be a sample of bivariate random variables centered at zero. The Pearson correlation estimator is given by

$$P = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}}.$$

It is asymptotically efficient if the data are normally distributed. A drawback of the Pearson estimator is that it is relatively sensitive to outliers, which can be understood from its influence function, which is unbounded, see Devlin et al. (1975).

The Quadrant and Kendall estimators are based on a link between the correlation and the concordance-probability $q = \Pr[XY > 0]$. For a bivariate normal distribution with zero means we have the following identity:

$$q = \Pr[XY > 0] = \frac{\arcsin \rho}{\pi} + \frac{1}{2}, \tag{1}$$

where $\rho = \mathrm{corr}(X, Y)$, see Blomqvist (1950). Observe that the one-to-one mapping between q and ρ

is invariant to the variances of X and Y, such that (1) generally holds for a class of mixed Gaussian distributions, including the multivariate t-distribution. In fact, the identity (1) is valid for all symmetric elliptical distributions with an existing second moment, see a proof in Appendix. In this paper, we rely on (1), but an alternative link function could easily be used instead based on a non-elliptical distribution.

The Quadrant estimator is based on the sample estimate of q,

$$\hat{q}_Q = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{x_i y_i > 0\}},$$

where $\mathbf{1}_{\{.\}}$ represents the indicator function. Relying on (1), the corresponding estimate of the correlation is given by $Q = \rho(\hat{q}_Q)$, where

$$\rho(q) = \sin(\pi(q - \frac{1}{2})), \tag{2}$$

see Mosteller (1946).

The Kendall estimator involves two pairs of observations, (X_1, Y_1) and (X_2, Y_2). Because $\rho = \text{corr}(X_1, Y_1) = \text{corr}(X_2, Y_2)$ and the two pairs are independent we also have $\rho = \text{corr}(X_1 - X_2, Y_1 - Y_2)$, and this motivates the correlation estimator,

$$K = \rho(\hat{q}_K), \qquad \text{where} \quad \hat{q}_K = \frac{2}{n(n-1)}\sum_{i<j} \mathbf{1}_{\{(x_i - x_j)(y_i - y_j) > 0\}},$$

see Kendall (1938).

We want to highlight three advantages of the Quadrant and Kendall estimators over the Pearson estimator. First, Q and K are more robust to outliers, especially the large ones, because their influence functions are bounded, see Croux and Dehon (2010). Both Q and K are also less sensitive to outliers drawn from an entirely different distribution (with a small probability). Second, being constructed from simple binary variables has a computational advantage and then practical importance in a large-scale setting. Third, the Quadrant estimator is robust to time-varying volatilities, which is only partially true for the Kendall estimator.

In addition to the Quadrant and Kendall estimators, there are other commonly used nonparametric correlation estimators in literature, such as the Spearman rank correlation estimator, seeSpearman (1904); Moran (1948), and the Gaussian rank correlation estimator, see Boudt et al. (2012a). The

Spearman estimator also has a bounded influence function, but its bound is broader than that for Q and K, which shows it can be more sensitive to outliers. In fact, the Spearman estimator computes the Pearson correlation of the ranks instead of the concordance. The Gaussian rank estimator has an unbounded influence function as the Pearson estimator and is as efficient as the latter, see Boudt et al. (2012a). One can find comparisons between it and the Quadrant, Kendall, and Pearson estimators in the framework of high-frequency financial data in Vander Elst and Veredas (2016) and Boudt et al. (2012c). This paper will not discuss or include them in the following comparisons.

2.2 Subsampled Quadrant Correlation Estimator

We construct our new estimator by taking advantage of the fact that the relationship in (1) is not affected by marginal variances of the underlying distribution, and we make a simple extension of the Quadrant probability estimator :

$$\hat{q}_S = \frac{1}{n-S+1} \sum_{i=1}^{n-S+1} \mathbf{1}_{\{\sum_{j=0}^{S-1} x_{i+j} \sum_{j=0}^{S-1} y_{i+j} > 0\}}.$$

We call it by subsampled Quadrant estimator, analogous to the subsampled realized variance proposed by Mykland. The above estimator applies the Quadrant probability estimator to all consecutive sub-series of length S for $1 < S < n$. It shares similar ideas with the pre-averaging estimators and the moving block method in stationary bootstrap, see Künsch (1989); Liu et al. (1992), and variance estimators based on averaging sub-series, which are asymptotically equivalent to the Bartlett spectral density estimator (Bartlett (1946, 1950)), as mentioned in Politis and White (2004).

Note that $\hat{q}_S$ is a consistent estimator for quantity q, then we derive a corresponding consistent correlation estimator at assuming normal models as above:

$$Q_S = \sin(\pi(\hat{q}_S - \frac{1}{2}))$$

and call it subsampled Quadrant correlation estimator.

The subsampled Quadrant correlation estimator is appealing in several ways. First, the subsampled Quadrant estimator is simpler than the Kendall estimator to compute at the same convergence rate. That is an attractive advantage when we run daily correlation estimations on high-frequency financial data. Next, the subsampled Quadrant estimator, as an extension of the Quadrant estimator, inherits the latter's influence function and then keeps its robustness against outliers. The difference between the

subsampled Quadrant and the Quadrant estimators applied to sub-series of length S is that the former uses all possible consecutive sub-series and the latter only uses non-overlapping sub-series, making the former more accurate in a finite. Last but not least, the subsampled Quadrant estimator naturally fits in the framework of high-frequency financial data. When every pair of observations is a pair of one-second log asset returns, the subsampled Quadrant applied to all consecutive S-second log returns follows the spirit of the sparse sampling method.

3 Properties of Estimators

3.1 Efficiency

Combined with volatility estimators from high-frequency data, sparse sampling is commonly used to alleviate the distortions arising from market microstructure noise. For example, the realized variance estimated based on 5-minute returns is far more widely used than the one at higher frequencies. Sparse sampling helps by increasing the signal-to-noise ratio, which is critical for covariance estimation as well.

In this way, we will compare the asymptotic properties of the above correlation estimators associated with the sparse sampling method. Here, Q, K, and P are based on sparsely sampled observations, whereas Q_S is constructed from all consecutive sub-series that also construct sparsely sampled observations. S represents the frequency by which we construct sparsely sampled observations. Then, from $N+1$ price observations, we have $n = \lfloor N/S \rfloor$ pairs of intraday returns for Q, K, and P and $N-S+1$ pairs of returns for Q_S. With any fixed S, all estimators' convergence rates are characterized by n.

Proposition 1. *Under assumption of normality with mean zero and correlation coefficient ρ, asymptotically, we have the following results with fixed S,*

$$\begin{aligned}
\sqrt{n}(P-\rho) &\xrightarrow{d} N(0,V_P), \qquad \textit{with} \quad V_P = (1-\rho^2)^2, \\
\sqrt{n}(K-\rho) &\xrightarrow{d} N(0,V_K), \qquad \textit{with} \quad V_K = (1-\rho^2)(\tfrac{\pi^2}{9} - 4\arcsin^2(\tfrac{\rho}{2})). \\
\sqrt{n}(Q-\rho) &\xrightarrow{d} N(0,V_Q), \qquad \textit{with} \quad V_Q = (1-\rho^2)(\tfrac{\pi^2}{4} - \arcsin^2\rho).
\end{aligned}$$

The proofs for the Quadrant and Kendall estimators are provided in Croux and Dehon (2010). Next, we state the corresponding for Q_S.

Theorem 1. *Under assumption of normality with mean zero and correlation coefficient ρ, with fixed k, the Bartlett-Quadrant correlation estimator is asymptotically normally distributed*

$$\sqrt{n}(Q_S - \rho) \xrightarrow{d} N(0, V_{Q_S})$$

and moreover, its asymptotic variance is a function of both ρ and S

$$V_{Q_S} = (1 - \rho^2)\left[\frac{\frac{\pi^2}{4} - \arcsin^2\rho + 2\sum_{s=1}^{S-1}\arcsin^2(\frac{s}{S}) - \arcsin^2(\rho\frac{s}{S})}{S}\right].$$

See Appendix for proof.[1]

It is not surprising that the additional information will improve the efficiency of the Q_S estimator relative to Q.

For $S = 1$ we obviously have $V_{Q_S} = V_Q$ because Q_S simplifies to Q in this case. More generally, for $S > 1$, we can compare the asymptotic variance of the new estimator as functions with that of other estimators. The asymptotic variances depend on ρ and are shown in Figure 1. In the figure, we omit the Quadrant estimator's asymptotic variance because it has the lowest asymptotic efficiency among Kendall, Pearson, and itself. Moreover, the subsampled Quadrant estimator is more accurate than the Kendall estimator over small ρ's after we raise S to 5 and is at least as efficient as Kendall across all ρ's when S is 15 or higher. Compared with the Pearson estimator, the subsampled Quadrant estimator is more efficient, with S being chosen as 30 or higher when the actual correlation is less than 0.5. As ρ approaches 1, the Kendall and subsampled Quadrant estimators' asymptotic variances are almost the same when S is high enough, but they are substantially more significant than the one of the Pearson. It is worth mentioning that, in the framework where a single observation stands for a one-second log return, the subsampled Quadrant estimator is asymptotically superior to Kendall on log returns sampled at 15 seconds, half a minute, and 15 minutes.

[1]Note that for $S = 1$ we have $\frac{\sum_{s=-S+1}^{S-1}\arcsin^2(\frac{S-s}{S}) - \arcsin^2(\rho\frac{S-s}{S})}{S} = \frac{\pi^2}{4} - \arcsin^2\rho$ such that $V_{Q_S} = V_Q$ as expected.

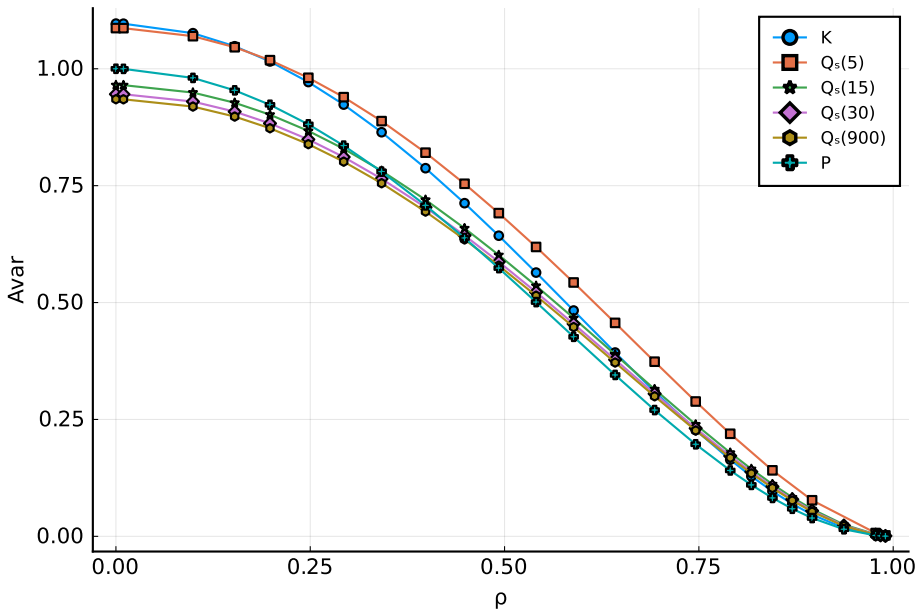


Figure 1: Plot asymptotic variances of the Kendall (K), Pearson (P), and subsampled Quadrant estimators with different choices of sub-series length S.

3.2 Influence Function

The influence function evaluates the infinitesimal effect on a statistical functional from a tiny amount of contamination placed on the underlying distribution. Specifically, let R be a statistical functional, then the influence function (IF) of R at a distribution H is defined as

$$\mathrm{IF}((x_0, y_0), R, H) = \lim_{\varepsilon \downarrow 0} \frac{R((1-\varepsilon)H + \varepsilon \boldsymbol{\Delta}_{(x_0,y_0)}) - R(H)}{\varepsilon}$$

where $\boldsymbol{\Delta}_{(x_0,y_0)}$ is a Dirac measure putting all its mass at point (x_0, y_0). Denote the bivariate normal distribution with zero mean, unit variance, and correlation coefficient ρ, by Φ_ρ. For the Quadrant and Kendall estimators, we have

$$R_Q(\Phi_\rho) = \sin\left(\pi(\mathrm{Prob}[XY > 0|\Phi_\rho] - \frac{1}{2})\right) \quad \text{and} \quad R_K(\Phi_\rho) = \sin\left(\pi(\mathrm{Prob}[(X_1 - X_2)(Y_1 - Y_2) > 0|\Phi_\rho] - \frac{1}{2})\right).$$

For the Pearson estimator, we have

$$R_P(\Phi_\rho) = \rho.$$

Proposition 2. *The influence functions of correlation estimators at Φ_ρ are given by*

$$\begin{aligned}
\mathrm{IF}((x_0,y_0),R_Q,\Phi_\rho) &= \pi\sqrt{1-\rho^2}(\mathbf{1}_{\{x_0y_0>0\}}-q)\\
\mathrm{IF}((x_0,y_0),R_K,\Phi_\rho) &= 2\pi\sqrt{1-\rho^2}(2\Phi(x_0,y_0)+1-\Phi(x_0)-\Phi(y_0)-q)\\
\mathrm{IF}((x_0,y_0),R_P,\Phi_\rho) &= x_0y_0-(\frac{x_0^2+y_0^2}{2})\rho
\end{aligned}$$

where $\Phi(\bullet,\bullet)$ and $\Phi(\bullet)$ are the joint cumulative density function and marginal cumulative density function of Φ_ρ.

The proofs are given in Devlin et al. (1975) and Croux and Dehon (2010). It is worth noting that only Pearson has an unbounded influence function, so it is viewed as not robust when outliers contaminate the data-generating process. The outliers discussed here could be some points far away from the data generated by the underlying distribution or some observations generated by another distribution.

In this paper, we focus on applying these estimators to sparse-sampled observations in the high-frequency financial data framework. Specifically, we consider the observations $(\tilde{X}_i,\tilde{Y}_i)$ over a window of length S instead of (X_i,Y_i), where the former is defined as

$$\tilde{X}_i=\sum_{j=i}^{i+S-1}X_j \quad \text{and} \quad \tilde{Y}_i=\sum_{j=i}^{i+S-1}Y_j.$$

Therefore, we should update the influence functions of these estimators for the sparse-sampled observations.

Proposition 3. *The influence function of Pearson estimator at Φ_ρ is given by*

$$\mathrm{IF}((x_0,y_0),R_P,\Phi_\rho)=x_0y_0-(\frac{x_0^2+y_0^2}{2})\rho.$$

The influence function of Kendall estimator at Φ_ρ is given by

$$\mathrm{IF}((x_0,y_0),R_K,\Phi_\rho)=2\pi\sqrt{1-\rho^2}S\Big[2\Phi(\frac{x_0}{\sqrt{2S-1}},\frac{y_0}{\sqrt{2S-1}})-\Phi(\frac{x_0}{\sqrt{2S-1}})-\Phi(\frac{y_0}{\sqrt{2S-1}})+1-q\Big].$$

The influence function of (subsampled) Quadrant estimator at Φ_ρ is given by

$$\mathrm{IF}((x_0,y_0),R_Q,\Phi_\rho)=\pi\sqrt{1-\rho^2}S\Big[2\Phi(\frac{x_0}{\sqrt{S-1}},\frac{y_0}{\sqrt{S-1}})-\Phi(\frac{x_0}{\sqrt{S-1}})-\Phi(\frac{y_0}{\sqrt{S-1}})+1-q\Big]$$

where $\Phi(\bullet,\bullet)$ *and* $\Phi(\bullet)$ *are the joint cumulative density function and marginal cumulative density function of* Φ_ρ.

See Appendix for proof. The Pearson estimator's influence function is invariant to the sparse sampling, while the sampling frequency determines the (subsampled) Quadrant and Kendall estimators' influence functions. Again, in contrast with Pearson, the bounded influence functions of non-parametric estimators guarantee robustness under data contamination.

3.3 Consistency

We note the asymptotic properties of correlation estimators by assuming that observations are independently and identically distributed. In the following discussion, we will relax this assumption and then compare the behaviors of those estimators. Specifically, we keep the independence and normality but allow observations to have varying variances, commonly used to characterize high-frequency financial data. The subsampled Quadrant estimator is constructed for estimating the correlation coefficient of such data, so it is necessary to investigate whether it still estimates the correct quantity in this scenario.

Constant variance is necessary for the consistency of Kendall and subsampled Quadrant estimators. We use a simple example to illustrate it.

Example 1. Consider observations $(x_i, y_i)_{i=1}^n$ with $x_i = \sigma_{xi}\tilde{x}_i$ and $y_i = \sigma_{yi}\tilde{y}_i$, where $(\tilde{x}_i, \tilde{y}_i)_{i=1}^n$ are independent and normally distributed with unit variances and correlation coefficient ρ. For Quadrant estimator, asymptotically we have

$$\hat{q}_Q \xrightarrow{P} \text{Prob}[\sigma_{xi}\tilde{x}_i\sigma_{yt}\tilde{y}_i > 0] = \text{Prob}[\tilde{x}_i\tilde{y}_i > 0] = q = \frac{\arcsin\rho}{\pi} + \frac{1}{2}$$

and $Q \xrightarrow{P} \rho$. Let $Z_i = \mathbf{1}_{\{\sum_{j=0}^{S-1} x_{i+j} \sum_{j=0}^{S-1} y_{i+j} > 0\}}$, $i = 1, \ldots, n-S+1$. Then

$$\begin{aligned}\mathbb{E}Z_i =& \text{Prob}[\sum_{i=1}^{S}\sigma_{xi}\tilde{x}_i \sum_{i=1}^{S}\sigma_{yi}\tilde{y}_i > 0] \\ =& \frac{1}{\pi}\arcsin(\frac{\rho\sum_{i=1}^{S}\sigma_{xi}\sigma_{yi}}{\sqrt{\sum_{i=1}^{S}\sigma_{xi}^2\sum_{i=1}^{S}\sigma_{yi}^2}}) + \frac{1}{2}.\end{aligned}$$

So the subsampled Quadrant estimator is not necessarily consistent to the correlation ρ. In a special case where (x_i, y_i) are student's t distributed, we have $Q_S \xrightarrow{P} \rho$ because of $\sum_{i=1}^{s}\sigma_{xi}\sigma_{yi} = \sqrt{\sum_{i=1}^{S}\sigma_{xi}^2\sum_{i=1}^{S}\sigma_{yi}^2}$. Analogously, the Kendall estimator is constructed based on random variables $Z_{ij} = \mathbf{1}_{\{(x_i-x_j)(y_i-y_j)>0\}}$

for $i \neq j$, and then

$$\mathbb{E}Z_{ij} = \frac{1}{\pi}\arcsin\left(\frac{\rho(\sigma_{xi}\sigma_{yi} + \sigma_{xj}\sigma_{yj})}{\sqrt{(\sigma_{xi}^2 + \sigma_{xj}^2)(\sigma_{yi}^2 + \sigma_{yj}^2)}}\right) + \frac{1}{2}.$$

Specially, the limit of the Kendall estimator is consistent when $\sigma_{xi}\sigma_{yi}+\sigma_{xj}\sigma_{yj} = \sqrt{(\sigma_{xi}^2 + \sigma_{xj}^2)(\sigma_{yi}^2 + \sigma_{yj}^2)}$ for any $i \neq j$.

Among three nonparametric estimators, the Quadrant estimator is naturally consistent, while the other two are only consistent under restrictive conditions. For instance, for each i, $\sum_{i=1}^{S}\sigma_{xi}\sigma_{yi} \approx \sqrt{\sum_{i=1}^{S}\sigma_{xi}^2\sum_{i=1}^{S}\sigma_{yi}^2}$ for small S, then the subsampled Quadrant estimator over every consecutive S observations might be consistent. Nonetheless, the Kendall estimator could still be potentially impacted even in that situation as it involves two pairs of observations with a relatively large time gap.

Noting that we are interested in the behaviors of the above estimators in the high-frequency framework, we investigate the conditions of the consistency asking in stochastic models that represent the processes of asset log prices. Let $(X_t)_{t\in[0,1]}$ be a bivariate Itô semimartingale process characterized by the equation

$$X_t = X_0 + \int_0^t a_u du + \int_0^t \sigma_u dW_u \tag{3}$$

where $a = (a^{(1)}, a^{(2)})$ is a locally bounded predictable drift function and

$$\sigma = \begin{pmatrix} \sigma^{(1)} & 0 \\ 0 & \sigma^{(2)} \end{pmatrix}$$

is a cadlag volatility process. Normally, we assume

$$\int_0^1 \sigma_u^2 du < \infty.$$

Here W denotes a 2-dimensional Wiener process and $\mathrm{Corr}(dW^{(1)}, dW^{(2)}) = \rho$. Specifically, X is defined on the filtered space $(\Omega, \mathcal{F}, (\mathcal{F})_{t\in[0,1]}, \mathbb{P})$, where all involved processes a, σ, and W are adapted.

The process X is observed at discrete times $t \in \{0, 1/T, \ldots, 1\}$. Denote the increment in X over interval $[(i-1)/T, i/T)$ by

$$\Delta_i^T X = X_{i/T} - X_{(i-1)/T}.$$

Then the subsampled estimator for the quantity $q = \frac{\arcsin\rho}{\pi} + \frac{1}{2}$ is defined based on all possible increments

over a time interval of length s/T:

$$\hat{q}_S = \frac{1}{T-S+1}\sum_{i=1}^{T-S+1}\zeta_i, \quad \text{where } \zeta_i = \mathbf{1}_{\{\sum_{j=0}^{S-1}\Delta_{j+i}^T X^{(1)}\sum_{j=0}^{S-1}\Delta_{j+i}^T X^{(2)}>0\}}. \tag{4}$$

Note that the increment in X is not necessarily centered at 0 because of the drift process a. However, a is not predictable, and its increment can be removed if we demean the observed increment in X. So, from now on, we take $a = 0$ for simplicity. We assume the smoothness of volatility in the following assumption and then state the consistency of the subsampled Quadrant estimator.

Assumption 1. *The cadlag process σ satisfies, for any $\varepsilon > 0$, there is $\delta > 0$ such that for any $t_1, t_2 \in [0,1]$ and $0 < t_2 - t_1 < \delta$*

$$\left| \frac{\int_{t_1}^{t_2}\sigma_u^{(1)}\sigma_u^{(2)}du}{\sqrt{\int_{t_1}^{t_2}\sigma_u^{(1)2}du\int\sigma_u^{(2)2}du}} - 1 \right| < \varepsilon.$$

Assumption 1 requires local smoothness of each volatility process and comparability of both processes. In other words, the change in two processes over a short time interval is of the same order. For example, if two polynomial functions in $[t_1, t_2)$ uniformly approximate σ for any $t_2 - t_1 < \delta$ and have the equal highest orders, then it is clear to see the result in the assumption is attainable.

Theorem 2. *Under the representation of X in (3) and Assumption 1, $\hat{q}_s$ defined in (4) converges in probability to the quantity q given fixed s. Moreover, $Q_S \overset{def}{=} \sin(\pi(\hat{q}_S - \frac{1}{2})) \xrightarrow{p} \rho$.*

Remark 1. The Kendall estimator considers the product of the difference in two pairs of increments $(\Delta_i^T X^{(1)} - \Delta_j^T X^{(1)})(\Delta_i^T X^{(2)} - \Delta_j^T X^{(2)})$, and its probability of being positive is analogous to the one in Example 1:

$$\begin{aligned}
&\text{Prob}[(\Delta_i^T X^{(1)} - \Delta_j^T X^{(1)})(\Delta_i^T X^{(2)} - \Delta_j^T X^{(2)}) > 0] \\
=&\frac{1}{\pi}\arcsin\left(\rho\frac{\int_{(i-1)/T}^{i/T}\sigma_u^{(1)}\sigma_u^{(2)}du + \int_{(j-1)/T}^{j/T}\sigma_u^{(1)}\sigma_u^{(2)}du}{\sqrt{(\int_{(i-1)/T}^{i/T}\sigma_u^{(1)2}ds + \int_{(j-1)/T}^{j/T}\sigma_u^{(1)2}du)(\int_{(i-1)/T}^{i/T}\sigma_u^{(2)2}ds + \int_{(j-1)/T}^{j/T}\sigma_u^{(2)2}du)}}\right) + \frac{1}{2}.
\end{aligned}$$

We need more restrictive conditions on volatility processes such that their integrals over time intervals are pretty similar to guarantee that the above probability converges to the same thing for all i and j. We also need comparability between two processes as in Assumption 1 to ensure the limit is

q. Without those conditions, Kendall estimates the actual correlation scaled by some quantity close to the averaged instant association between volatilities over $[0, 1]$.

Remark 2. In this scenario, the Pearson correlation estimator has the following limit

$$P \xrightarrow{p} \rho \times \frac{\int_0^1 \sigma_u^{(1)} \sigma_u^{(2)} du}{\sqrt{\int_0^1 \sigma_u^{(1)2} dt \int_0^1 \sigma_u^{(2)2} du}}.$$

To assure the consistency to ρ, condition $\int_0^1 \sigma_u^{(1)} \sigma_u^{(2)} du = \sqrt{\int_0^1 \sigma_u^{(1)2} du \int_0^1 \sigma_u^{(2)2} du}$ is needed. Here, the targeting correlation coefficient is the one between the bivariate Weiner process, while Pearson estimates that scaled by the association between the volatility processes.

3.4 Robustness

In the high-frequency financial framework, it is usual to consider the contamination of the observations from the current microstructure issues, such as noise, rounding error, non-synchronous trading, and jumps. Such contamination indeed boosts the complexity of correlative estimation as described in Epps (1979). Nevertheless, the nonparametric estimators only consider the concordance of signs or ranks, which are less impacted by the issues than Pearson, so they show less bias when applied to high-frequency data. Their robustness is supported by their bounded influence functions as well. When we add a one-time jump, a tiny amount of another distribution contaminates the underlying distribution of observations, and consequently, the impact on correlation estimation is circumscribed for the nonparametric estimators.

3.4.1 Microstructure noise

The previous discussion assumes that the process $(X_t)_{t\in[0,1]}$ is observable, which is not the case in practice, unfortunately. Suppose we instead observe $(Y_t)_{t\in[0,1]}$ represented by

$$Y_t = X_t + \epsilon_t$$

where $(\epsilon_t)_{t\in[0,1]}$ is 2-dimensional i.i.d. noise process with $\mathbb{E}[\epsilon_t] = 0$ and $\mathbb{E}[\epsilon_t \epsilon_t'] = \Psi$. In this way, all estimators are applied to the observed increment over $[(i-1)/T, i/T)$

$$\Delta_i^T Y = Y_{i/T} - Y_{(i-1)/T} = \Delta_i^T X + \epsilon_{i/T} - \epsilon_{(i-1)/T}$$

for $i = 1, \ldots, T$. There are two perspectives to see the robustness of the subsampled Quadrant estimator.

First, the observed increment $\Delta_i^T Y$ has the opposite sign of $\Delta_i^T X$ only if the decrease in the noise ϵ_t is more prominent than $\Delta_i^T X$ over the time interval. Nevertheless, the literature often views noise's variance as a small fraction of the volatility of X (see Bandi and Russell (2006)), so the Quadrant and subsampled Quadrant estimator of the quantity q should be invariant to the added noise process. Since the Kendall estimator counts the rank concordance, it might be affected by the opposite rank of two pairs of increments in the noise. The level of noise's variance relative to the volatility still governs the influence, but it is more likely to be significant enough to flip the rank.

Second, Sparse sampling is capable of reducing the impact of noise and bias (Barndorff-Nielsen and Shephard (2007)), but it also reduces the sample size and consequently lowers the precision. The subsampled Quadrant estimator follows the spirit of sparse sampling since it uses increments over sub-series of length s/T. However, it avoids shrinking the sample size by taking all consecutive sub-series related to the subsampling method introduced by Zhang et al. (2005) and Zhang (2006).

3.4.2 Discretization

The prices of assets do not take all real values and are measured in a unit (or tick). Denoting one tick by $\alpha > 0$, then the observed price process can be expressed as $(\alpha \lfloor Y_t/\alpha \rfloor)_{t\in[0,1]}$. Consequently, rounding error is characterized as the difference between the rounded and actual increment over $[(i-1)/T, i/T)$. For a positive increment, $\Delta_i^T Y$, the corresponding rounded increment $\Delta_i^T \alpha \lfloor Y/\alpha \rfloor$ will be non-negative and equals zero whenever $\lfloor Y_{(i-1)/T}/\alpha \rfloor = \lfloor Y_{i/T}/\alpha \rfloor$. Therefore, for the Quadrant and subsampled Quadrant estimators, we only need to be cautious with zero returns that should not be observed much in actual prices if α is relatively negligible compared to the prices. As for the Kendall estimator, those zero returns will flip the rank of two increments in one component such as

$$\Delta_i^T Y^{(1)} > \Delta_j^T Y^{(1)} \text{ but } \lfloor Y^{(1)}_{(i-1)/T}/\alpha \rfloor = \lfloor Y^{(1)}_{i/T}/\alpha \rfloor,\ \lfloor Y^{(1)}_{(j-1)/T}/\alpha \rfloor > \lfloor Y^{(1)}_{j/T}/\alpha \rfloor.$$

Thus the probability that returns sampled over unit intervals are over the tick-size controls the bias caused by the rounding process.

3.4.3 Asynchronous trading

Not all assets are traded synchronously, and there can be many missing observations when sampling at high frequencies. Synchronized prices can be constructed by interpolation methods either in calendar time or event time, such as refresh time sampling, see Barndorff-Nielsen et al. (2011).

Let $\{\tau_k^{(j)}\}_{k=0}^{N_j}$ be the time stamps at which the j-th process is observed, $j = 1, 2$. In the following simulation study and empirical analysis, we use the event time approach and sample returns over time points that are the union of the two individual sets of time stamps, $\{\tau_i\}_{i=0}^{N} = \{\tau_k^{(1)}\}_{k=0}^{N_1} \cup \{\tau_k^{(2)}\}_{k=0}^{N_2}$. This sampling scheme should not harm the Quadrant-based estimators because, as long as $|\tau_{i+1} - \tau_i|$ is short enough, they do not concern the change in volatilities and still estimate the quantity q under Assumption 1. However, because of asynchronous trading, each process is possibly unobserved at some time stamps of $\{\tau_i\}_{i=0}^{N}$. The literature offers two ways to substitute those unobserved prices, previous-tick and linear interpolation.

If the process j is unobservable at time τ_i, the i-th observation is constructed as

$$\tilde{Y}_{\tau_i}^{(j)} = Y_{\tau_k^{(j)}}^{(j)}$$

where $k = \max(s|\tau_s^{(j)} \leq \tau_i)$, that is, prices are constructed by projecting the nearest past observation to the i-th point of the time grid. This interpolation method is called previous-tick. This method possibly generates zero returns, which are uninformative for the Quadrant-type estimators.

Besides, we can also use the linear interpolation over two observations over a time interval, including τ_i to construct the i-th observation

$$\tilde{Y}_{\tau_i}^{(j)} = Y_{\tau_k^{(j)}}^{(j)} + \frac{\tau_i - \tau_k^{(j)}}{\tau_{k'}^{(j)} - \tau_k^{(j)}}(Y_{\tau_{k'}^{(j)}}^{(j)} - Y_{\tau_k^{(j)}}^{(j)})$$

where $k = \max(s|\tau_s^{(j)} \leq \tau_i)$ and $k' = \min(s|\tau_s^{(j)} > \tau_i)$, for $i = 1, \ldots, N$ and $j = 1, 2$. In this way, $\Delta_{\tau_i}^N \tilde{Y}^{(j)}$ is exactly $Y_{\tau_i}^{(j)} - Y_{\tau_{i-1}}^{(j)}$ when $\tau_i, \tau_{i-1} \in \{\tau_k^{(j)}\}_{k=0}^{N_j}$, and is positively related to $Y_{\tau_{k'}^{(j)}}^{(j)} - Y_{\tau_k^{(j)}}^{(j)}$ if at least one of τ_i and τ_{i+1} at which the process j is not observed. If the process j is not observed at τ_i or τ_{i-1}, then its increment over time interval $[\tau_{i-1}, \tau_i)$ after interpolation is

$$\Delta_{\tau_i}^N \tilde{Y}^{(j)} = \tilde{Y}_{\tau_i}^{(j)} - \tilde{Y}_{\tau_{i-1}}^{(j)} = \left(\frac{\tau_i - \tau_k^{(j)}}{\tau_{k'}^{(j)} - \tau_k^{(j)}} - \frac{\tau_{i-1} - \tau_k^{(j)}}{\tau_{k'}^{(j)} - \tau_k^{(j)}} \right)(Y_{\tau_{k'}^{(j)}}^{(j)} - Y_{\tau_k^{(j)}}^{(j)}).$$

Under appropriate assumptions on $\{\tau_i\}_{i=0}^N$ (see Christensen et al. (2013)), the sign of observed increment $Y^{(j)}_{\tau^{(j)}_{k'}} - Y^{(j)}_{\tau^{(j)}_k}$ should be in line with $Y^{(j)}_{\tau_i} - Y^{(j)}_{\tau_k}$, then linear interpolation will not harm the precision of Quadrant and subsampled Quadrant estimators. Nevertheless, the influence on ranks of increments is intricate to determine.

In our simulation study, we compare the results under all combinations of sampling schemes and synchronization approaches and find similar patterns as well as values[2], so we only present the results for the event time scheme and previous-tick interpolation in the rest of this paper.

4 Simulation Study

We compare the Pearson, Quadrant, Kendall, and subsampled Quadrant estimators in simulation studies that are designed to replicate empirical situations that are encountered in empirical analyses with high-frequency data. Specifically, we generate two logarithmic price processes, $X^{(j)}_t$, $j = 1, 2$, for $t = 0, 1/T, \ldots, 1$, where $T = 23400$ which corresponds to second-by-second observation over 6.5 hours (a typical trading day).

Thus the highest available sampling frequency is one second, and we are unable to use the subsampled Quadrant estimator at that frequency. To assess the performance of the subsampled Quadrant estimator and make further comparisons with the other three estimators, we also estimate correlations using log returns sampled at lower frequencies to avoid microstructure noise in the financial market.

In this section, we denote two log-price processes by $X^{(j)}_t$ for $j = 1, 2$ and $t = 0, 1/T \ldots, 1$, and denote observed one-second log returns by $\Delta^T_i X = X_{i/T} - X_{(i-1)/T}$. Suppose we aim to compare those estimators on observed S-second log-returns, which are written as $\tilde{x}_l = \sum_{i=(l-1)S+1}^{lS} \Delta^T_i X$ for $l = 1, \ldots, n$ with $n = \lfloor T/S \rfloor$, then we compute estimates as follows:

$$\hat{q}_Q = \frac{1}{n} \sum_{l=1}^{n} \mathbf{1}_{\{\tilde{x}^{(1)}_l \tilde{x}^{(2)}_l > 0\}}$$

$$\hat{q}_K = \frac{2}{n(n-1)} \sum_{l<l'} \mathbf{1}_{\{(\tilde{x}^{(1)}_l - \tilde{x}^{(1)}_{l'})(\tilde{x}^{(2)}_l - \tilde{x}^{(2)}_{l'}) > 0\}}$$

[2]The optimal choices of sampling frequency are lower under the calendar time scheme than the event time, as we expect. The unit time interval under the event time scheme is generally longer than one second, so the event time scheme embraces a favor of low frequency essentially.

$$\bullet = \sin(\pi(\hat{q}_\bullet - \frac{1}{2})),\ \bullet = Q, K$$

and

$$P = \frac{\sum_{l=1}^{n} \tilde{x}_l^{(1)} \tilde{x}_l^{(2)}}{\sqrt{\sum_{l=1}^{n} \tilde{x}_l^{(1)2} \sum_{l=1}^{n} \tilde{x}_l^{(2)2}}}.$$

In contrast, the subsampled Quadrant estimate used in our comparison is computed based on $T-S+1$ observations:

$$\hat{q}_S = \frac{1}{T-S+1} \sum_{t=1}^{T-S+1} \mathbf{1}_{\{\sum_{j=0}^{S-1} \Delta_{i+j}^T X^{(1)} \sum_{j=0}^{S-1} \Delta_{i+j}^T X^{(2)} > 0\}}$$

and

$$Q_S = \sin(\pi(\hat{q}_S - \frac{1}{2})).$$

4.1 Time-varying Volatility

The first simulation is to investigate the inconsistency of estimators when returns are generated with time-varying volatility. Here we consider the classic model commonly used for high-frequency financial data, the Heston model.

For $j = 1, 2$, we consider log-prices are generated by

$$\begin{aligned} dX_t^{(j)} &= \mu_j dt + \sigma_t^{(j)} dW_t^{(j)} \\ d\sigma_t^{(j)2} &= \kappa_j(\bar{\sigma}_j^2 - \sigma_t^{(j)2})dt + s_j \sigma_t^{(j)} dB_t^{(j)}, \end{aligned} \tag{5}$$

where $W_t^{(j)}$ and $B_t^{(j)}$ are Brownian motions with $\text{Cov}(dW_t^{(j)}, dB_{it}) = \varrho_i dt$ and $\text{Cov}(dW_t^{(1)}, dW_t^{(2)}) = \rho dt$. The model is calibrated as in Aït-Sahalia et al. (2010), given by Table 1. Also, we assign initial values for $\sigma_0^{(j)2}$ drawn from a Gamma distribution $\Gamma(2\kappa_j\bar{\sigma}_j^2/s_j^2, s_j^2/2\kappa_j)$. The initial log-prices are set to $X_0^{(1)} = \log(100)$ and $X_0^{(2)} = \log(40)$. Here, we leave simulating the practical microstructure issues for later since our goal is to compare consistency among the correlation estimators in this subsection.

Table 1: Parameters calibration for the Heston model

	μ_j	$\bar{\sigma}_j^2$	κ_j	s_j	ϱ_j
$i=1$	0.05	0.16	3	0.8	-0.6
$i=2$	0.03	0.09	2	0.5	-0.75

Since the biases of those estimators might vary for underlying ρ's, we set ρ to be 0.25, 0.5, and

0.75 in the simulation study. With calibration choices in Table 1, this data-generating process leads to dependence between volatilities process $(\sigma_t^{(1)2}, \sigma_t^{(2)2})$ and Brownian motions $(W_t^{(1)}, W_t^{(2)})$, capturing by parameters ϱ_j, then we know the Pearson estimator is inconsistent to ρ in this design. On the contrary, the updating motions for volatilities are mean-reverting, indicating that they continuously evolve to get close to their means $\bar{\sigma}_j^2$. So the Quadrant and subsampled Quadrant estimators might be less affected when applied on returns sampled at relatively high frequencies, while the Kendall estimator is acceptable only when the initial values $\sigma_0^{(j)2}$ are close to the means. We plot the averages of estimates against sampling frequencies to see the asymptotic limits of estimators in Figures 2-4, and the results are entirely in line with the above conjecture.

Along the x-axis from right to left, we increase the frequency returns sampled and the number of observations in every replication. In this way, we expect to see the asymptotic limits of estimators as points close to the y-axis. Generally, we observe that the averages of estimates get closer to the actual values as we increase the sampling frequency across three plots.

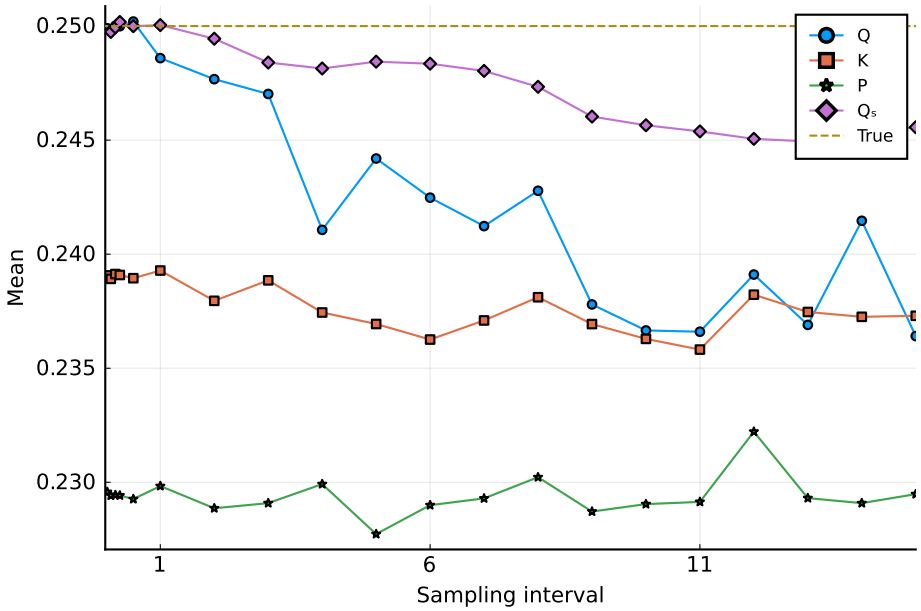


Figure 2: Plot the averages of correlation estimates against the frequencies at which returns are sampled, where the prices are generated in a Heston model with underlying correlation $\rho = 0.25$.

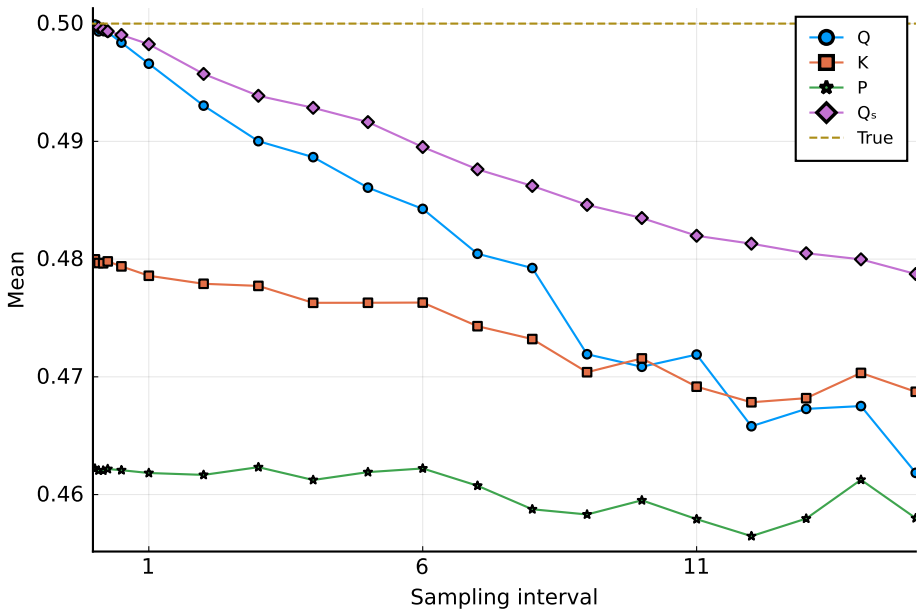


Figure 3: Plot the averages of correlation estimates against the frequencies at which returns are sampled, where the prices are generated in a Heston model with underlying correlation $\rho = 0.5$.

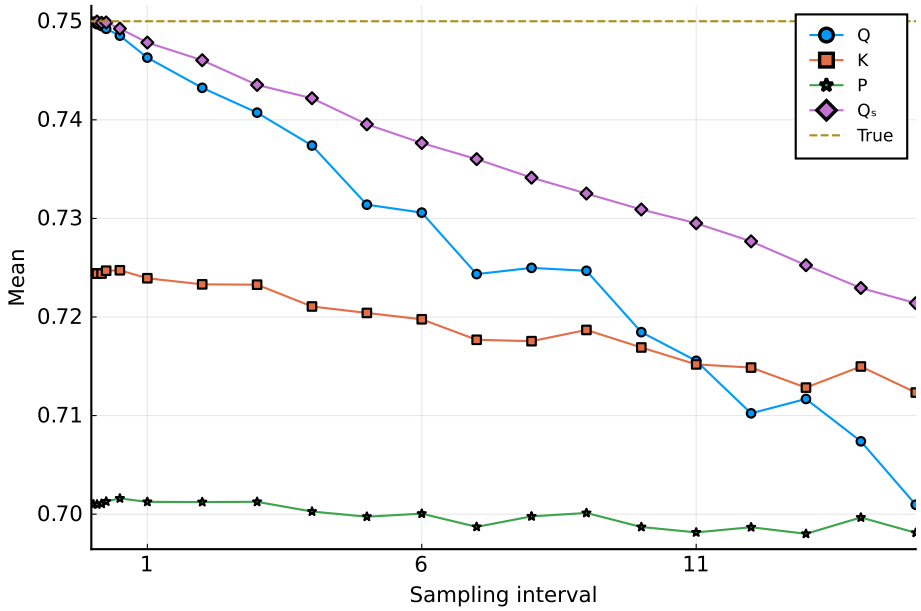


Figure 4: Plot the averages of correlation estimates against the frequencies at which returns are sampled, where the prices are generated in a Heston model with underlying correlation $\rho = 0.75$.

For all ρs, the averages of the Quadrant estimates are equal to the actual values when returns are sampled at one second, as shown in Example 1. The subsampled Quadrant estimator presents the lowest level of bias combined with a five-second sampling frequency when the actual correlation is 0.5 and 0.75. As the underlying correlation is 0.25, it has nearly zero bias when we sample returns at every

15 intervals. Also, we observe that the Kendall and Pearson estimators are associated with apparent biases at the limit, whereas the former's bias is about half of that of the latter. Additionally, biases increase for both estimators if we raise the underlying correlation. When the actual correlation is 0.75, the averages of the Pearson estimates are all around 0.7, no matter at which frequency we sample the returns.

4.2 Microstructure Issues

In this subsection, we investigate the robustness of nonparametric correlation estimators to three microstructure issues existing in actual high-frequency financial data: rounding error, asynchronous trading, and independent noise. Those are viewed as the main reasons behind the Epps effect in literature. To accentuate the effect of those contaminations, we will focus on the Levy model that assumes constant volatility.

To be more specific, we generate two latent log-price paths following a fundamental diffusion process, for $j = 1, 2$,

$$dX_t^{(j)} = \sigma^{(j)} dW_t^{(j)}$$

where $X_t^{(j)}$s are prices of two assets in logarithm, and $W_t^{(j)}$s are two standard Brownian motions with correlation ρ. In our study, we choose $(0.15, 0.45)$ respectively for $\sigma^{(1)}$ and $\sigma^{(2)}$, and consider 0.25, 0.5, and 0.75 as the choices of correlation coefficient ρ. Next, we simulate the microstructure issues separately and jointly to compare the accuracy as well as the bias of the correlation estimators.

First, we add independent noise on the efficient price

$$Y_t^{(j)} = X_t^{(j)} + \epsilon_t^{(j)}$$

where $\epsilon_t^{(j)} \sim N(0, \omega^2)$ for $j = 1, 2$. The variance of the noise is set to increase with the volatility of the efficient price (e.g., Bandi and Russell (2006)), $\omega^2 = \xi^2 \sqrt{T^{-1} \sum_{i=1}^{T} \sigma_{i/T}^{(j)4}}$, where $\xi^2 = 0.001$. Next, we round off the contaminated prices with the tick size of one cent and update prices by certain

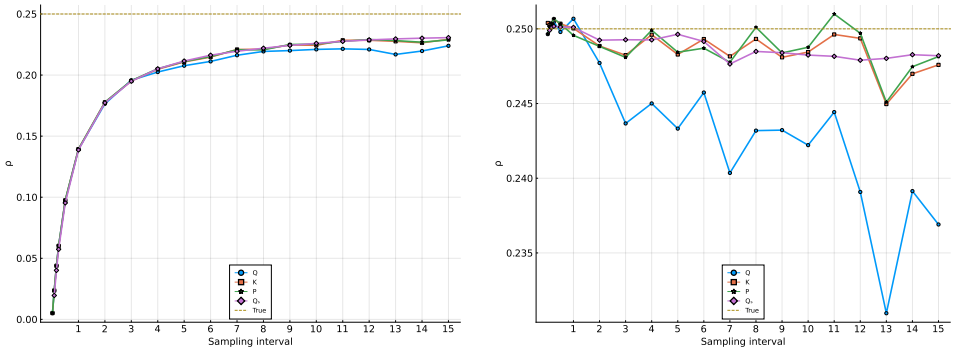
probabilities to generate asynchrony

$$\tilde{Y}_t^{(j)} = \begin{cases} \lfloor \exp(Y_t^{(j)})/0.01 \rfloor \times 0.01 & \text{with probability } p_j^{\text{tr}} \\ \tilde{Y}_{t-1}^{(j)} & \text{otherwise} \end{cases}.$$

We consider various updating probabilities for the two observed price processes. One is traded more frequently, $p_h^{\text{tr}} = 0.8$, and the other is traded at a lower frequency, $p_l^{\text{tr}} = 0.5$. In this way, the prices are not updated every second or updated simultaneously. Hence, we need to synchronize observed prices before applying correlation estimators. The sampling scheme we use here is the event time and measures the sampling frequency by sampling interval, which refers to the number of intervals on the generated time grid between the two sampled prices. For instance, five sampling intervals equals returns over 301 consecutive transactions. We make up the missing observations on the time grid generated by two price processes with the previous-tick interpolation.

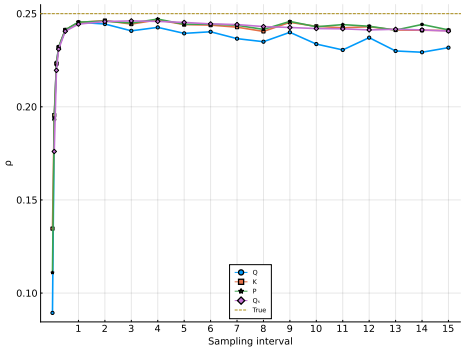
Figures 5-7 present the signature plot of correlation estimates against sampling frequency under various sources of microstructure issues. All correlation estimators exhibit the pattern described as the Epps effect while facing independent noise and asynchronous trading. There is no evident difference among the subsampled Quadrant, Kendall, and Pearson estimators in terms of bias over any underlying correlations and sampling frequencies. Instead, the Quadrant estimator suffers more from loss of sample size when we sample less frequently.

Independent noise, as we speculated, drives the estimated correlations toward zero when its size is relatively substantial. That explains why we witness more considerable bias at high sampling frequencies (less than 60 intervals) for higher underlying correlations. Rounding error does not induce bias at high sampling frequencies as the initial prices we choose here are immensely more than one tick, so the rounding procedure only moderately impacts them. Similarly, asynchronous trading also leads to bias at high sampling frequencies, as we create a significant proportion of zero returns after previous-tick synchronization. Nonetheless, the correlation estimates converge to the actual value swiftly in contrast to the case of independent noise. The right bottom subfigures in Figures 5-7 bring about the signature plots under all three microstructure issues mentioned above. It is worth mentioning that the bias in this scenario primarily originates from the independent noise, as a result of the more extensive noise relative to the efficient returns sampled less frequently than every 60 intervals.

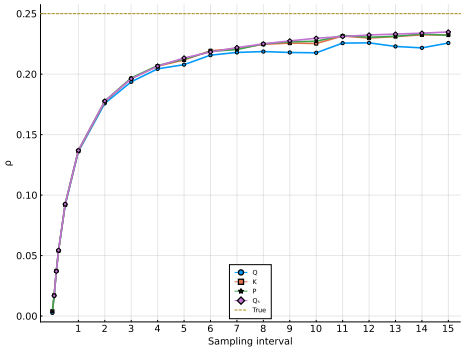


(a) Independent noise

(b) Rounding Error

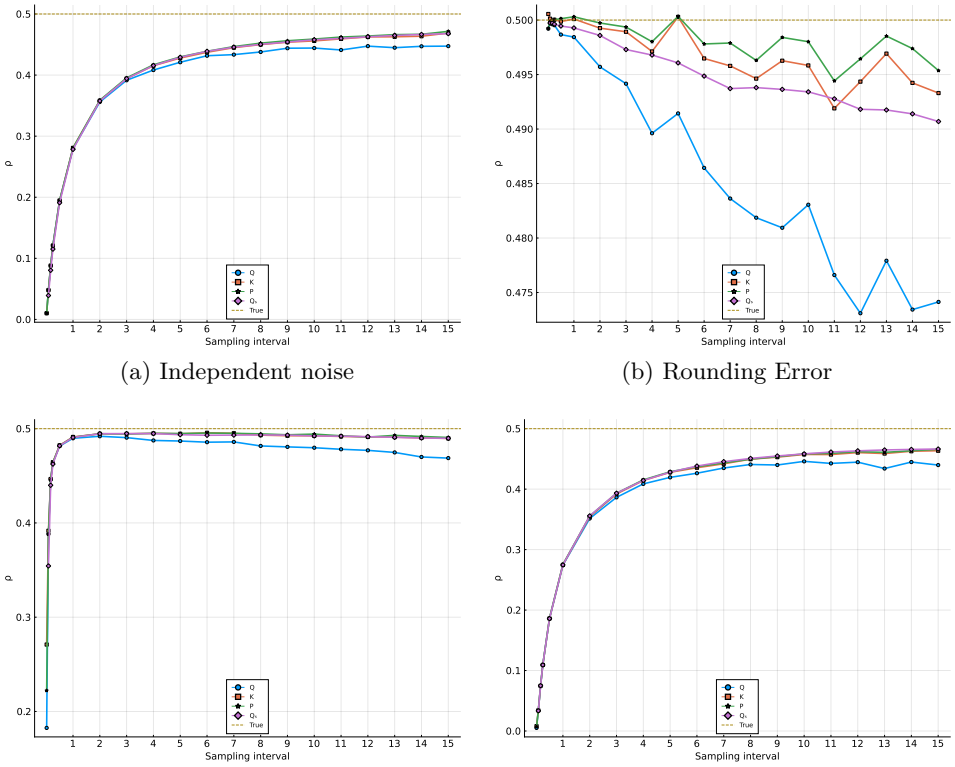


(c) Asynchronous trading



(d) All microstructure issues

Figure 5: Plot the averages of correlation estimates against the frequencies at which returns are sampled, where the prices are generated in a Levy model with underlying correlation $\rho = 0.25$ and different microstructure issues.



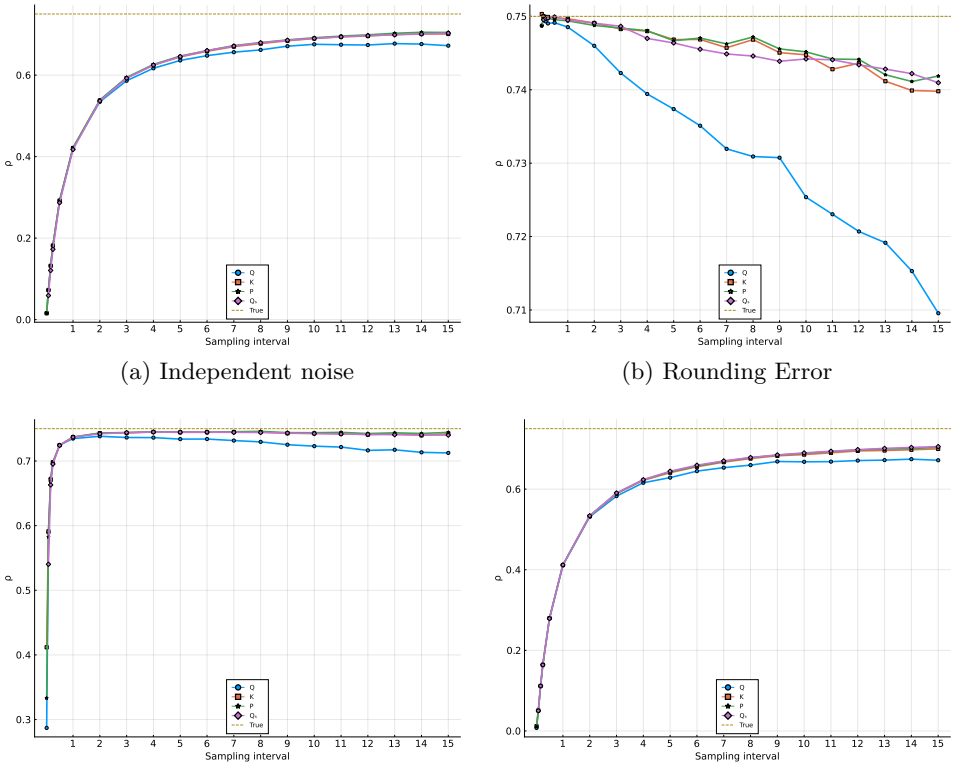
(a) Independent noise

(b) Rounding Error

(c) Asynchronous trading

(d) All microstructure issues

Figure 6: Plot the averages of correlation estimates against the frequencies at which returns are sampled, where the prices are generated in a Levy model with underlying correlation $\rho = 0.5$ and different microstructure issues.



(a) Independent noise (b) Rounding Error

(c) Asynchronous trading (d) All microstructure issues

Figure 7: Plot the averages of correlation estimates against the frequencies at which returns are sampled, where the prices are generated in a Levy model with underlying correlation $\rho = 0.75$ and different microstructure issues.

Table 2 reports the results when all three types of issues are considered. We divide it into three panels, corresponding to three possible underlying correlation coefficients: 0.25, 0.5, and 0.75. In each panel, estimators are compared based on mean squared errors of correlation estimates obtained from returns sampled at different frequencies, from one second to five minutes. The first finding is that, in every panel, the mean squared errors of all estimators show a U-shape against sampling frequencies. The mean squared error is high for high-frequency returns because the magnitude of microstructure noise is more significant than efficient returns over short time intervals, and it also is relatively high for low-frequency because of lacking enough observations within a trading day. Meanwhile, the frequency at which the minimum mean squared errors appear shifts to the right as the underlying correlation increases, implying that the higher the actual correlation, the more severe distortion microstructure noise causes on high-frequency estimation. That can also be seen from the considerably increasing mean squared errors from top to bottom for high sampling frequencies like one-interval and five-interval.

Second, we observe that the subsampled Quadrant estimator performs the best for all correlations.

Table 2: Mean squared errors of correlation estimators on returns generated in Levy model with microstructure issues ($\times 10^{-2}$).

S	1	5	10	15	30	60	120	180	240	300	360	420	480	540	600
							$\rho = 0.25$								
Quadrant	6.127	5.487	4.654	4.014	2.847	1.923	1.710	2.132	2.586	3.163	3.593	4.146	4.624	5.097	5.497
Kendall	6.054	5.451	4.586	3.931	2.650	1.576	1.060	1.118	1.284	1.508	1.707	1.979	2.213	2.520	2.772
Sub Q		5.450	4.551	3.843	2.508	1.345	0.715	0.643	0.717	0.849	1.005	1.179	1.352	1.545	1.739
Pearson	6.069	5.450	4.575	3.922	2.624	1.541	1.007	1.021	1.145	1.328	1.495	1.726	1.928	2.154	2.395
							$\rho = 0.5$								
Quadrant	24.499	21.823	18.218	15.432	10.192	5.671	3.289	2.805	2.737	2.937	3.231	3.689	4.070	4.285	4.528
Kendall	24.220	21.739	18.147	15.349	10.009	5.341	2.577	1.810	1.533	1.513	1.609	1.802	1.804	2.000	2.114
Sub Q		21.787	18.093	15.290	9.898	5.136	2.236	1.405	1.128	1.044	1.038	1.087	1.161	1.255	1.351
Pearson	24.289	21.720	18.130	15.328	9.954	5.284	2.499	1.700	1.410	1.362	1.409	1.553	1.573	1.696	1.804
							$\rho = 0.75$								
Quadrant	55.132	49.072	40.970	34.500	22.495	11.993	5.579	3.878	3.033	3.041	2.801	2.797	3.127	2.968	3.221
Kendall	54.498	48.921	40.803	34.387	22.238	11.676	5.036	3.007	2.118	1.751	1.522	1.397	1.389	1.320	1.367
Sub Q		48.993	40.763	34.367	22.217	11.480	4.771	2.695	1.808	1.374	1.140	1.012	0.944	0.916	0.921
Pearson	54.641	48.885	40.760	34.345	22.175	11.619	4.960	2.910	2.002	1.612	1.377	1.239	1.177	1.128	1.122

In each trading day, we generate $T = 23400$ seconds returns. The sampling frequency here represents the number of basic sampling intervals the returns are sampled over. The above results are based on 5000 replications.

Based on the minimum mean squared error values, the subsampled Quadrant associated with 180, 360, and 540 intervals are our best estimation choices for the actual correlation being 0.25, 0.5, and 0.75, respectively. The Kendall estimator behaves similarly to the Pearson estimator. However, the Quadrant estimator always shows the highest mean squared error regardless of frequency or underlying correlation. Moreover, its mean squared errors explode rapidly when sampling frequency decreases since it relies on sample size strongly. Combined with the signature plots, the subsampled Quadrant estimator's precision benefits from its efficiency gain since it embraces the same level of bias as Kendall and Pearson. The comparison between the Quadrant and the subsampled Quadrant estimators again emphasizes the improvement of the subsampling approach.

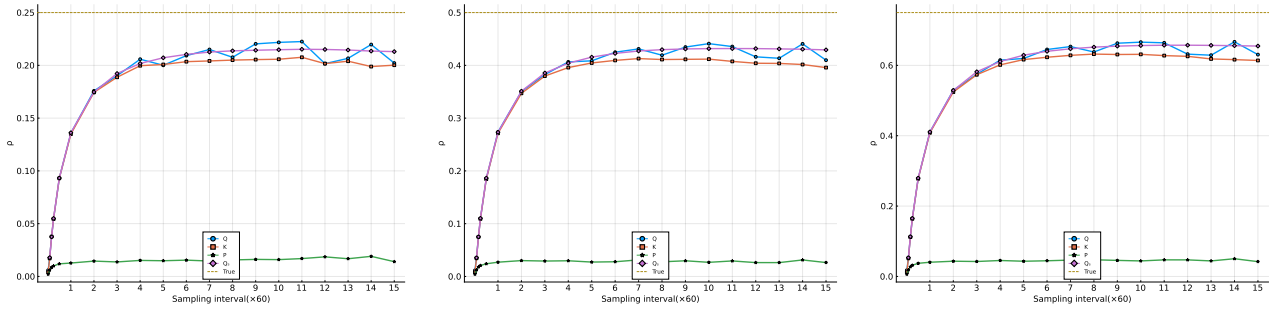
4.3 Microstructure Issues and Jumps

Besides microstructure issues, we may observe jumps in high-frequency prices. Now we consider two possible types of jumps, individual jumps and synchronous jumps on both sides. We observe log-prices added noise and jumps

$$Y_t = X_t + \epsilon_t + \sum_{s \leq t} J_s$$

where J_t denotes the Poisson jump process with intensity λ and size uniformly drawn from $[-2, -1] \bigcup [1, 2]$ divided by $\sqrt{2\lambda}$. Here, we consider individual jumps with intensity $\lambda^j = 1$ and cojumps with intensity $\lambda^c = 1$. That implies, on average, one asynchronous jump and one synchronous jump happen on each of the two price processes at the end of a trading day.

We first compare the bias of correlation estimators when only individual jumps and all sources of microstructure issues hit the prices in Figure 8. All estimators display a similar pattern as in the scenarios where we add microstructure issues only. Consequently, the bias brought about by the independent noise continuously dominates in this scenario. Among the nonparametric estimators, now the subsampled Quadrant converges to some quantity closer to the underlying correlation than the Kendall estimator. Across three possible underlying correlations, Pearson estimates some quantity else, mainly determined by the individual jumps.



(a) $\rho = 0.25$ (b) $\rho = 0.5$ (c) $\rho = 0.75$

Figure 8: Plot the average correlation estimates against the frequency at which returns are sampled when the prices are generated in a Levy model with independent noise, rounding error, asynchronous trading, and individual jumps.

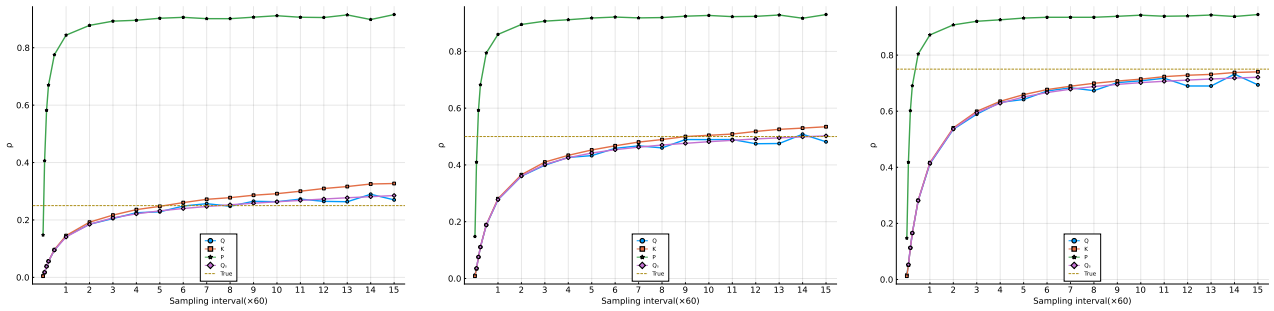
The comparison of the accuracy of correlation estimators in this scenario is given in Table 3. The tremendous change shows up on the Pearson estimator. Overall, the mean squared errors of the Pearson estimator climb when we add jumps and will not diminish to a competitive level on low-frequency returns because the Pearson estimator converges to another quantity instead of the actual correlations. However, for nonparametric estimators, there is little difference with adding jumps. Such distinct impact of jumps can be explained by the robustness of estimators facing extreme outliers, and it has known that only the Pearson estimator has an unbounded influence function. After ruling out the Pearson estimator, the subsampled Quadrant estimator is the most favorable one associated with 180, 300, and 480 sampling intervals for correlation being 0.25, 0.5, and 0.75, slightly different from the best frequency choices in Table 2.

Now we consider the influence of cojumps on correlation estimation. The Pearson estimator quickly converges to its limit, over 0.75, because of the perfect correlation between the cojumps. That perfect correlation also makes the nonparametric estimators display minute upward bias at low sampling frequencies. Their bounded influence functions predict the invariance to jumps but are insufficient to explain their behaviors in the finite sample. The proportion of outliers caused by jumps is rather sizable when we have limited sample sizes, so the nonparametric estimators exhibit dissimilar bias here. Even so, the subsampled Quadrant maintains the most negligible bias if we sample returns less frequently.

Table 3: Mean squared errors of correlation estimators on returns generated in the Levy model with microstructure issues and individual jumps ($\times 10^{-2}$).

S	1	5	10	15	30	60	120	180	240	300	360	420	480	540	600
							$\rho = 0.25$								
Quadrant	5.989	5.456	4.625	4.002	2.799	1.940	1.773	2.192	2.582	3.221	3.797	4.510	4.904	5.678	6.035
Kendall	6.058	5.454	4.564	3.911	2.630	1.624	1.132	1.208	1.293	1.600	1.870	2.181	2.394	2.691	2.970
Sub Q		5.405	4.531	3.822	2.486	1.363	0.760	0.696	0.776	0.925	1.099	1.280	1.480	1.683	1.885
Pearson	6.147	5.974	5.864	5.824	5.766	5.883	6.047	6.285	6.582	6.667	7.154	7.335	7.651	7.806	7.874
							$\rho = 0.5$								
Quadrant	23.893	21.615	18.187	15.405	10.251	5.700	3.322	2.893	2.885	3.259	3.541	3.752	4.530	4.895	5.313
Kendall	24.226	21.703	18.139	15.369	10.114	5.493	2.809	2.092	1.941	1.969	2.101	2.211	2.519	2.738	3.015
Sub Q		21.607	18.073	15.217	9.865	5.204	2.374	1.587	1.323	1.249	1.274	1.357	1.462	1.594	1.750
Pearson	24.572	23.788	23.267	23.013	22.785	22.630	22.586	22.867	23.058	23.479	23.610	23.702	24.279	24.212	24.686
							$\rho = 0.75$								
Quadrant	53.669	48.641	40.820	34.442	22.532	12.063	5.712	4.123	3.188	3.292	2.926	2.975	3.631	3.304	3.600
Kendall	54.467	48.832	40.782	34.443	22.456	11.906	5.429	3.527	2.717	2.383	2.309	2.280	2.314	2.463	2.560
Sub Q		48.558	40.586	34.249	22.175	11.565	4.996	2.987	2.149	1.737	1.547	1.461	1.432	1.449	1.491
Pearson	55.290	53.498	52.268	51.681	50.955	50.619	50.491	50.811	50.698	51.188	51.275	51.220	51.486	51.873	52.434

In each trading day, we generate $T = 23400$ seconds returns. The sampling frequency here represents the number of basic sampling intervals the returns are sampled over. The above results are based on 5000 replications.



(a) $\rho = 0.25$ (b) $\rho = 0.5$ (c) $\rho = 0.75$

Figure 9: Plot the average correlation estimates against the frequency at which returns are sampled when the prices are generated in a Levy model with independent noise, rounding error, asynchronous trading, and cojumps.

Furthermore, in Table 4, we present the mean squared errors of estimators in this scenario. Similar to the results in Table 3, nonparametric estimators' precision matches the one without jumps. In this case, the subsampled Quadrant delivers the best job with 180, 300, and 600 intervals sampling frequencies. Pearson shows lower mean squared errors for the underlying correlation being 0.75 because 0.75 is closer to its actual limit.

4.4 Microstructure Issues and Time-varying Volatility

We investigate the effect of time-varying volatilities and microstructure issues, including noise, rounding error from discretization, and asynchronous trading. We still simulate efficient log prices as in the Heston model (5) associated with parameter choice in Table 1, denoted by X_t, for $t = 0, 1/T, \ldots, 1$ and $T = 23400$ as the number of seconds in a trading day. Then, on each draw, we add independent noise, rounding error, and asynchronous trading, characterized as above. Again, we simulate the Heston model with correlations of 0.25, 0.5, and 0.75.

Taking note of the asynchrony, we apply the event time sampling scheme here to unify the time grids of the two processes because, as Chang et al. (2021) pointed out, the correlation estimates converge faster under event time sampling. Two remarks should be mentioned here. The unit sampling interval is probably longer than one second, and the total number of the unified grid points can consequently be less than 23400. Besides, we cannot sample returns at a fixed frequency since the interval between two observed time points is not determined. Instead, we sample returns over more basic sampling intervals to use sparse sampling. On the unified time grid, we synchronize prices with the previous-tick interpolation.

Table 4: Mean squared errors of correlation estimators on returns generated in the Levy model with microstructure issues and cojumps ($\times 10^{-2}$).

S	1	5	10	15	30	60	120	180	240	300	360	420	480	540	600
							$\rho = 0.25$								
Quadrant	5.977	5.478	4.591	3.940	2.725	1.783	1.647	1.980	2.416	2.915	3.419	3.986	4.562	5.074	5.764
Kendall	6.048	5.442	4.509	3.839	2.514	1.387	0.859	0.884	1.045	1.292	1.494	1.796	2.058	2.272	2.600
Sub Q		5.414	4.513	3.782	2.430	1.263	0.622	0.537	0.608	0.730	0.886	1.058	1.244	1.433	1.618
Pearson	3.900	7.576	15.852	21.593	30.347	36.752	40.485	41.973	42.527	43.138	43.472	43.367	43.489	43.818	44.137
							$\rho = 0.5$								
Quadrant	23.872	21.626	18.179	15.357	10.052	5.550	2.976	2.472	2.466	2.780	2.919	3.289	3.724	3.924	4.433
Kendall	24.201	21.696	18.098	15.235	9.811	5.068	2.259	1.420	1.227	1.202	1.218	1.321	1.497	1.591	1.775
Sub Q		21.576	18.008	15.171	9.743	4.986	2.078	1.232	0.941	0.850	0.862	0.913	0.997	1.100	1.204
Pearson	15.249	6.001	5.571	7.167	10.906	14.166	16.316	17.161	17.542	17.867	18.065	18.058	18.169	18.359	18.458
							$\rho = 0.75$								
Quadrant	53.733	48.661	40.687	34.409	22.296	11.893	5.435	3.614	2.660	2.668	2.269	2.262	2.703	2.550	2.567
Kendall	54.496	48.844	40.655	34.285	22.034	11.347	4.712	2.655	1.766	1.334	1.108	1.008	0.930	0.948	0.949
Sub Q		48.613	40.580	34.146	21.980	11.294	4.636	2.553	1.668	1.228	0.990	0.857	0.784	0.744	0.730
Pearson	39.118	16.081	6.816	4.259	2.637	2.736	3.245	3.461	3.590	3.650	3.732	3.686	3.825	3.824	3.787

In each trading day, we generate $T = 23400$ seconds returns. The sampling frequency here represents the number of basic sampling intervals the returns are sampled over. The above results are based on 5000 replications.

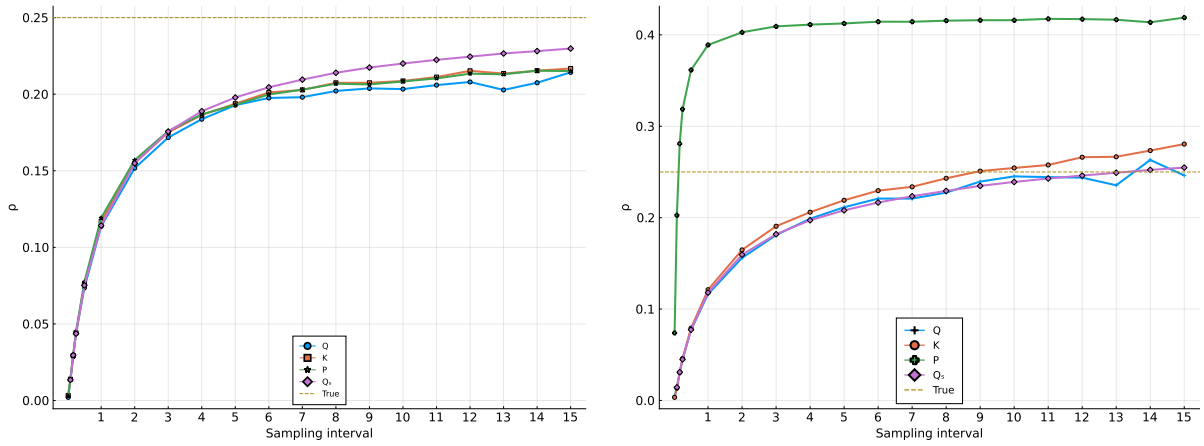
Table 5: Mean squared errors of correlation estimators on returns generated in the Heston model with microstructure issues ($\times 10^{-2}$).

S	1	5	10	15	30	60	120	180	240	300	360	420	480	540	600
							$\rho = 0.25$								
Quadrant	6.153	5.640	5.014	4.419	3.456	2.537	2.213	2.429	2.858	3.262	3.712	4.386	4.933	5.154	5.771
Kendall	6.094	5.603	4.924	4.325	3.205	2.089	1.472	1.404	1.533	1.681	1.902	2.231	2.456	2.696	2.981
Sub Q		5.609	4.876	4.274	3.085	1.908	1.076	0.855	0.829	0.885	0.993	1.135	1.288	1.468	1.663
Pearson	6.106	5.591	4.896	4.285	3.135	2.003	1.379	1.312	1.398	1.535	1.726	1.948	2.180	2.339	2.573
							$\rho = 0.5$								
Quadrant	24.604	22.400	19.525	17.108	12.607	8.121	5.081	4.006	3.837	3.978	3.895	4.386	4.766	4.821	5.079
Kendall	24.369	22.326	19.387	16.937	12.175	7.504	4.195	2.970	2.588	2.360	2.295	2.404	2.523	2.548	2.682
Sub Q		22.386	19.398	16.993	12.214	7.424	3.853	2.520	1.915	1.606	1.456	1.400	1.403	1.447	1.507
Pearson	24.419	22.297	19.303	16.820	11.925	7.227	3.977	2.820	2.459	2.222	2.157	2.247	2.307	2.378	2.452
							$\rho = 0.75$								
Quadrant	55.364	50.350	43.627	38.162	27.597	17.182	9.481	6.554	5.325	4.676	4.396	4.045	4.323	4.055	4.275
Kendall	54.841	50.167	43.383	37.813	26.877	16.115	8.340	5.408	4.030	3.312	2.884	2.590	2.448	2.330	2.329
Sub Q		50.329	43.503	38.097	27.227	16.396	8.251	5.155	3.607	2.729	2.194	1.844	1.619	1.467	1.375
Pearson	54.944	50.105	43.217	37.497	26.299	15.425	7.898	5.134	3.836	3.193	2.805	2.516	2.353	2.262	2.224

In each trading day, we generate $T = 23400$ seconds returns. The sampling frequency here represents the number of basic sampling intervals the returns are sampled over. The above results are based on 5000 replications.

Table 5 provides the mean squared errors of the correlation estimates. Because of microstructure issues, all estimators have substantially low precision at high sampling frequencies compared to the previous case. Across all correlations considered here, the subsampled Quadrant estimator presents the lowest mean squared error at different frequencies. Especially when the actual correlation is 0.75, the subsampled Quadrant estimator appears more precise with the decreasing sampling frequency, which suggests the subsampled Quadrant estimator may attain its best accuracy at sampling frequencies lower than 600 intervals, or equivalently, the sample size smaller than 39. The Kendall estimator is more accurate than the Quadrant for all underlying correlations and at all sampling frequencies, while it is slightly outperformed by the Pearson estimator overall. All estimators' best sampling frequencies (corresponding to the least mean squared errors) are close or the same, declining against the underlying correlations.

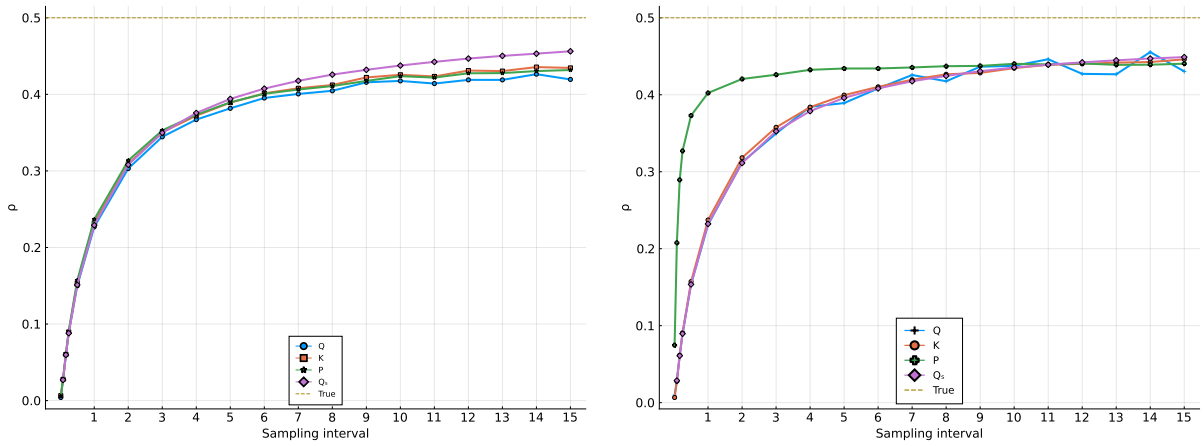
We are more interested in the bias of estimators in this simulation study as their deviations are predictable based on asymptotic variances, so we plot the averages of estimates in Figures 10-12. With the presence of the microstructure features, we still observe the Epps effect from the plot. At the highest sampling frequency, the averages of the three estimators are near zero. However, the bias is still immense when we sample returns at every ten intervals. The subsampled Quadrant is associated with a relatively less significant bias than the other three alternatives, which helps us to understand the former's higher precision in Table 5. The Kendall and Pearson estimators present close averages across all frequencies as well as correlations.



(a) Triple microstructure issues

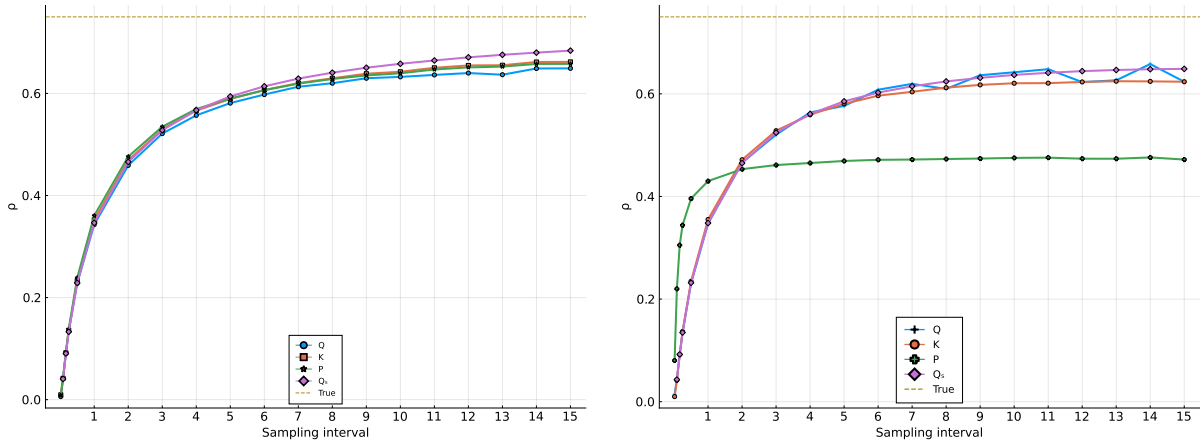
(b) Triple microstructure issues and jumps

Figure 10: Plot the average correlation estimators on returns generated in the Heston model with independent noise, discretization, asynchrony, and jumps against different sampling frequencies. The underlying correlation is $\rho = 0.25$.



(a) Triple microstructure issues (b) Triple microstructure issues and jumps

Figure 11: Plot the average correlation estimators on returns generated in the Heston model with independent noise, discretization, asynchrony, and jumps against different sampling frequencies. The underlying correlation is $\rho = 0.5$.



(a) Triple microstructure issues (b) Triple microstructure issues and jumps

Figure 12: Plot the average correlation estimators on returns generated in the Heston model with independent noise, discretization, asynchrony, and jumps against different sampling frequencies. The underlying correlation is $\rho = 0.75$.

The alternative combination of trading frequencies is considered as well. The mean squared errors and bias against sampling frequencies are similar except for lower accuracy and more considerable bias when we use trading probabilities 0.5 and 0.2. The results are provided in Appendix.

In addition, we add individual jumps as well as cojumps on the prices altered by the microstructure issues. The perfect correlation by cojumps is mitigated by the independence induced by individual jumps. That explains why the Pearson estimator is converging to some value around 0.45 across different underlying correlations. For the same reason, the nonparametric estimators are slightly higher than the true correlation (0.25) when the sampling interval length is more than 600. Table 6 implies that the nonparametric estimators' precision has not changed significantly by adding sources of jumps. The subsample Quadrant picks similar sampling frequencies to attain the least mean squared errors. On the contrary, the Pearson estimator shows exploding mean squared errors after the enormous bias.

5 Empirical Application

We now use the estimators mentioned above to explore the intraday correlations between stocks and the market. The analysis uses high-frequency prices for twenty-two stocks evenly from eleven sectors, constituents of the S&P 500 market index, and the SPY ETF on the S&P 500 index as our market proxy. The sample period runs from January 1, 2015, to December 31, 2021, delivering 1763 trading days. The data is the collection of transaction prices taken from the TAQ database through the Wharton Research Data Services (WRDS) system. We cleaned the data and only kept prices by second, following the instruction in Barndorff-Nielsen et al. (2011).

We applied the tick-time sampling scheme for all correlation estimators and then used the duration between two transactions in a row as the unit sampling interval that almost always lasts longer than a second. Table 7 presents the summary statistics of the selected stocks as well as the market. SPY is the most frequently traded stock in our collection. On average, it only offers observations as half of the total number of seconds in a trading day, which is linked to a trading probability of about 0.5. In our collection of stocks, we include stocks with varying trading frequencies. AAPL has 11706 prices recorded by every second across a trading day on average, in contrast to 2478 transactions of AMT. From the Duration time listed in Table 7, we have an estimation of the actual time lengths of the unit sampling intervals for the selected stocks. Generally, we expect lower sampling frequency for less frequently traded stocks when we use returns sampled by unit. However, transactions do not imply updated prices, demonstrated by the average number of zero one-unit returns across the day in Table 7. The last two columns of Table 7 provide the estimated noise variances and integrated variances for each stock. The noise sizes are close except for AMT, which is traded the least frequently, so our

Table 6: Mean squared errors of correlation estimators on returns generated in the Heston model with microstructure issues and jumps ($\times 10^{-2}$).

S	1	5	10	15	30	60	120	180	240	300	360	420	480	540	600
							$\rho = 0.25$								
Quadrant	6.038	5.648	4.941	4.389	3.336	2.436	2.138	2.259	2.781	3.174	3.666	4.379	4.727	5.600	5.951
Kendall	6.090	5.624	4.880	4.258	3.095	1.977	1.332	1.224	1.319	1.567	1.726	2.070	2.323	2.707	2.911
Sub Q		5.570	4.817	4.224	3.010	1.809	0.998	0.779	0.749	0.811	0.929	1.088	1.266	1.458	1.666
Pearson	4.388	4.769	6.949	8.338	10.425	11.844	12.699	13.110	13.359	13.579	13.877	14.026	14.280	14.508	14.331
							$\rho = 0.5$								
Quadrant	24.068	22.324	19.395	17.051	12.336	7.884	4.703	3.932	3.389	3.730	3.829	3.970	4.417	4.693	5.092
Kendall	24.346	22.324	19.275	16.844	11.926	7.230	3.862	2.790	2.281	2.153	2.158	2.270	2.337	2.555	2.730
Sub Q		22.232	19.297	16.865	12.041	7.258	3.746	2.464	1.870	1.602	1.482	1.451	1.479	1.526	1.606
Pearson	19.340	12.999	11.024	10.537	10.216	10.106	10.060	10.190	10.225	10.364	10.645	10.846	10.961	10.913	10.981
							$\rho = 0.75$								
Quadrant	54.145	50.094	43.568	38.057	27.346	16.879	9.087	6.597	5.092	4.920	4.112	4.065	4.627	4.228	4.407
Kendall	54.783	50.183	43.399	37.712	26.749	15.905	8.241	5.521	4.338	3.699	3.296	3.215	3.111	3.098	3.172
Sub Q		49.988	43.305	37.878	26.926	16.260	8.316	5.343	3.895	3.078	2.608	2.324	2.151	2.044	1.990
Pearson	46.170	32.561	26.202	23.760	20.703	18.792	17.678	17.430	17.338	17.358	17.297	17.529	17.700	17.820	17.858

In each trading day, we generate $T = 23400$ seconds returns. The sampling frequency here represents the number of basic sampling intervals the returns are sampled over. The above results are based on 5000 replications.

variance estimator is dominated by the variance of efficient returns in this scenario. In addition, we observe relatively high integrated variances for AMD, AAL, HAL, and TSLA, implying the potential presence of jumps.

Table 7: Descriptive statistics for the selected S&P 500 stocks and S&P 500 ETF.

Sector	Ticker	Transactions	Duration (in seconds)	Zero returns	$\hat{\sigma}_\epsilon^2$	$\hat{IV}$
	SPY	12056	2.275	3488	1.71×10^{-9}	6.57×10^{-5}
Utilities	D	3137	8.037	1114	1.92×10^{-8}	1.37×10^{-4}
	DUK	3329	7.467	1132	1.76×10^{-8}	1.35×10^{-4}
Real estate	AMT	2478	10.178	474	3.54×10^{-6}	1.76×10^{-4}
	PLD	2606	9.771	879	3.16×10^{-8}	1.69×10^{-4}
Materials	LYB	2804	9.226	664	5.28×10^{-8}	3.33×10^{-4}
	NEM	4056	6.670	1659	3.55×10^{-8}	3.77×10^{-4}
Information	AAPL	11706	2.271	3067	6.18×10^{-9}	1.86×10^{-4}
Technology	AMD	6139	14.865	2133	2.70×10^{-7}	8×10^{-4}
Industrials	AAL	4088	6.829	1692	5.43×10^{-8}	7.42×10^{-4}
	UNP	3715	6.862	867	2.40×10^{-8}	1.85×10^{-4}
Health Care	JNJ	5523	4.571	1720	8.58×10^{-9}	1.08×10^{-4}
	MRK	5119	5.056	2090	1.11×10^{-8}	1.37×10^{-4}
Financials	JPM	7977	3.217	2888	8.76×10^{-9}	1.78×10^{-4}
	WFC	5772	4.495	2689	1.29×10^{-8}	2.28×10^{-4}
Energy	HAL	4591	5.564	2074	4.79×10^{-8}	6.08×10^{-4}
	XOM	6794	3.831	2753	9.94×10^{-9}	2×10^{-4}
Consumer	PG	5271	4.860	1962	8.42×10^{-9}	1.09×10^{-4}
Staples	WMT	5654	4.461	2116	8.47×10^{-9}	1.18×10^{-4}
Consumer	TSLA	7485	5.121	557	5.17×10^{-8}	7×10^{-4}
Discretionary	AMZN	5869	5.028	393	2×10^{-8}	2.11×10^{-4}
Communication	DIS	6273	4.193	1819	1.01×10^{-8}	1.71×10^{-4}
Services	FB	9225	2.805	1891	1.01×10^{-8}	2.38×10^{-4}

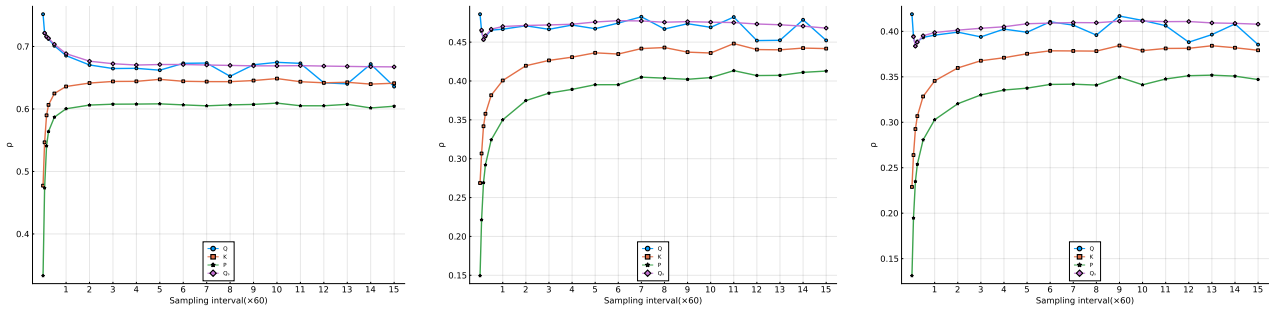
The first column describes the industry sector of the selected stocks. Transaction means the average number of daily transactions during the sampling period, from Jan 1, 2015, to Dec 31, 2021. Duration presents the average waiting time in seconds between two consecutive transactions on a stock. We count the average number of zero returns caused by two transactions in a row within a trading day. To estimate the variance of the independent noise in stock prices, we use half of the realized variance of returns sampled at every event time. In contrast, the integrated variance of stock is approximated by the realized variance of returns sampled at 300 unit intervals (event time), which corresponds to 5-minute returns under the calendar time scheme.

5.1 Integrated Correlation Estimation

In this subsection, we estimate the correlation between the continuous part in the prices of two stocks if it is constant during a trading day, or we aim to estimate the integrated correlations over a trading day if it is time-changing. We synchronized prices between two stocks in the same sector as well as stocks and the market using the previous-tick approach. Each trading day, we sample the synchronized prices over

every S unit intervals (a proxy for S-second sampling frequency under the calendar sampling scheme) with various choices of S. Tables 8 and 9 present the summary statistics of estimates based on returns sampled at every 180 unit intervals across 1763 trading days. Overall, the Quadrant and subsampled Quadrant correlation estimates are over the other two at each percentile and on average. The medians and means from Table 8 tell us that most stocks are correlated to the market with correlation coefficients around 0.5 except D and NEM. On the contrary, the estimated correlation between D and DUK is as high as 0.75, from Table 9. However, the correlation estimates for Sector Materials (LYB and NEM) are still close to zero. A lower sampling frequency, $S = 300$, is also taken into consideration, but it delivers a parallel pattern (in Appendix).

We compare the average of estimates against sampling frequency for all estimators in Figures 13 and 14 as representatives of high and less frequently traded stocks, respectively. Different from the plots in the previous simulation study, the Quadrant and subsampled Quadrant estimators appear in reverse pattern compared to the Kendall and Pearson. Across all selected stocks, the Quadrant-type estimators provide higher correlation estimates at any sampling frequency. They both converge to distinct stable levels swiftly. For the stocks in Information Technology and Materials shown in the figures, their correlation estimates converge within 180 sampling intervals. The estimates of the subsampled Quadrant are more smooth than those of the Quadrant since the former uses more information and gains efficiency. The Pearson estimator is always the closest to zero at the extremely high frequency and decreases most drastically due to the increased sampling frequency. The Kendall estimator is associated with an extra downward bias for the frequently traded stocks when we sample returns with S less than 60, as shown in Figure 13.



(a) AAPL v.s. SPY (b) AMD v.s. SPY (c) AAPL v.s. AMD

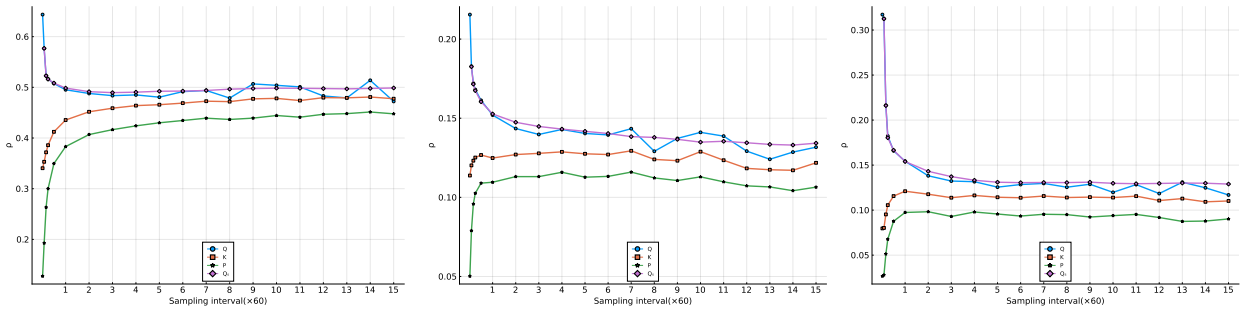
Figure 13: From left to right, we plot the average of correlation estimates between AAPL and SPY, AMD and SPY, and AAPL and AMD against sampling frequency.

Table 8: Summary statistics of correlation estimates between selected stocks and SPY based on returns sampled with $S = 180$.

Ticker	Q				K				Q_s				P			
	25th	Median	Mean	75th	25th	Median	Mean	75th	25th	Median	Mean	75th	25th	Median	Mean	75th
D	0.109	0.263	0.271	0.443	0.104	0.230	0.249	0.379	0.127	0.261	0.277	0.412	0.081	0.201	0.222	0.350
DUK	0.089	0.243	0.251	0.400	0.086	0.210	0.230	0.361	0.105	0.243	0.256	0.400	0.059	0.184	0.204	0.334
AMT	0.216	0.355	0.369	0.528	0.205	0.330	0.346	0.474	0.232	0.353	0.373	0.503	0.173	0.294	0.312	0.441
PLD	0.263	0.399	0.402	0.536	0.244	0.361	0.371	0.495	0.285	0.400	0.409	0.533	0.216	0.323	0.337	0.453
LYB	0.355	0.486	0.484	0.629	0.342	0.461	0.459	0.583	0.368	0.488	0.490	0.611	0.299	0.420	0.416	0.534
NEM	-0.024	0.142	0.140	0.304	-0.016	0.127	0.128	0.266	-0.002	0.150	0.145	0.286	-0.014	0.113	0.113	0.243
AAPL	0.568	0.681	0.665	0.780	0.548	0.653	0.644	0.748	0.579	0.683	0.672	0.775	0.510	0.619	0.608	0.720
AMD	0.309	0.486	0.467	0.638	0.291	0.426	0.426	0.579	0.334	0.477	0.472	0.622	0.241	0.376	0.384	0.527
AAL	0.309	0.460	0.451	0.603	0.300	0.419	0.419	0.536	0.329	0.457	0.456	0.575	0.263	0.375	0.379	0.491
UNP	0.399	0.528	0.522	0.645	0.379	0.501	0.501	0.620	0.411	0.530	0.529	0.647	0.342	0.464	0.465	0.587
JNJ	0.343	0.486	0.482	0.644	0.316	0.455	0.456	0.600	0.347	0.485	0.487	0.630	0.270	0.415	0.419	0.564
MRK	0.330	0.486	0.484	0.645	0.317	0.456	0.455	0.596	0.353	0.498	0.493	0.644	0.278	0.416	0.418	0.560
JPM	0.524	0.669	0.638	0.780	0.499	0.630	0.612	0.741	0.531	0.667	0.645	0.772	0.456	0.595	0.576	0.712
WFC	0.443	0.583	0.573	0.721	0.418	0.554	0.542	0.676	0.458	0.595	0.581	0.723	0.383	0.522	0.508	0.644
HAL	0.294	0.443	0.430	0.568	0.278	0.400	0.402	0.529	0.312	0.436	0.435	0.560	0.253	0.372	0.374	0.495
XOM	0.375	0.528	0.508	0.645	0.357	0.488	0.482	0.614	0.381	0.518	0.513	0.646	0.325	0.459	0.453	0.587
PG	0.263	0.407	0.411	0.568	0.242	0.377	0.386	0.532	0.273	0.417	0.419	0.564	0.205	0.342	0.352	0.498
WMT	0.304	0.443	0.442	0.588	0.279	0.404	0.414	0.544	0.317	0.435	0.450	0.585	0.236	0.356	0.373	0.503
TSLA	0.307	0.440	0.430	0.568	0.291	0.411	0.414	0.532	0.309	0.437	0.435	0.559	0.271	0.384	0.387	0.500
AMZN	0.492	0.609	0.611	0.741	0.493	0.606	0.596	0.710	0.517	0.620	0.619	0.727	0.457	0.574	0.565	0.681
DIS	0.411	0.565	0.552	0.716	0.402	0.532	0.530	0.676	0.431	0.570	0.561	0.704	0.356	0.495	0.493	0.645
FB	0.486	0.607	0.587	0.716	0.464	0.567	0.567	0.683	0.492	0.595	0.595	0.708	0.426	0.535	0.531	0.647

Table 9: Summary statistics of correlation estimates between selected stocks within the sector based on returns sampled with $S = 180$.

Ticker	Q				K				Q_s				P			
	25th	Median	Mean	75th	25th	Median	Mean	75th	25th	Median	Mean	75th	25th	Median	Mean	75th
Utilities	0.684	0.770	0.753	0.840	0.656	0.731	0.718	0.791	0.704	0.775	0.762	0.833	0.630	0.713	0.698	0.782
Real Estate	0.394	0.528	0.513	0.649	0.379	0.482	0.477	0.587	0.425	0.526	0.518	0.627	0.324	0.456	0.440	0.571
Materials	-0.014	0.135	0.132	0.278	-0.001	0.121	0.114	0.232	0.016	0.140	0.137	0.250	-0.027	0.103	0.093	0.218
Information Technology	0.239	0.402	0.394	0.568	0.212	0.364	0.368	0.516	0.259	0.400	0.403	0.553	0.163	0.329	0.330	0.485
Industrials	0.168	0.300	0.303	0.443	0.165	0.266	0.278	0.384	0.186	0.293	0.306	0.414	0.128	0.243	0.245	0.365
Health Care	0.443	0.565	0.553	0.681	0.412	0.523	0.519	0.630	0.458	0.568	0.560	0.663	0.375	0.495	0.485	0.609
Financials	0.658	0.750	0.735	0.836	0.636	0.722	0.706	0.790	0.671	0.760	0.744	0.827	0.610	0.702	0.684	0.779
Energy	0.476	0.603	0.586	0.716	0.449	0.558	0.550	0.664	0.489	0.598	0.591	0.702	0.418	0.539	0.528	0.649
Consumer Staples	0.263	0.399	0.399	0.539	0.250	0.360	0.367	0.485	0.292	0.400	0.406	0.522	0.204	0.329	0.331	0.458
Consumer Discretionary	0.216	0.355	0.363	0.500	0.226	0.342	0.352	0.476	0.238	0.346	0.365	0.482	0.207	0.326	0.328	0.453
Communication Services	0.172	0.329	0.331	0.486	0.170	0.300	0.312	0.454	0.198	0.333	0.339	0.483	0.136	0.260	0.273	0.417



(a) LYB v.s. SPY (b) NEM v.s. SPY (c) LYB v.s. NEM

Figure 14: From left to right, we plot the average of correlation estimates between LYB and SPY, NEM and SPY, and LYB and NEM against sampling frequency.

5.2 Intra-day Correlations and Market Beta Estimation

We now focus on the appearance of correlation estimators applied to returns within a local window for instantaneous correlations at each timestamp (shifted to origin) of a trading day. With the fixed window length l and sampling frequency S, the Quadrant, Kendall, and Pearson estimates at time t are defined as follows

$$\hat{q}_{Q,t} = \frac{1}{n}\sum_{k=1}^{n} \mathbf{1}_{\{\sum_{i=(k-1)S+1}^{kS} \Delta_i^T X^{(1)} \sum_{i=(k-1)s+1}^{kS} \Delta_i^T X^{(2)} > 0\}} \mathbf{1}_{\{t-l+S \le kS \le t\}}$$

$$\hat{q}_{K,t} = \frac{2}{n(n-1)} \sum_{k<k'}^{n} \mathbf{1}_{\{(\sum_{i=(k-1)S+1}^{kS} \Delta_i^T X^{(1)} - \sum_{i=(k'-1)S+1}^{k'S} \Delta_i^T X^{(1)})(\sum_{i=(k-1)S+1}^{kS} \Delta_i^T X^{(2)} - (\sum_{i=(k'-1)S+1}^{k'S} \Delta_i^T X^{(2)}) > 0\}}$$
$$\times \mathbf{1}_{\{t-l+s \le kS < k'S \le t\}}$$

with $\bullet = \sin(\pi(\hat{q}_{\bullet,t} - \frac{1}{2}))$, $\bullet = Q_t, K_t$,

$$P_t = \frac{\sum_{k=1}^{n} \sum_{i=(k-1)S+1}^{kS} \Delta_i^T X^{(1)} \sum_{i=(k-1)S+1}^{kS} \Delta_i^T X^{(2)} \mathbf{1}_{\{t-l+s \le kS \le t\}}}{\sqrt{\sum_{k=1}^{n} \sum_{i=(k-1)S+1}^{kS} \Delta_i^T X^{(1)2} \mathbf{1}_{\{t-l+s \le kS \le t\}} \sum_{k=1}^{n} \sum_{i=(k-1)S+1}^{kS} \Delta_i^T X^{(2)2} \mathbf{1}_{\{t-l+s \le kS \le t\}}}}$$

where $n = \lfloor T/S \rfloor$ and $t = l, \ldots, T$. The subsampled Quadrant estimator at time t uses all sub-series of length S within the local window $[t-l, t]$

$$\hat{q}_{Q_S,t} = \frac{1}{N}\sum_{k=1}^{N} \mathbf{1}_{\{\sum_{i=k}^{k+S-1} \Delta_i^T X^{(1)} \sum_{i=k}^{k+S-1} \Delta_i^T X^{(2)} > 0\}} \mathbf{1}_{\{t-l \le k < k+S-1 \le t\}}$$

with $Q_{S,t} = \sin(\pi(\hat{q}_{Q_S,t} - \frac{1}{2}))$ and $N = T - S + 1$.

We show the pattern of intra-day correlation of stocks from Sector Information Technology and Materials in Figures 15 and 16. The average correlation estimates for all four selected stocks are climbing over time. The range of estimates is in line with the previous integrated correlation estimates. Significantly, the intra-day correlations of NEM and SPY start from 5% and end up around a quarter, indicating a weak linear relationship between them as the integrated correlation is about 0.15, as pointed out in Table 8. The subsampled Quadrant estimates are slightly above the other three except for NEM and SPY, coinciding with the signature plot for integrated correlations. The smoothness of the subsampled Quadrant is evident again.

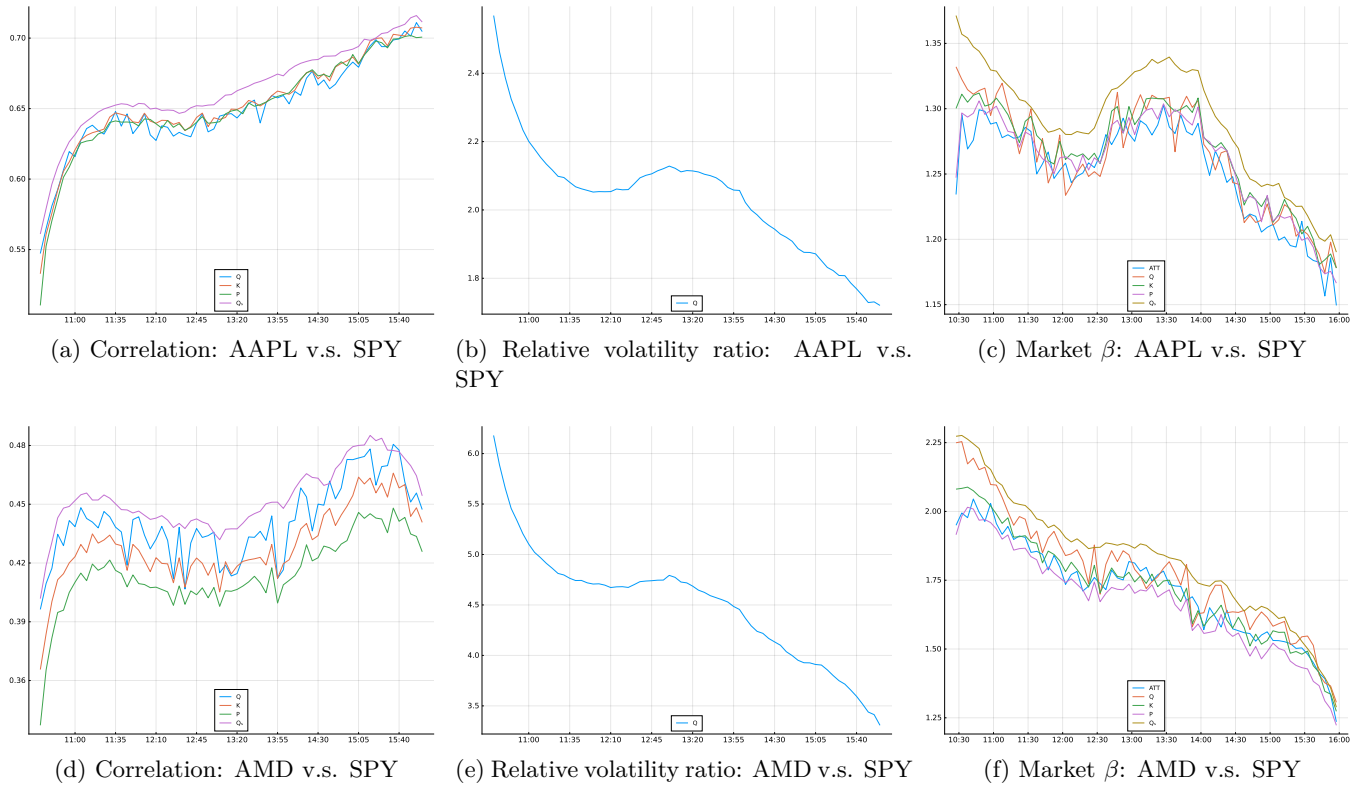
Additionally, we estimate the time-varying market beta at each timestamp, which captures the covariances of asset returns with the systematic risk factors. Normally, researchers measure betas by estimating linear regressions using daily (or lower frequency) market returns against asset returns. Alternatively, we utilize the correlation estimators by decomposing beta

$$\beta_t = \rho_t \times \frac{\sigma_t^{(j)}}{\sigma_t^{(0)}}$$

where j and 0 denote stocks and market, respectively. Consequently, we estimate ρ_t by applying the above estimators within a local window. To approximate the relative volatility ratio, we use the sum of the absolute value of subsampled returns

$$\text{RelV}_t = \frac{\sum_{k=1}^{N} \sum_{i=k}^{k+S-1} |\Delta_i^T X^{(j)}| \mathbf{1}_{\{\sum_{i=k}^{k+S-1} |\Delta_i^T X^{(j)}| < \nu^{(j)}\}} \mathbf{1}_{\{t-l \le k < k+S-1 \le t\}}}{\sum_{k=1}^{N} \sum_{i=k}^{k+S-1} |\Delta_i^T X^{(0)}| \mathbf{1}_{\{\sum_{i=k}^{k+S-1} |\Delta_i^T X^{(0)}| < \nu^{(0)}\}} \mathbf{1}_{\{t-l \le k < k+S-1 \le t\}}}$$

where $\nu^{(\bullet)}$ are thresholds for ruling out jumps. Specifically, for $\bullet = 0, j$, $\nu^{(\bullet)} = 4\sqrt{BV^{(\bullet)}}/T^{0.49}$, and $BV^{(\bullet)} = \frac{\pi}{2} \sum_{i=2}^{T} |\Delta_i^T X^{(\bullet)}||\Delta_{i-1}^T X^{(\bullet)}|$ is the so-called bipower variation serving as a nonparametric estimate of daily integrated volatility, see Barndorff-Nielsen and Shephard (2004*b*).



(a) Correlation: AAPL v.s. SPY

(b) Relative volatility ratio: AAPL v.s. SPY

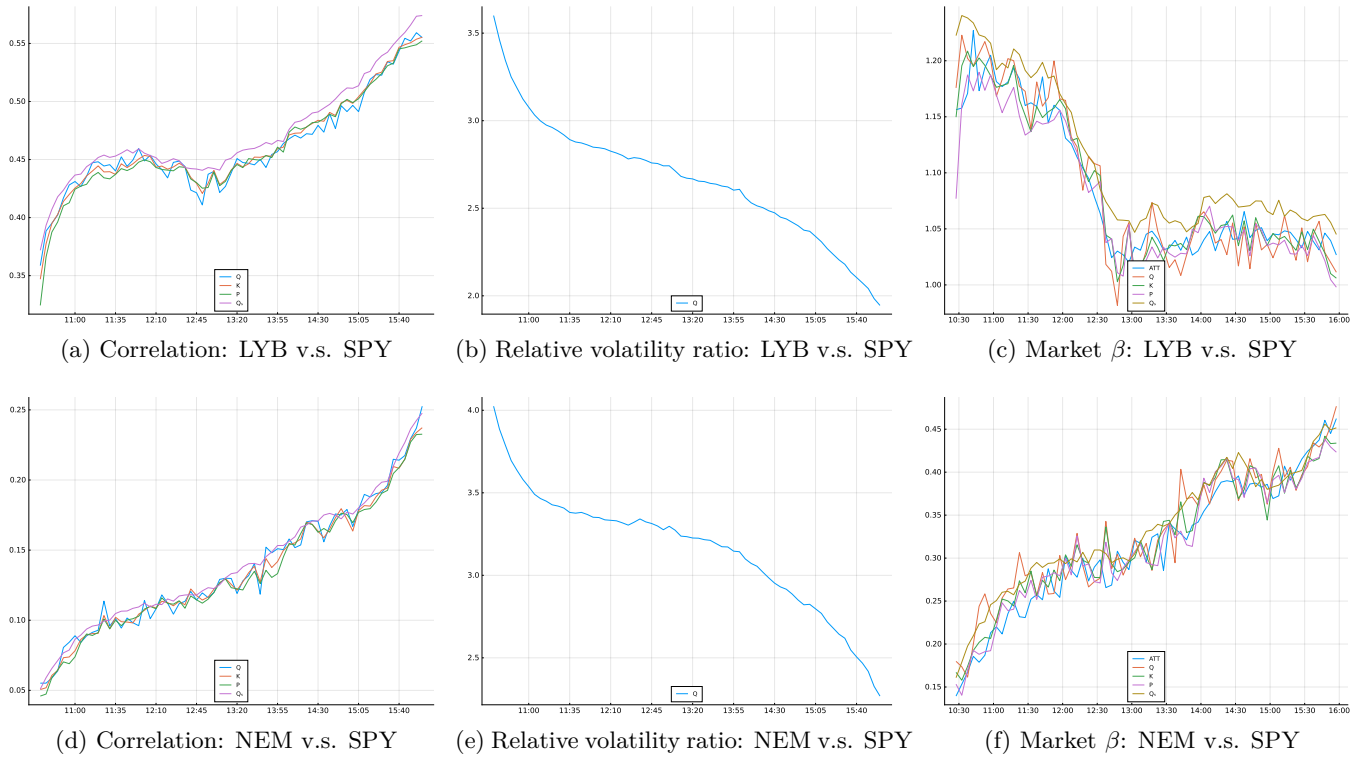
(c) Market β: AAPL v.s. SPY

(d) Correlation: AMD v.s. SPY

(e) Relative volatility ratio: AMD v.s. SPY

(f) Market β: AMD v.s. SPY

Figure 15: The left column plots the average correlation estimates between AAPL and AMD against the market on returns sampled at 180 unit intervals over local windows of 1 hour. The middle column plots the average of local relative volatility ratio estimates. The right column plots the average of estimated βs by combining the correlation estimates and local relative volatility ratio estimates and by ATT.



(a) Correlation: LYB v.s. SPY (b) Relative volatility ratio: LYB v.s. SPY (c) Market β: LYB v.s. SPY

(d) Correlation: NEM v.s. SPY (e) Relative volatility ratio: NEM v.s. SPY (f) Market β: NEM v.s. SPY

Figure 16: The left column plots the average correlation estimates between LYB and NEM against the market on returns sampled at 180 unit intervals over local windows of 1 hour. The middle column plots the average of local relative volatility ratio estimates. The right column plots the average of estimated βs by combining the correlation estimates and local relative volatility ratio estimates and by ATT.

Andersen et al. (2021) (ATT) estimated the time-varying betas using the ratio of covariance estimates of stocks and the market and variance estimates of the market within a local window, and they demonstrated the variation in patterns of betas across a trading day. In Figure 15, we compare the aforementioned correlation-based estimators and ATT's approach, where ATT's approach was extracted on returns sampled under the calendar scheme as they suggested. Although the estimated betas move up and down over time, they all exhibit a downward trend at most times, in contrast to the declining correlation estimates. Because of the subsampled Quadrant estimator, the estimated beta with it is a bit above the other approaches as well. Figure 16 compares a pair of low-frequency traded stocks, LYB and NEM. All betas evolve closely, and the distances shrink between the subsampled Q and the others. Unlike the other three stocks, NEM displays an increasing beta across a trading day. The opposite pattern is not driven by the stock's comparatively low trading frequency because LYB, which is traded less than 12% of a trading day on average, presents declining beta estimates like frequently traded stocks, such as AAPL. A comprehensive comparison of 22 stocks is presented in Appendix.

6 Robust Estimation of Correlation Matrices

We have introduced a robust estimator of a single correlation coefficient. A generalization to the multivariate case will be appealing in many applications. We can, obviously, estimate correlation coefficients pairwisely, but combining these unnecessarily result in a positive definite correlation matrix.

Thus consider the case where X is a d-dimensional random variable with $d > 2$, and we seek an estimate of its correlation matrix. For this situation, we propose to project the matrix with individual estimates, $\hat{R}$, onto the set of positive definite correlation matrices, denoted by S_+^d. In regards to the Frobenius norm, we seek a solution to the following optimization problem

$$\begin{aligned} &\min_{X \in \mathcal{R}^{d \times d}} ||X - \hat{R}||^2 \\ &\text{s.t.} X_{ii} = 1, \ i = 1, \ldots, d \cdot \\ &X \in S_+^d \end{aligned}$$

Higham (2002) shows a gradient method to find the solutions to the above problem with weighted Frobenius norms. In addition, Qi and Sun (2006) provide solutions based on a Newton-type method that enables quadratic convergence faster than the linear convergence of the gradient method proposed by Higham.

7 Concluding Remarks

We propose a novel subsampled Quadrant estimator that is robust to contaminated high-frequency financial data and reasonably efficient. We compared the Q_S estimator to the standard realized correlation, Pearson, and two other robust estimators: Quadrant and Kendall. Under the Itô semimartingale model for the price process, we have shown that Q_S is a consistent estimator for the integrated correlation. We compare the precision of the estimators in a Monte Carlo study, which covers scenarios when volatilities are constant or time-varying, and when efficient prices are observable or contaminated by noise, rounding error from discretization, and asynchronous trading.

Our main conclusion is that the subsampled Quadrant inherits the simplicity and consistency of the Quadrant, including the when facing time-varying volatility models. As a sign-concordance-based estimator, it is less sensitive to microstructure contamination than Pearson. As it takes advantage of using more information from the data with overlapping sub-series, the subsampled Quadrant improves

the efficiency of the Quadrant estimator and achieves higher efficiency than the Kendall estimator.

Empirically, we find that the subsampled Quadrant estimates present reversed patterns for increasing the sampling frequency compared to the Pearson and Kendall estimates. In another example, we propose a new approach to estimate the intra-day market betas, which embraces the superiority of the subsampled Quadrant estimator. Furthermore, we compare the intra-day motions of betas estimated by the new approach combined with all estimators of interest.

In the future, the following research questions will be of our interest. When both time-varying volatility and microstructure issues exist, all estimators are still biased at sampling frequency as low as every 600 sampling intervals, so a de-biasing correction on the estimators is expected. We look forward to seeing the robustness of the correlation matrix estimator built upon pairwise nonparametric correlation estimators in high dimensions. The nonparametric estimators can be extended as realized measures in a dynamic model of correlations.

Appendix of Proofs

Proof. [Symmetric elliptical distributions]: Assume (X, Y) are drawn from a symmetric elliptical distribution with location parameter μ and dispersion matrix Σ, then by Cambanis et al. (1981), (X, Y) have a stochastic representation:

$$\begin{pmatrix} X \\ Y \end{pmatrix} = \mu + R\Sigma^{\frac{1}{2}}U$$

where $R \geq 0$ is independent of U, and U is uniformly distributed on the unit circle, which indicates $U = (\cos\theta, \sin\theta)$ with density function $f(\theta) = \frac{1}{2\pi}$ for $-\pi \leq \theta \leq \pi$.

If $\mathbb{E}[R^2] < \infty$, then it can be verified that the correlation ρ between X and Y is the correlation coefficient in Σ. Let

$$\Sigma^{\frac{1}{2}} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

then $\rho = \frac{ac+bd}{\sqrt{a^2+b^2}\sqrt{c^2+d^2}}$.

$X > 0$ implies $a\cos\theta + b\sin\theta > 0$, or equivalently, $\sin(\theta + \phi_1) > 0$ with $\sin\phi_1 = \frac{a}{\sqrt{a^2+b^2}}$. Similarly, we have $\sin(\theta + \phi_2) > 0$ with $\sin\phi_2 = \frac{c}{\sqrt{c^2+d^2}}$ for $Y > 0$. Without loss of generality, we assume $\phi_1 \leq \phi_2$

and obtain the quadrant probability

$$\text{Prob}[X > 0, Y > 0] = \text{Prob}[-\phi_1 < \theta < \pi - \phi_2] = \frac{\pi - (\phi_2 - \phi_1)}{2\pi}.$$

By $\arccos \rho = \phi_2 - \phi_1$,

$$\text{Prob}[X > 0, Y > 0] = \frac{\frac{\pi}{2} + \arcsin \rho}{2\pi} = \frac{1}{4} + \frac{\arcsin \rho}{2\pi}.$$

□

Proof. [Theorem 1]: Write $Z_i^{(S)} = \mathbf{1}_{\{\sum_{\mathbf{j=0}}^{\mathbf{S-1}} \mathbf{x_{i+j}} \sum_{\mathbf{j=0}}^{\mathbf{S-1}} \mathbf{y_{i+j}} > \mathbf{0}\}}$, then

$$\hat{q}_S = \frac{1}{n - S + 1} \sum_{i=1}^{n-S+1} Z_i^{(S)}$$

sand

$$\text{Var}[\hat{q}_S] = (\frac{1}{n - S + 1})\text{Var}(Z_i^{(S)}) + (\frac{1}{n - S + 1})^2 2 \sum_{i<j} \text{Cov}(Z_i^{(S)}, Z_j^{(S)}).$$

□

Denote $x_i + x_{i+S-1}$ by $\tilde{x}_i$ and $y_i + y_{i+S-1}$ by $\tilde{y}_i$, and we assume that (x_i, y_i) is drawn from normal distribution with mean zero, unit marginal variances, and correlation ρ.

- $\text{Var}[Z_i^{(S)}] = \Pr(\tilde{x}_i \tilde{y}_i > 0)[1 - \Pr(\tilde{x}_i \tilde{y}_i > 0)] = q(1 - q) = \frac{1}{4} - \frac{\arcsin^2 \rho}{\pi^2}$, where $q = \frac{\arcsin \rho}{\pi} + \frac{1}{2}$.
- If $j > i + S - 1$ then $\text{Cov}(Z_i^{(S)}, Z_j^{(S)}) = 0$.
- If $i < j \leq i + S - 1$, then

$$\begin{aligned}
\text{Cov}(Z_i^{(S)}, Z_j^{(S)}) =& \mathbb{E} Z_i^{(S)} Z_j^{(S)} - \mathbb{E} Z_i^{(S)} \mathbb{E} Z_j^{(S)} \\
=& \Pr(\tilde{x}_i \tilde{y}_i > 0, \tilde{x}_j \tilde{y}_j > 0) - q^2 \\
=& 2\Pr(\tilde{x}_i > 0, \tilde{y}_i > 0, \tilde{x}_j > 0, \tilde{y}_j > 0) + 2\Pr(\tilde{x}_i > 0, \tilde{y}_i > 0, \tilde{x}_j < 0, \tilde{y}_j < 0) - q^2.
\end{aligned}$$

Next, we compute above two probabilities following the procedure proposed by Plackett (1954). Suppose a normally distributed random vector is associated with correlation matrix R and zero means, then we denote the integral over the first positive quadrant by $\Gamma(R)$.

a) For $1 < j \leq S$, the correlation matrix of vector $(\tilde{x}_1, \tilde{y}_1, \tilde{x}_j, \tilde{y}_j)$ is

$$R^{++} = \begin{pmatrix} 1 & \rho & \frac{S-j+1}{S} & \frac{S-j+1}{S}\rho \\ \rho & 1 & \frac{S-j+1}{S}\rho & \frac{S-j+1}{S} \\ \frac{S-j+1}{S} & \frac{S-j+1}{S}\rho & 1 & \rho \\ \frac{S-j+1}{S}\rho & \frac{S-j+1}{S} & \rho & 1 \end{pmatrix}$$

and its inverse matrix is

$$C^{++} = \frac{1}{(1-\rho^2)[1-(\frac{S-j+1}{S})^2]} \begin{pmatrix} 1 & -\rho & -\frac{S-j+1}{S} & \frac{S-j+1}{S}\rho \\ -\rho & 1 & \frac{S-j+1}{S}\rho & -\frac{S-j+1}{S} \\ -\frac{S-j+1}{S} & \frac{S-j+1}{S}\rho & 1 & -\rho \\ \frac{S-j+1}{S}\rho & -\frac{S-j+1}{S} & -\rho & 1 \end{pmatrix}.$$

Write $X^{++}_{pq} = \frac{C^{++}_{pq}}{\sqrt{C^{++}_{pp}C^{++}_{qq}}}$, then we form following matrix

$$X^{++} = \begin{pmatrix} 1 & -\rho & -\frac{S-j+1}{S} & \frac{S-j+1}{S}\rho \\ -\rho & 1 & \frac{S-j+1}{S}\rho & -\frac{S-j+1}{S} \\ -\frac{S-j+1}{S} & \frac{S-j+1}{S}\rho & 1 & -\rho \\ \frac{S-j+1}{S}\rho & -\frac{S-j+1}{S} & -\rho & 1 \end{pmatrix}.$$

Moreover, we denote matrix R^{++} with $\rho = 0$ and $\rho = 1$ by K^{++} and L^{++} respectively, and let $\mathcal{S}$ be the set containing elements' positions in the upper triangle on which matrix K^{++} and L^{++} have different values. So $\mathcal{S} = \{(1,2),(1,4),(2,3),(3,4)\}$. Then

$$\Gamma(R^{++}) = \Gamma(K^{++}) + \frac{1}{4\pi^2} \sum_{(i,j)\in\mathcal{S}} \int_0^\rho \frac{L^{++}_{ij} \arccos(X^{++}_{pq})dt}{\sqrt{1-(L^{++}_{ij})^2t^2}}$$

where $p < q$ and are different from (i,j).

$$\begin{aligned} \Gamma(K^{++}) =& \Pr(\tilde{x}_1 > 0, \tilde{x}_j > 0) \times \Pr(\tilde{y}_1 > 0, \tilde{y}_j > 0) \\ =& [\frac{1}{4} + \frac{\arcsin(\frac{S-j+1}{S})}{2\pi}]^2 = \frac{1}{16} + \frac{\arcsin(\frac{S-j+1}{S})}{4\pi} + \frac{\arcsin^2(\frac{S-j+1}{S})}{4\pi^2}. \end{aligned}$$

$$\int_0^\rho \frac{L_{12}^{++}\arccos(X_{34}^{++})dt}{\sqrt{1-(L_{12}^{++})^2t^2}} = \int_0^\rho \frac{L_{34}^{++}\arccos(X_{12}^{++})dt}{\sqrt{1-(L_{34}^{++})^2t^2}} = \int_0^\rho \frac{\arccos(-t)dt}{\sqrt{1-t^2}}$$
$$= \int_0^\rho \frac{\frac{\pi}{2}-\arcsin(-t)dt}{\sqrt{1-t^2}}$$
$$\overset{(t=\sin\theta)}{=} \int_0^{\arcsin\rho} \frac{\frac{\pi}{2}+\theta}{\cos\theta}\cos\theta d\theta$$
$$= \frac{\pi}{2}\arcsin\rho + \frac{1}{2}\arcsin^2\rho$$

and

$$\int_0^\rho \frac{L_{14}^{++}\arccos(X_{23}^{++})dt}{\sqrt{1-(L_{14}^{++})^2t^2}} = \int_0^\rho \frac{L_{23}^{++}\arccos(X_{14}^{++})dt}{\sqrt{1-(L_{23}^{++})^2t^2}} = \int_0^\rho \frac{\frac{S-j+1}{S}\arccos(\frac{S-j+1}{S}t)dt}{\sqrt{1-(\frac{S-j+1}{S})^2t^2}}$$
$$= \frac{S-j+1}{S}\int_0^\rho \frac{\frac{\pi}{2}-\arcsin(\frac{S-j+1}{S}t)dt}{\sqrt{1-(\frac{S-j+1}{S})^2t^2}}$$
$$\overset{(\frac{S-j+1}{S}t=\sin\theta)}{=} \int_0^{\arcsin\rho} \frac{\frac{\pi}{2}-\theta}{\cos\theta}\cos\theta d\theta$$
$$= \frac{\pi}{2}\arcsin(\frac{S-j+1}{S}\rho) - \frac{1}{2}\arcsin^2(\frac{S-j+1}{S}\rho).$$

Then

$$\Gamma(R^{++}) = \frac{1}{16} + \frac{1}{4\pi}[\arcsin\rho + \arcsin(\frac{S-j+1}{S}) + \arcsin(\frac{S-j+1}{S}\rho)]$$
$$+ \frac{1}{4\pi^2}[\arcsin^2\rho + \arcsin^2(\frac{S-j+1}{S}) - \arcsin^2(\frac{S-j+1}{S}\rho)].$$

b) Let $\tilde{u}_j = -\tilde{x}_j$ and $\tilde{v}_j = -\tilde{y}_j$ then the correlation matrix of vector $(\tilde{x}_1, \tilde{y}_1, \tilde{u}_j, \tilde{v}_j)$ is

$$R^{+-} = \begin{pmatrix} 1 & \rho & -\frac{S-j+1}{S} & -\frac{S-j+1}{S}\rho \\ \rho & 1 & -\frac{S-j+1}{S}\rho & -\frac{S-j+1}{S} \\ -\frac{S-j+1}{S} & -\frac{S-j+1}{S}\rho & 1 & \rho \\ -\frac{S-j+1}{S}\rho & -\frac{S-j+1}{S} & \rho & 1 \end{pmatrix}$$

and our goal is to compute $\Gamma(R^{+-})$. The corresponding X matrix is

$$X^{+-} = \begin{pmatrix} 1 & -\rho & \frac{S-j+1}{s} & -\frac{S-j+1}{S}\rho \\ -\rho & 1 & -\frac{S-j+1}{S}\rho & \frac{S-j+1}{S} \\ \frac{S-j+1}{s} & -\frac{S-j+1}{S}\rho & 1 & -\rho \\ -\frac{S-j+1}{S}\rho & \frac{S-j+1}{s} & -\rho & 1 \end{pmatrix}.$$

After forming matrices K^{+-} and L^{+-} based on R^{+-}, we derive the set $\mathcal{S}$ as $\{(1,2),(1,4),(2,3),(3,4)\}$.

First, we have

$$\begin{aligned}\Gamma(K^{+-}) =&\Pr(\tilde{x}_1 > 0, \tilde{u}_j > 0) \times \Pr(\tilde{y}_1 > 0, \tilde{v}_j > 0)\\ =&[\frac{1}{4} + \frac{\arcsin(-\frac{S-j+1}{S})}{2\pi}]^2 = \frac{1}{16} - \frac{\arcsin(\frac{S-j+1}{S})}{4\pi} + \frac{\arcsin^2(\frac{S-j+1}{S})}{4\pi^2}.\end{aligned}$$

Second,

$$\int_0^\rho \frac{L_{12}^{++}\arccos(X_{34}^{++})dt}{\sqrt{1-(L_{12}^{++})^2t^2}} = \int_0^\rho \frac{L_{34}^{++}\arccos(X_{12}^{++})dt}{\sqrt{1-(L_{34}^{++})^2t^2}} = \int_0^\rho \frac{\arccos(-t)dt}{\sqrt{1-t^2}} = \frac{\pi}{2}\arcsin\rho + \frac{1}{2}\arcsin^2\rho$$

and

$$\begin{aligned}\int_0^\rho \frac{L_{14}^{++}\arccos(X_{23}^{++})dt}{\sqrt{1-(L_{14}^{++})^2t^2}} = \int_0^\rho \frac{L_{23}^{++}\arccos(X_{14}^{++})dt}{\sqrt{1-(L_{23}^{++})^2t^2}} =& \int_0^\rho \frac{-\frac{S-j+1}{S}\arccos(-\frac{S-j+1}{S}t)dt}{\sqrt{1-(\frac{S-j+1}{S})^2t^2}}\\ =& -\frac{S-j+1}{S}\int_0^\rho \frac{\frac{\pi}{2}-\arcsin(-\frac{S-j+1}{S}t)dt}{\sqrt{1-(\frac{S-j+1}{S})^2t^2}}\\ \overset{(\frac{S-j+1}{S}t=\sin\theta)}{=}& -\int_0^{\arcsin\rho}\frac{\frac{\pi}{2}+\theta}{\cos\theta}\cos\theta d\theta\\ =& -\frac{\pi}{2}\arcsin(\frac{S-j+1}{S}\rho) - \frac{1}{2}\arcsin^2(\frac{S-j+1}{S}\rho).\end{aligned}$$

So

$$\begin{aligned}\Gamma(R^{+-}) =&\frac{1}{16} + \frac{1}{4\pi}[\arcsin\rho - \arcsin(\frac{S-j+1}{S}) - \arcsin(\frac{S-j+1}{S}\rho)]\\ &+\frac{1}{4\pi^2}[\arcsin^2\rho + \arcsin^2(\frac{S-j+1}{S}) - \arcsin^2(\frac{S-j+1}{S}\rho)].\end{aligned}$$

Thus,

$$\begin{aligned}\mathrm{Var}[\hat{q}_s] =&\frac{1}{n-S+1}(\frac{1}{4} - \frac{\arcsin^2\rho}{\pi^2}) + 2(\frac{1}{n-S+1})^2\sum_{h=1}^{S-1}(n-S+1-h)\mathrm{Cov}(Z_i^{(S)}, Z_{i+h}^{(S)})\\ =&\frac{1}{n-S+1}(\frac{1}{4} - \frac{\arcsin^2\rho}{\pi^2}) + 2(\frac{1}{n-S+1})^2\\ &\times\sum_{h=1}^{S-1}(n-S+1-h)\frac{1}{\pi^2}[\arcsin^2(\frac{S-h}{S}) - \arcsin^2(\frac{S-h}{S}\rho)].\end{aligned}$$

Therefore, for fixed s,

$$\mathrm{Avar}[\hat{q}_S] = \frac{1}{4} - \frac{\arcsin^2\rho}{\pi^2} + \frac{2}{\pi^2}\sum_{h=1}^{S-1}[\arcsin^2(\frac{S-h}{S}) - \arcsin^2(\frac{S-h}{S}\rho)].$$

Proof. [Proposition 3]: For $\tilde{X}_i = \sum_{j=i}^{i+S-1} X_j$ and $\tilde{Y}_i = \sum_{j=i}^{i+S-1} Y_j$, they follows a bivariate normal

distribution with variances S and covariance $S\rho$ since $(X_i, Y_i) \sim \Phi_\rho$. For each pair (X_i, Y_i), there is probability ε of they are exactly at point (x_0, y_0). Define a function G in terms of ε and another function g

$$G(\varepsilon, g) = \sum_{h=0}^{S} (1-\varepsilon)^{S-h} \varepsilon^h \binom{S}{h} g(h) \quad \text{and} \quad G_\varepsilon(0, g) = \frac{\partial G}{\partial \varepsilon}\Big|_{\varepsilon=0} = s[g(1) - g(0)].$$

□

1. The Pearson estimator converges to

$$R_p = \frac{\mathbb{E}[\tilde{X}\tilde{Y}] - \mathbb{E}[\tilde{X}]\mathbb{E}[\tilde{Y}]}{\sqrt{(\mathbb{E}[\tilde{X}^2] - \mathbb{E}[\tilde{X}]^2)(\mathbb{E}[\tilde{Y}^2] - \mathbb{E}[\tilde{Y}]^2)}}$$

   in probability. Under the contaminated distribution of $(\tilde{X}, \tilde{Y})$,

$$\mathbb{E}[\tilde{X}\tilde{Y}] = G(\varepsilon, U),\ \mathbb{E}[\tilde{X}] = G(\varepsilon, V_x),\ \text{and } \mathbb{E}[\tilde{X}^2] = G(\varepsilon, W_x)$$

   with

$$U(h) = (S-h)\rho + hx_0y_0$$
$$V_x(h) = hx_0$$
$$W_x(h) = S - h + hx_0^2.$$

   Then

$$G(0, U) = S\rho \qquad G(0, V_x) = 0 \qquad G(0, W_x) = S$$
$$G_\varepsilon(0, U) = S(x_0y_0 - \rho) \quad G_\varepsilon(0, V_\varepsilon) = Sx_0 \quad G_\varepsilon(0, W_x) = S(x_0^2 - 1).$$

   The influence function of the Pearson under the sparse sampling method is

$$\text{IF}((x_0, y_0), R_P, \Phi_\rho) = \frac{S(x_0y_0 - \rho)S - S\rho\frac{1}{2}S(x_0^2 + y_0^2 - 2)}{S^2} = x_0y_0 - \rho(\frac{x_0^2 + y_0^2}{2}).$$

2. The Kendall probability estimator is associated with the functional

$$\tilde{R}_K = \mathbb{E}[(\tilde{X}_1 - \tilde{X}_2)(\tilde{Y}_1 - \tilde{Y}_2) > 0] = \tilde{G}(\varepsilon, F_K)$$

with

$$\tilde{G}(\varepsilon, F_K) = \sum_{h_1=0}^{S-h_1} \sum_{h_2=0}^{S-h_2} (1-\varepsilon)^{2S-h_1-h_2} \varepsilon^{h_1+h_2} \binom{S}{h_1} \binom{S}{h_2} F_K(h_1, h_2)$$

and

$$\begin{aligned} F_K(h_1, h_2) &= \text{Prob}[(\sum_{i=1}^{S-h_1} X_i - h_1 x_0 - \sum_{i=1}^{S-h_2} X_{S+i} + h_2 x_0)(\sum_{i=1}^{S-h_1} Y_i - h_1 y_0 - \sum_{i=1}^{S-h_2} Y_{S+i} + h_2 y_0) > 0] \\ &= \text{Prob}[(\sum_{i=1}^{S-h_1} X_i - \sum_{i=1}^{S-h_2} X_{S+i} - (h_1 - h_2) x_0)(\sum_{i=1}^{S-h_1} Y_i - \sum_{i=1}^{S-h_2} Y_{S+i} - (h_1 - h_2) y_0) > 0] \\ &= 2\Phi(\frac{(h_1 - h_2)x_0}{\sqrt{2S - h_1 - h_2}}, \frac{(h_1 - h_2)y_0}{\sqrt{2S - h_1 - h_2}}) - \Phi(\frac{(h_1 - h_2)x_0}{\sqrt{2S - h_1 - h_2}}) - \Phi(\frac{(h_1 - h_2)y_0}{\sqrt{2S - h_1 - h_2}}) + 1 \end{aligned}$$

when $h_1 + h_2 \geq 1$ and $F_K(0,0) = q$. Thus, the Kendall estimator's influence function is

$$\text{IF}((x,y), R_K, \Phi_\rho) = \pi\sqrt{1-\rho^2} 2S \Big[2\Phi(\frac{x_0}{\sqrt{2S-1}}, \frac{y_0}{\sqrt{2S-1}}) - \Phi(-\frac{x_0}{\sqrt{2S-1}}) - \Phi(-\frac{y_0}{\sqrt{2S-1}}) + 1 - q \Big]$$

3. Note that the statistical functional of Quadrant probability estimator at the contaminated distribution is

$$\tilde{R}_S = G(\varepsilon, F_Q)$$

with

$$\begin{aligned} F_Q(h) &= \text{Prob}[(\sum_{i=1}^{S-h} X_i + h x_0)(\sum_{i=1}^{S-h} Y_i + h y_0)] \\ &= 2\Phi(\frac{h x_0}{\sqrt{S-h}}, \frac{h y_0}{\sqrt{S-h}}) - \Phi(\frac{h x_0}{\sqrt{S-h}}) - \Phi(\frac{h y_0}{\sqrt{S-h}}) + 1 \end{aligned}$$

when $h \geq 1$ and $F_Q(0) = q$. Then the influence function of the Quadrant estimator with length S is

$$\text{IF}((x,y), R_S, \Phi_\rho) = \pi\sqrt{1-\rho^2} S \Big[2\Phi(\frac{x_0}{\sqrt{S-1}}, \frac{y_0}{\sqrt{S-1}}) - \Phi(\frac{x_0}{\sqrt{S-1}}) - \Phi(\frac{y_0}{\sqrt{S-1}}) + 1 - q \Big].$$

Lemma 1. *Suppose Assumption 1 holds, for all* $1 \leq i \leq T - S + 1$,

$$\frac{1}{T-S+1} \sum_{i=1}^{T-S+1} \mathbb{E}[\zeta_i] \longrightarrow q \text{ as } T \to \infty.$$

Proof. For any $1 \le i \le T-S+1$ and $j=1,2$,

$$\sum_{j=0}^{S-1} \Delta_{j+i}^{T} X^{(i)} = \int_{i-1}^{i+S-1} \sigma_u^{(j)} dW_u^{(j)} \stackrel{def}{=} Z^{(j)}. \tag{6}$$

A partition $\{t_k^n\}_{k=0}^n$ of $[i-1, i+S-1)$ satisfies $i-1 = t_0^n < t_1^n < \cdots < t_n^n = i+S-1$ with $\max_k |t_k^n - t_{k-1}^n| \to 0$ when $n \to \infty$. Then Equation (6) is also defined as the limit of Riemann sums over above partitions, i.e. as $n \to \infty$

$$Z_n^{(j)} \stackrel{def}{=} \sum_{k=1}^{n} \sigma_{t_{k-1}^n}^{(j)} (W_{t_k^n}^{(j)} - W_{t_{k-1}^n}^{(j)}) \xrightarrow{p} Z^{(j)}.$$

Since Z_n is normally distributed, then

$$\begin{aligned}
\text{Prob}(Z_n^{(1)} Z_n^{(2)} > 0) &= \int_{\mathbb{R}} \int_{\mathbb{R}} \mathbf{1}_{\{z_1 z_2 > 0\}} f_n(z_1, z_2) dz_1 dz_2 \\
&= \frac{1}{\pi} \arcsin\left(\rho \frac{\sum_{k=1}^n \sigma_{t_{k-1}^n}^{(1)} \sigma_{t_{k-1}^n}^{(2)} (t_k^n - t_{k-1}^n)}{\sqrt{\sum_{k=1}^n \sigma_{t_{k-1}^n}^{(1)2} (t_k^n - t_{k-1}^n) \sum_{k=1}^n \sigma_{t_{k-1}^n}^{(2)2} (t_k^n - t_{k-1}^n)}}\right) + \frac{1}{2}
\end{aligned}$$

where f_n is the density function of Z_n. Let f be the density function of Z, then we have f_n converging to f pointwisely and then $\mathbf{1}_{\{z_1 z_2 > 0\}} f_n \to \mathbf{1}_{\{z_1 z_2 > 0\}} f$. By $\lim_{n\to\infty} \int_{\mathbb{R}^2} f_n(Z) dZ = 1 = \int_{\mathbb{R}^2} f(Z) dZ$ and dominating convergence theorem,

$$\begin{aligned}
\text{Prob}(Z^{(1)} Z^{(2)} > 0) &= \lim_{n\to\infty} \text{Prob}(Z_n^{(1)} Z_n^{(2)} > 0) \\
&= \frac{1}{\pi} \arcsin\left(\rho \frac{\int_{(i-1)/T}^{(i+S-1)/T} \sigma_u^{(1)} \sigma_u^{(2)} du}{\sqrt{\int_{(i-1)/T}^{(i+S-1)/T} \sigma_u^{(1)2} du \int_{(i-1)/T}^{(i+S-1)/T} \sigma_u^{(2)2} du}}\right) + \frac{1}{2}.
\end{aligned}$$

Moreover, there are $\{\theta_{T,i}\}_{i=1}^T$ lying between $[\frac{\int_{(i-1)/T}^{(i+S-1)/T} \sigma_u^{(1)} \sigma_u^{(2)} du}{\sqrt{\int_{(i-1)/T}^{(i+S-1)/T} \sigma_u^{(1)2} du \int_{(i-1)/T}^{(i+S-1)/T} \sigma_u^{(2)2} du}}, 1]$ for each i such that

$$\begin{aligned}
|\mathbb{E}[\zeta_i] - q| &= \frac{1}{\pi} \frac{\rho}{\sqrt{1-\rho^2\theta_{T,i}^2}} \left| \frac{\int_{(i-1)/T}^{(i+S-1)/T} \sigma_u^{(1)} \sigma_u^{(2)} du}{\sqrt{\int_{(i-1)/T}^{(i+S-1)/T} \sigma_u^{(1)2} du \int_{(i-1)/T}^{(i+S-1)/T} \sigma_u^{(2)2} du}} - 1 \right| \\
&\le \frac{1}{\pi} \frac{\rho}{\sqrt{1-\rho^2}} \left| \frac{\int_{(i-1)/T}^{(i+S-1)/T} \sigma_u^{(1)} \sigma_u^{(2)} du}{\sqrt{\int_{(i-1)/T}^{(i+S-1)/T} \sigma_u^{(1)2} du \int_{(i-1)/T}^{(i+S-1)/T} \sigma_u^{(2)2} du}} - 1 \right| \\
&\le \varepsilon.
\end{aligned}$$

The last step holds if we choose large enough T satisfying $S/T < \delta$ which is defined as in Assumption

1 for given $\varepsilon > 0$. Furthermore,

$$\left| \frac{1}{T-S+1} \sum_{i=1}^{T-S+1} \mathbb{E}[\zeta_i] - q \right| < \varepsilon.$$

□

Proof. [Theorem 2]: Because ζ_i is a binary variable, we have

$$|\mathrm{Cov}(\zeta_i, \zeta_{i+h})| \leq 1,\ 0 \leq h \leq S-1$$

and

$$\mathrm{Cov}(\zeta_i, \zeta_{i+h}) = 0,\ h \geq S.$$

Thus

$$\begin{aligned}
\mathrm{Var}(\frac{1}{T-S+1} \sum_{i=1}^{T-S+1} \zeta_i) =& \frac{\sum_{i=1}^{T-S+1} \mathrm{Var}(\zeta_i)}{(T-S+1)^2} + \frac{1}{T^2} \sum_{h=1}^{s-1} (T-S+1)\mathrm{Cov}(\zeta_1, \zeta_1+h) \\
&\leq \frac{1}{T-S+1} + \frac{(2T-S)(S-1)}{(T-S+1)^2} \\
&= o(1).
\end{aligned}$$

For any $\varepsilon > 0$,

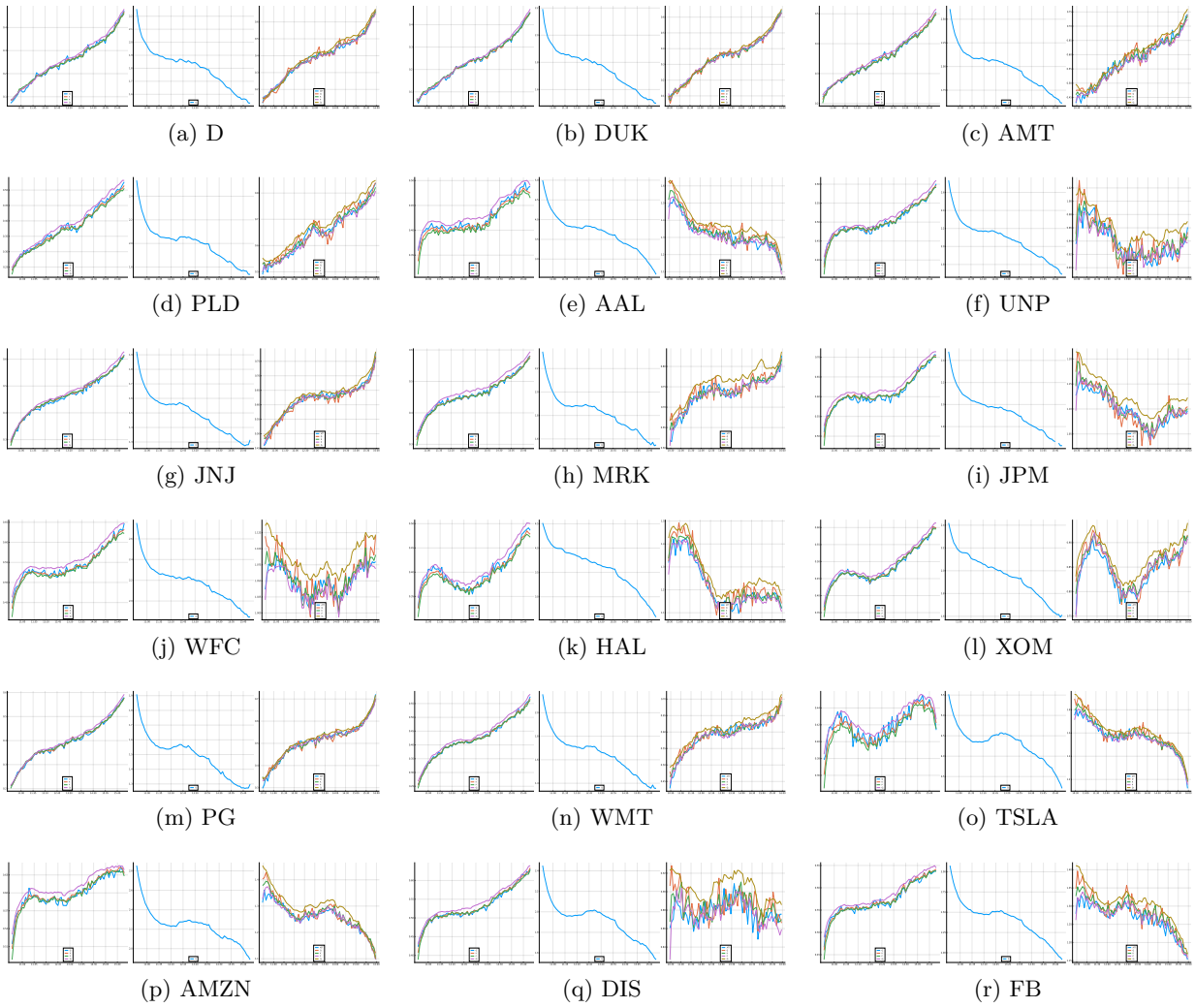
$$\begin{aligned}
\mathrm{Prob}(|\hat{q}_S - q| > \varepsilon) &= \mathrm{Prob}(|\hat{q}_S - \frac{1}{T-S+1} \sum_{i=1}^{T-S+1} \mathbb{E}[\zeta_i] + \frac{1}{(T-S+1)} \sum_{i=1}^{T-S+1} \mathbb{E}[\zeta_i] - q| > \varepsilon) \\
&\leq \mathrm{Prob}(|\hat{q}_S - \frac{1}{T-S+1} \sum_{i=1}^{T-S+1} \mathbb{E}[\zeta_i]| + |\frac{1}{T-S+1} \sum_{i=1}^{T-S+1} \mathbb{E}[\zeta_i] - q| > \varepsilon) \\
&\leq \frac{\mathrm{Var}(\hat{q}_S)}{(\varepsilon - |\frac{1}{T-S+1} \sum_{i=1}^{T-S+1} \mathbb{E}[\zeta_i] - q|)^2}
\end{aligned}$$

The last step is the result of Chebyshev's inequality. Choose $T > s/\delta$ with δ defined in Assumption 1, then by Lemma 1

$$|\frac{1}{T-S+1} \sum_{i=1}^{T-S+1} \mathbb{E}[\zeta_i] - q| = 0.$$

Hence $\hat{q}_S \xrightarrow{p} q$. By continuous mapping theorem and the mapping between q and ρ, we have $Q_S \xrightarrow{p} \rho$. □

Appendix of Figures



(a) D (b) DUK (c) AMT

(d) PLD (e) AAL (f) UNP

(g) JNJ (h) MRK (i) JPM

(j) WFC (k) HAL (l) XOM

(m) PG (n) WMT (o) TSLA

(p) AMZN (q) DIS (r) FB

Figure 17: For each stock, plot average of correlation estimates (left), relative volatility ratio estimates (middle), and market β estimates (right) against the market with returns sampled at 180 unit intervals over local windows of 1 hour.

Table 10: Summary statistics of correlation estimates between selected stocks and SPY based on returns sampled with $s = 300$.

Ticker	Q				K				Q_s				P			
	25th	Median	Mean	75th	25th	Median	Mean	75th	25th	Median	Mean	75th	25th	Median	Mean	75th
D	0.121	0.278	0.270	0.429	0.095	0.228	0.249	0.393	0.117	0.261	0.274	0.419	0.077	0.206	0.224	0.367
DUK	0.040	0.245	0.248	0.429	0.070	0.215	0.230	0.387	0.101	0.244	0.255	0.406	0.042	0.188	0.205	0.350
AMT	0.200	0.355	0.366	0.524	0.208	0.342	0.354	0.496	0.236	0.364	0.377	0.513	0.179	0.311	0.324	0.460
PLD	0.262	0.415	0.398	0.568	0.246	0.375	0.377	0.510	0.272	0.398	0.405	0.539	0.214	0.341	0.345	0.469
LYB	0.355	0.500	0.481	0.632	0.333	0.471	0.466	0.593	0.368	0.492	0.492	0.622	0.308	0.437	0.430	0.560
NEM	-0.040	0.121	0.140	0.335	-0.017	0.130	0.128	0.275	-0.012	0.144	0.142	0.287	-0.037	0.115	0.113	0.256
AAPL	0.568	0.693	0.662	0.799	0.548	0.660	0.647	0.763	0.580	0.679	0.671	0.781	0.501	0.618	0.608	0.730
AMD	0.318	0.500	0.467	0.632	0.302	0.437	0.436	0.589	0.334	0.485	0.476	0.622	0.248	0.396	0.395	0.546
AAL	0.278	0.439	0.441	0.588	0.293	0.425	0.423	0.557	0.322	0.450	0.451	0.579	0.262	0.392	0.387	0.516
UNP	0.355	0.512	0.515	0.693	0.383	0.508	0.504	0.635	0.413	0.532	0.530	0.655	0.352	0.482	0.474	0.606
JNJ	0.317	0.500	0.473	0.632	0.314	0.460	0.456	0.609	0.334	0.485	0.484	0.633	0.263	0.421	0.421	0.583
MRK	0.291	0.500	0.470	0.632	0.304	0.462	0.453	0.606	0.338	0.490	0.485	0.639	0.274	0.425	0.420	0.575
JPM	0.500	0.632	0.624	0.799	0.493	0.637	0.612	0.751	0.522	0.661	0.638	0.775	0.453	0.599	0.578	0.723
WFC	0.429	0.568	0.560	0.729	0.413	0.554	0.540	0.689	0.448	0.585	0.573	0.717	0.379	0.519	0.508	0.656
HAL	0.278	0.429	0.415	0.568	0.271	0.403	0.402	0.540	0.297	0.431	0.427	0.561	0.253	0.377	0.376	0.507
XOM	0.355	0.500	0.499	0.652	0.346	0.493	0.484	0.628	0.373	0.518	0.506	0.644	0.326	0.467	0.456	0.603
PG	0.234	0.429	0.410	0.574	0.237	0.388	0.387	0.545	0.266	0.407	0.413	0.564	0.190	0.356	0.354	0.508
WMT	0.278	0.437	0.443	0.628	0.279	0.411	0.419	0.563	0.306	0.437	0.447	0.595	0.227	0.372	0.377	0.523
TSLA	0.278	0.429	0.430	0.568	0.286	0.422	0.417	0.548	0.301	0.435	0.435	0.574	0.265	0.395	0.391	0.519
AMZN	0.500	0.632	0.614	0.749	0.494	0.622	0.605	0.721	0.521	0.630	0.626	0.744	0.472	0.591	0.574	0.698
DIS	0.424	0.568	0.547	0.707	0.392	0.551	0.534	0.684	0.424	0.565	0.557	0.709	0.355	0.510	0.500	0.657
FB	0.453	0.581	0.586	0.722	0.458	0.575	0.569	0.690	0.483	0.601	0.594	0.714	0.421	0.542	0.534	0.663

Table 11: Summary statistics of correlation estimates between selected stocks within the sector based on returns sampled with $s = 300$.

Ticker	Q				K				Q_s				P			
	25th	Median	Mean	75th	25th	Median	Mean	75th	25th	Median	Mean	75th	25th	Median	Mean	75th
Utilities	0.652	0.749	0.738	0.845	0.661	0.740	0.724	0.808	0.692	0.767	0.751	0.828	0.640	0.731	0.709	0.799
Real Estate	0.355	0.500	0.496	0.632	0.362	0.489	0.476	0.604	0.392	0.520	0.505	0.621	0.315	0.463	0.441	0.592
Materials	-0.040	0.121	0.126	0.278	-0.018	0.121	0.114	0.247	0.010	0.130	0.131	0.255	-0.043	0.107	0.096	0.248
Information Technology	0.223	0.423	0.399	0.568	0.216	0.372	0.375	0.539	0.263	0.410	0.408	0.566	0.170	0.340	0.338	0.503
Industrials	0.121	0.278	0.291	0.429	0.150	0.273	0.277	0.402	0.176	0.295	0.301	0.418	0.116	0.250	0.250	0.390
Health Care	0.423	0.568	0.534	0.693	0.415	0.532	0.521	0.644	0.447	0.564	0.549	0.663	0.368	0.516	0.493	0.633
Financials	0.632	0.749	0.717	0.829	0.636	0.723	0.709	0.801	0.661	0.751	0.734	0.821	0.617	0.714	0.693	0.792
Energy	0.429	0.568	0.563	0.693	0.441	0.558	0.549	0.668	0.470	0.584	0.577	0.695	0.421	0.545	0.529	0.658
Consumer Staples	0.259	0.398	0.388	0.536	0.250	0.368	0.371	0.494	0.277	0.394	0.399	0.522	0.198	0.343	0.339	0.486
Consumer Discretionary	0.200	0.355	0.367	0.500	0.225	0.361	0.359	0.488	0.237	0.366	0.373	0.500	0.197	0.340	0.334	0.471
Communication Services	0.200	0.355	0.338	0.500	0.174	0.321	0.317	0.469	0.196	0.335	0.342	0.495	0.133	0.282	0.280	0.437

References

Aït-Sahalia, Y., Fan, J. and Xiu, D. (2010), 'High-frequency covariance estimates with noisy and asynchronous financial data', *Journal of the American Statistical Association* **105**(492), 1504–1517.

Andersen, T. G., Bollerslev, T., Diebold, F. X. and Labys, P. (2001), 'The distribution of realized exchange rate volatility', *Journal of the American statistical association* **96**(453), 42–55.

Andersen, T. G., Bollerslev, T., Diebold, F. X. and Labys, P. (2003), 'Modeling and forecasting realized volatility', *Econometrica* **71**(2), 579–625.

Andersen, T. G., Thyrsgaard, M. and Todorov, V. (2019), 'Time-varying periodicity in intraday volatility', *Journal of the American Statistical Association* .

Andersen, T. G., Thyrsgaard, M. and Todorov, V. (2021), 'Recalcitrant betas: Intraday variation in the cross-sectional dispersion of systematic risk', *Quantitative Economics* **12**(2), 647–682.

Bandi, F. M. and Russell, J. R. (2005), 'Realized covariation, realized beta and microstructure noise', *Unpublished paper, Graduate School of Business, University of Chicago* **122**.

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

Bandi, F. M. and Russell, J. R. (2008), 'Microstructure noise, realized variance, and optimal sampling', *The Review of Economic Studies* **75**(2), 339–369.

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A. and Shephard, N. (2011), 'Multivariate realised kernels: consistent positive semi-definite estimators of the covariation of equity prices with noise and non-synchronous trading', *Journal of Econometrics* **162**(2), 149–169.

Barndorff-Nielsen, O. E. and Shephard, N. (2004*a*), 'Econometric analysis of realized covariation: High frequency based covariance, regression, and correlation in financial economics', *Econometrica* **72**(3), 885–925.

Barndorff-Nielsen, O. E. and Shephard, N. (2004*b*), 'Power and bipower variation with stochastic volatility and jumps', *Journal of financial econometrics* **2**(1), 1–37.

Barndorff-Nielsen, O. E. and Shephard, N. (2007), *Variation, Jumps, and High-Frequency Data in Financial Econometrics*, Vol. 3 of *Econometric Society Monographs*, Cambridge University Press, pp. 328–372.

Bartlett, M. S. (1946), 'On the theoretical specification and sampling properties of autocorrelated time-series', *Supplement to the Journal of the Royal Statistical Society* **8**(1), 27–41.

Bartlett, M. S. (1950), 'Periodogram analysis and continuous spectra', *Biometrika* **37**(1/2), 1–16.

Blomqvist, N. (1950), 'On a measure of dependence between two random variables', *The Annals of Mathematical Statistics* pp. 593–600.

Boudt, K., Cornelissen, J. and Croux, C. (2012*a*), 'The gaussian rank correlation estimator: robustness properties', *Statistics and Computing* **22**(2), 471–483.

Boudt, K., Cornelissen, J. and Croux, C. (2012*b*), 'Jump robust daily covariance estimation by disentangling variance and correlation components', *Computational Statistics & Data Analysis* **56**(11), 2993–3005. 1st issue of the Annals of Computational and Financial Econometrics Sixth Special Issue on Computational Econometrics.
URL: *https://www.sciencedirect.com/science/article/pii/S0167947311001691*

Boudt, K., Cornelissen, J. and Croux, C. (2012*c*), 'Jump robust daily covariance estimation by disentangling variance and correlation components', *Computational Statistics & Data Analysis* **56**(11), 2993–3005.

Boudt, K., Croux, C. and Laurent, S. (2011), 'Outlyingness weighted covariation', *Journal of Financial Econometrics* **9**(4), 657–684.

Cambanis, S., Huang, S. and Simons, G. (1981), 'On the theory of elliptically contoured distributions', *Journal of Multivariate Analysis* **11**(3), 368–385.

Chang, P., Pienaar, E. and Gebbie, T. (2021), 'The epps effect under alternative sampling schemes', *Physica A: Statistical Mechanics and its Applications* **583**, 126329.

Christensen, K., Kinnebrock, S. and Podolskij, M. (2010), 'Pre-averaging estimators of the ex-post covariance matrix in noisy diffusion models with non-synchronous data', *Journal of econometrics* **159**(1), 116–133.

Christensen, K., Podolskij, M. and Vetter, M. (2013), 'On covariation estimation for multivariate continuous itô semimartingales with noise in non-synchronous observation schemes', *Journal of Multivariate Analysis* **120**, 59–84.

Corsi, F., Peluso, S. and Audrino, F. (2015), 'Missing in asynchronicity: a kalman-em approach for multivariate realized covariance estimation', *Journal of Applied Econometrics* **30**(3), 377–397.

Croux, C. and Dehon, C. (2010), 'Influence functions of the spearman and kendall correlation measures', *Statistical methods & applications* **19**(4), 497–515.

Delattre, S. and Jacod, J. (1997), 'A central limit theorem for normalized functions of the increments of a diffusion process, in the presence of round-off errors', *Bernoulli* pp. 1–28.

Devlin, S. J., Gnanadesikan, R. and Kettenring, J. R. (1975), 'Robust estimation and outlier detection with correlation coefficients', *Biometrika* **62**(3), 531–545.

Epps, T. W. (1979), 'Comovements in stock prices in the very short run', *Journal of the American Statistical Association* **74**(366a), 291–298.

Griffin, J. E. and Oomen, R. C. (2011), 'Covariance measurement in the presence of non-synchronous trading and market microstructure noise', *Journal of Econometrics* **160**(1), 58–68.

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

Higham, N. J. (2002), 'Computing the nearest correlation matrix-a problem from finance', *IMA journal of Numerical Analysis* **22**(3), 329–343.

Jacod, J., Li, Y., Mykland, P. A., Podolskij, M. and Vetter, M. (2009), 'Microstructure noise in the continuous case: the pre-averaging approach', *Stochastic processes and their applications* **119**(7), 2249–2276.

Kendall, M. G. (1938), 'A new measure of rank correlation', *Biometrika* **30**(1/2), 81–93.

Künsch, H. R. (1989), 'The jackknife and the bootstrap for general stationary observations', *Annals of Statistics* **17**(3), 1217–1241.

Li, Y. and Mykland, P. A. (2015), 'Rounding errors and volatility estimation', *Journal of Financial Econometrics* **13**(2), 478–504.

Li, Y., Zhang, Z. and Li, Y. (2018), 'A unified approach to volatility estimation in the presence of both rounding and random market microstructure noise', *Journal of Econometrics* **203**(2), 187–222.

Liu, R. Y., Singh, K. et al. (1992), 'Moving blocks jackknife and bootstrap capture weak dependence', *Exploring the limits of bootstrap* **225**, 248.

Malliavin, P. and Mancino, M. E. (2002), 'Fourier series method for measurement of multivariate volatilities', *Finance and Stochastics* **6**(1), 49–61.

Mancini, C. and Gobbi, F. (2012), 'Identifying the brownian covariation from the co-jumps given discrete observations', *Econometric Theory* **28**(2), 249–273.

Martens, M. (2004), 'Estimating unbiased and precise realized covariances', *Available at SSRN 556118* .

Mastromatteo, I., Marsili, M. and Zoi, P. (2011), 'Financial correlations at ultra-high frequency: theoretical models and empirical estimation', *The European Physical Journal B* **80**(2), 243–253.

Moran, P. (1948), Rank correlation and permutation distributions, *in* 'Mathematical Proceedings of the Cambridge Philosophical Society', Vol. 44, Cambridge University Press, pp. 142–144.

Mosteller, F. (1946), 'On some useful "inefficient" statistics', *The Annals of Mathematical Statistics* **17**(4), 377–408.

Münnix, M. C., Schäfer, R. and Guhr, T. (2011), 'Statistical causes for the epps effect in microstructure noise', *International Journal of Theoretical and Applied Finance* **14**(08), 1231–1246.

Plackett, R. L. (1954), 'A reduction formula for normal multivariate integrals', *Biometrika* **41**(3/4), 351–360.

Podolskij, M. and Vetter, M. (2009), 'Estimation of volatility functionals in the simultaneous presence of microstructure noise and jumps', *Bernoulli* **15**(3), 634–658.

Politis, D. N. and White, H. (2004), 'Automatic block-length selection for the dependent bootstrap', *Econometric reviews* **23**(1), 53–70.

Precup, O. V. and Iori, G. (2007), 'Cross-correlation measures in the high-frequency domain', *European Journal of Finance* **13**(4), 319–331.

Qi, H. and Sun, D. (2006), 'A quadratically convergent newton method for computing the nearest correlation matrix', *SIAM journal on matrix analysis and applications* **28**(2), 360–385.

Raymaekers, J. and Rousseeuw, P. J. (2021), 'Fast robust correlation for high-dimensional data', *Technometrics* **63**(2), 184–198.
URL: *https://doi.org/10.1080/00401706.2019.1677270*

Renò, R. (2003), 'A closer look at the epps effect', *International Journal of theoretical and applied finance* **6**(01), 87–102.

Rosenbaum, M. (2009), 'Integrated volatility and round-off error', *Bernoulli* **15**(3), 687–720.

Rousseeuw, P. J. (1984), 'Least median of squares regression', *Journal of the American Statistical Association* **79**(388), 871–880.
URL: *https://www.tandfonline.com/doi/abs/10.1080/01621459.1984.10477105*

Shephard, N. and Xiu, D. (2017), 'Econometric analysis of multivariate realised qml: estimation of the covariation of equity prices under asynchronous trading', *Journal of Econometrics* **201**(1), 19–42.

Shevlyakov, G. and Smirnov, P. (2011), 'Robust estimation of the correlation coefficient: An attempt of survey', *Austrian Journal of Statistics* **40**(1&2), 147–156.

Spearman, C. (1904), "'General intelligence," objectively determined and measured.', *American Journal of Psychology* **15**, 201–292.

Tóth, B. and Kertész, J. (2007), Modeling the epps effect of cross correlations in asset prices, *in* 'Noise and Stochastics in Complex Systems and Finance', Vol. 6601, SPIE, pp. 89–97.

Tóth, B. and Kertész, J. (2009), 'The epps effect revisited', *Quantitative Finance* **9**(7), 793–802.

Vander Elst, H. and Veredas, D. (2016), 'Smoothing it out: Empirical and simulation results for disentangled realized covariances', *Journal of Financial Econometrics* **15**(1), 106–138.

Voev, V. and Lunde, A. (2007), 'Integrated covariance estimation using high-frequency data in the presence of noise', *Journal of Financial Econometrics* **5**(1), 68–104.

Xu, W., Hou, Y., Hung, Y. and Zou, Y. (2013), 'A comparative analysis of spearman's rho and kendall's tau in normal and contaminated normal models', *Signal Processing* **93**(1), 261–276.
URL: *https://www.sciencedirect.com/science/article/pii/S0165168412002721*

Zhang, L. (2006), 'Efficient estimation of stochastic volatility using noisy observations: A multi-scale approach', *Bernoulli* **12**(6), 1019–1043.

Zhang, L. (2011), 'Estimating covariation: Epps effect, microstructure noise', *Journal of Econometrics* **160**(1), 33–47.

Zhang, L., Mykland, P. A. and Aït-Sahalia, Y. (2005), 'A tale of two time scales: Determining integrated volatility with noisy high-frequency data', *Journal of the American Statistical Association* **100**(472), 1394–1411.