# Robust Estimation of Realized Correlation: New Insight about Intraday Fluctuations in Market Betas

**Peter Reinhard Hansen**[a] **and Yiyao Luo**[b]

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

[b] *University of Mississippi*[*]

*October 30, 2023*

**Abstract**

Time-varying volatility is an inherent feature of most economic time-series, which causes standard correlation estimators to be inconsistent. The quadrant correlation estimator is consistent but very inefficient. We propose a novel subsampled quadrant estimator that improves efficiency while preserving consistency and robustness. This estimator is particularly well-suited for high-frequency financial data and we apply it to a large panel of US stocks. Our empirical analysis sheds new light on intra-day fluctuations in market betas by decomposing them into time-varying correlations and relative volatility changes. Our results show that intraday variation in betas is primarily driven by intraday variation in correlations.

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

[*]Corresponding author: Yiyao Luo. Email: yluo3@olemiss.edu Mailing Address: University of Mississippi, Department of Economics, Oxford, MS, 38677, USA. We are grateful for helpful comments from Ron Gallant and seminar participants at UNC, Duke University, and the 2023 Annual Meeting of SoFiE in Seoul, Korea.

# 1 Introduction

The correlation is a measure of association between two variables that plays a central role in many empirical methods. The correlation is most commonly estimated by the sample correlation, which is known as Pearson's $r$. Other classical correlation estimators include the Quadrant estimator, the Kendall's tau, Spearman's rank correlation, and the Gaussian rank correlation estimator, see Kruskal (1958) for the relationships between these measures and an historical account of their developments. The choice of estimator involves a tradeoffs between robustness and efficiency. This tradeoff is influenced by many features of the underlying distribution, including heteroskedasticity that is particularly important for many economic applications.

In this paper, we propose a new robust correlation estimator that is well-suited for heteroskedastic time-series, such as high-frequency financial data. Time-varying volatility and market microstructure noise are innate features of high-frequency financial data, and both features undermine the reliability of standard correlation estimators. We compare the sensitivity of correlation estimators to departures from homoskedasticity and show that the Quadrant estimator is the only estimator that is robust to heteroskedasticity, among the classical estimators. The other estimators are inconsistent, except in very special cases. Unfortunately, the Quadrant estimator is rather inefficient. We recover much efficiency by combining the Quadrant estimator with subsampling and this makes it possible to improve efficiency while retaining consistency. We derive theoretical properties of the new estimator and study it using simulation designs that mimic empirical high-frequency financial data. We show that the realized correlation can be very biased as documented in our empirical analysis. We apply the new estimator to high-frequency data for 22 assets and an exchange-traded fund that tracks the S&P 500 index. The empirical results suggest that the new estimator is more accurate than other estimators, with the improvements likely resulting from better bias properties. We combine intraday correlation estimates with estimates of relative volatility to form an estimate of intraday market beta, as analyzed in Andersen et al. (2021). We find substantial variation in market betas within the trading day, with some stocks having increasing betas over the trading hours, while others tend to have decreasing betas. Our empirical results corroborate the finding in Andersen et al. (2021), even though we use different estimation methods and a different (narrower) estimation window. Our estimation approach enables us to decompose the time variation in betas into time variation in correlation and time-variation in relative volatility. Interestingly, we find that the variation in betas is mainly driven by time-variation in

correlations. Relative to the market, all assets in our analysis have increasing correlations and decreasing relative volatilities over the trading hour. The declines in relative volatilities are very similar across assets. The relative volatility during the last hour of active trading is typically between 50%-75% of relative volatility during the first hour of trading. There is far more variation across assets in terms of their correlations with the market. For many assets their market correlation is 2-5 times larger during the last hour than during the first hour. These assets have nearly linearly increasing market betas during the trading hours. Another set of assets, which are characterized by high market correlations, have their correlations increase by much less than 100% during the day. These assets have, on average, decreasing market betas during the trading hours. Thus, we document that intraday variation in both correlation and relative volatility contribute to the variation in market betas, but the variation across assets is primarily driven their time-variation in correlations with the market.

Time-varying volatility in high-frequency financial data is well documented, see e.g. Andersen and Bollerslev (1998*b*). Similarly, it is well documented that market microstructure noise can harm realized measures of volatility, see Zhou (1996, 1998), Zhang et al. (2005), Bandi and Russell (2006), and Hansen and Lunde (2006). Market microstructure noise is defined as the difference between the observed prices and true prices. The latter are characterized by having certain martingale properties, whereas the former typically entails some degree of predictability. Market microstructure noise arises from many intricate aspects of high-frequency data. For instance, noise can arise as artifacts of imputation methods and recording and rounding errors. These issues are all important for correlation estimation, see e.g. Renò (2003), Precup and Iori (2007), and Münnix et al. (2011), and Tóth and Kertész (2007, 2009). The lack of synchronicity in observation times induces a type of noise that is particularly important for covariance and correlation estimation. This will often manifest as the Epps effect, where the sample correlation decreases as the sampling frequency increases, see Epps (1979). Hayashi and Yoshida (2005) proposed an estimator that adjusts for asynchronicity and Voev and Lunde (2007) and Griffin and Oomen (2011) proposed related estimators that are robust to additional forms of noise. Jumps in prices and adversely effect empirical measures, including realized variances, covariances, and correlations. However, these effects can be alleviated by truncation methods, see Mancini (2009) and Raymaekers and Rousseeuw (2021).

A standard remedy for market microstructure noise in high-frequency data is sparse sampling. Andersen and Bollerslev (1998*a*) estimated realized variances using 5-minute intraday returns and this sampling frequency appears to offer a reasonably good compromise between bias and variance in many

applications, see e.g. Hansen and Lunde (2006), Bandi and Russell (2008), and Liu et al. (2015). Realized measures that utilize more information include the subsampled realized variance by Zhang et al. (2005), the realized kernel estimator by Barndorff-Nielsen et al. (2008), and the pre-averaging estimators by Jacod et al. (2009). These three approaches, subsampling, realized kernels, and pre-averaging, lead to the same class of estimators, aside from minor differences caused by end-effects, see Barndorff-Nielsen et al. (2011*b*).

Realized correlations are often computed from multivariate estimators, such as those proposed in Malliavin and Mancino (2002), Barndorff-Nielsen and Shephard (2004*a*), Christensen et al. (2010), Ait-Sahalia et al. (2010), Barndorff-Nielsen et al. (2011*a*), and Christensen et al. (2013), among others. If volatility varies over the period for which estimators are computed, then the resulting estimator will be inconsistent, aside from special cases, as we detail in Section 3.

### 1.1 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 Population and Empirical Measures of Correlation

We begin by reviewing classical correlation measures, starting with the Pearson correlation.

### 2.1 Population Measures

For two random variables, $X$ and $Y$, with finite variances, the correlation coefficient is defined by

$$\rho = \frac{\sigma_{XY}}{\sqrt{\sigma_X^2 \sigma_Y^2}}, \qquad \text{where} \quad \sigma_{XY} = \text{cov}(X,Y) = \mathbb{E}[(X-\mu_X)(Y-\mu_Y)],$$

$\mu_X = \mathbb{E}(X)$, $\mu_Y = \mathbb{E}(Y)$, $\sigma_X^2 = \text{var}(X)$, and $\sigma_Y^2 = \text{var}(Y)$.

Nowadays, the "correlation" is commonly understood to mean $\rho = \sigma_{XY}/\sqrt{\sigma_X^2 \sigma_Y^2}$, but $\rho$ is just

one of several classical population measures of the correlation. Another measure is defined from sign-concordances,

$$\tau = \mathbb{E}[\text{sgn}\{(X - \xi_X)(Y - \xi_Y)\}],$$

where $\text{sgn}(x)$ denotes the sign of $x$, and $\xi_X$ and $\xi_Y$ are the medians of $X$ and $Y$, respectively. The parameter $\tau$ is given from the quadrant probabilities of the recentered variables, $\tilde{X} = X - \xi_X$ and $\tilde{Y} = Y - \xi_Y$, since $\tau = \Pr[Z > 0] - \Pr[Z < 0]$, where $Z = \tilde{X}\tilde{Y}$, and $\tau$ is the population quantity that is estimated by the quadrant estimators we use below. For spherical and continuously distributed variables, we have $q \equiv \Pr[Z > 0] = 1 - \Pr[Z < 0] = (\tau + 1)/2$, such that $\tau = 2q - 1$. A closely related population measure is Kendall's tau, which is given by

$$\tau_K = \mathbb{E}[\text{sgn}\{(X_1 - X_2)(Y_1 - Y_2)\}],$$

where $(X_1, Y_1)$ and $(X_2, Y_2)$ are independent and distributed as $(X, Y)$. For a continuous bivariate distribution with cdf, $F$, it can be shown that $\tau_K = \mathbb{E}[4\{F(X, Y) - \frac{1}{4}\}]$ whereas $\tau = 4[F(\xi_x, \xi_y) - \frac{1}{4}]$. The two quantities, $\tau$ and $\tau_K$, are identical for elliptical distributions.

Other classical correlation measures include the Gaussian rank correlation and Spearman's rank correlation, where the latter estimates $\eta = 12\{\mathbb{E}[F(X)G(Y)] - \frac{1}{4}\}$, where $F$ and $G$ are the cumulative distribution functions for $X$ and $Y$, respectively. We do not include these estimators in our comparison, because they are not competitive for various reasons discussed later in the paper.

The population measures, $\rho$, $\tau$, $\tau_K$, and $\eta$ are closely related and all have values ranging between $-1$ and 1. The exact relation between these quantities depends on the bivariate distribution of $(X, Y)$. For elliptical distributions we have $\tau(\rho) = \frac{2}{\pi} \arcsin \rho$, such that the inverse mapping is:

$$\rho = \sin(\tfrac{\pi}{2}\tau). \tag{1}$$

This link function was derived in Greiner (1909, p.236), albeit the identity is implicit from results in Sheppard (1899), who first related quadrant probabilities to the correlation. Greiner derived the result under the assumption that $(X, Y)$ are normally distributed, but (1) is valid for a broader class of distributions that includes all symmetric elliptical distributions for which the correlation is well defined, such as the multivariate $t$-distribution with degrees of freedom greater than two, see Proposition A.1. The link function is also unaffected by skewness as defined by the moments and cumulants of odd order,

see Kendall (1949). The link function in (1) makes it possible to translate an estimate of $\tau$ into an estimate of $\rho$. In general, the link function between $\tau$ and $\rho$ depends on actual bivariate distribution and we have shown three examples of the link function in Figure 1.

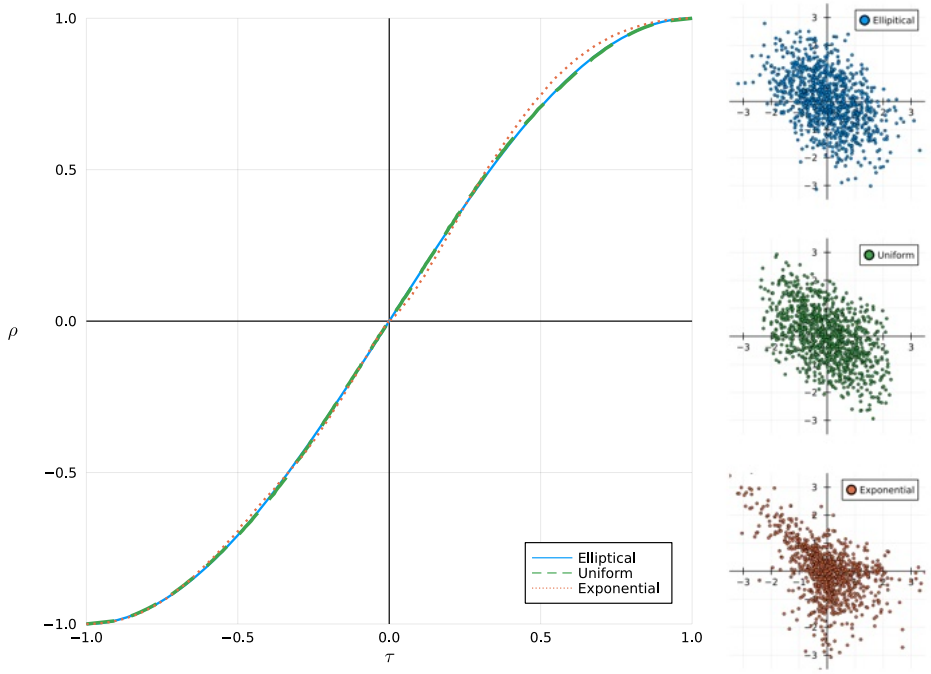



Figure 1: The mapping from $\tau = \mathbb{E}[\text{sgn}(X - \xi_x)(Y - \xi_y)]$ to $\rho = \text{corr}(X, Y)$ for three bivariate distributions.

The blue line represents Greiner's link function, (1), while the green dashed and the red dotted lines represent link functions for two non-elliptical distributions. The green dashed link function, labelled "Uniform" is based on the following bivariate distribution

$$X = U_1 - U_2 \quad Y = \rho X + \sqrt{1 - \rho^2}(U_3 - U_4), \qquad \rho \in [-1, 1]$$

where $U_1$,..., $U_4$ are independent and uniformly distributed on $[0, 1]$, and the red dotted link function, labelled "Exponential" is based on

$$X = aZ_0 - (1 - |a|)Z_1 \quad Y = |a|Z_0 + (1 - |a|)Z_2, \qquad a \in [-1, 1],$$

where $Z_0, Z_1, Z_2$ are independent and standard exponentially distributed. The correlation for this distribution is $\rho(a) = a|a|/[a^2 + (1 - |a|)^2]$. Examples of these distributions are shown in the right panels using scatterplots with 1,000 observations. The upper right panel is the bivariate normal distribution

with correlation $-0.5$. The middle right panel is that based on four uniformly distributed random variables with $\rho = -0.5$, and the lower right panel is based on the exponential random variables with $a = -0.5$, which happens to translate to the same correlation, $\rho(-0.5) = -0.5$. Thus Greiner's link function is a good approximation to the two other link functions in Figure 1, and may offers a good approximation to a broader class of distributions than the elliptical distributions. However, it is also possible to construct a bivariate distribution whose link function differs to a greater degree from that in (1).[1]

### 2.2 Classical Correlation Estimators: Pearson, Quadrant, and Kendall

Next, we introduce classical correlations estimators. To simplify the exposition we use $(x_i, y_i)$, $i = 1, \ldots, n$, to denote recentered variables, such that their sample means (or sample medians) are zero.[2]

The Pearson correlation estimator is the well-know sample correlation, which takes the form

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

This estimator is asymptotically efficient if the data are iid Gaussian. A drawback of the Pearson estimator is that it is sensitive to outliers, as we discuss below. More robust estimators of $\rho$ can be constructed from estimators of $\tau$, such as the quadrant estimator

$$\hat{\tau}_Q = \frac{1}{n} \sum_{i=1}^n \text{sgn}(x_i y_i),$$

and Kendall's tau coefficient

$$\hat{\tau}_K = \frac{2}{n(n-1)} \sum_{i<j} \text{sgn}([x_i - x_j][y_i - y_j]).$$

Quadrant-based estimation of the correlation was introduced in Sheppard (1899), with the relation between $\rho$ and $\tau$ spelled out in Greiner (1909). The asymptotic properties of the quadrant estimator were derived in Blomqvist (1950). Esscher (1924) introduced the $\hat{\tau}_K$ estimator and cited Greiner (1909) for the link function. This estimator was rediscovered in Kendall (1938) and is commonly known as

[1] For instance, pathological examples can be created by assigning small probabilities to extreme events. carefully shifting probability mass near zero to shift the binary distribution over signs, which will have negligible without having much impact on $\rho$.

[2] The sample mean is subtracted before applying the Pearson estimator and the sample median is subtracted if the Quadrant or Kendall estimators are used.

Kendall's tau coefficient and Kendall rank correlation coefficient. Note that $\hat{\tau}_K$ is the quadrant estimator applied to $\{(X_i - X_j, Y_i - Y_j)\}_{i<j}$, and it is easy to verify that $\rho = \text{corr}(X_1, Y_1) = \text{corr}(X_1 - X_2, Y_1 - Y_2)$, if $(X_1, Y_1)$ and $(X_2, Y_2)$ are independent and identically distribution.

In this paper, we employ Greiner's link function to map the estimators of $\tau$ to estimators of $\rho$. The estimators of $\rho$ are therefore defined by

$$Q = \sin(\tfrac{\pi}{2}\hat{\tau}_Q) \qquad \text{and} \qquad K = \sin(\tfrac{\pi}{2}\hat{\tau}_K),$$

respectively. A convenient feature of these two estimators, is that they bypass the need for estimating the variances of $X$ and $Y$. In fact, $Q$ and $K$ do not rely on $X$ and $Y$ having finite moments. For non-elliptical distributions, (1) may not be the appropriate link function, and this type of misspecification can therefore induce a bias in these estimators. Fortunately, Greiner's link function does appear to offer a good approximation beyond the class of elliptical distributions, as illustrated in Figure 1. In our empirical application we use sparsely sampled financial returns, which is an application where a Gaussian assumption has some theoretical justification.

### 2.3 A New Correlation Estimator

Our new estimator is motivated by the empirical situation one encounters with high-frequency financial data, where market microstructure noise, jumps, and time-varying volatility pose challenges to the validity of correlation estimators. While Pearson is the ideal estimator when the variables are distributed as a bivariate Gaussian distribution, it is inconsistent under more realistic and commonly accepted assumptions for intraday returns. The $K$ estimator is more robust, but also inconsistent under time-varying volatility, while $Q$ is is very inefficient. This motivates the estimator introduced below.

#### 2.3.1 Notation with High-Frequency Data

Let $X(t)$ and $Y(t)$ denote the the observed logarithmically transformed price processes over some period, such as a trading day. We denote the intraday returns over a time-interval with length $\delta$ by

$$\Delta_\delta X_t = X(t) - X(t - \delta),$$

and similarly for $\Delta_\delta Y_t$. In the context of high frequency data it is common to sample sparsely to mitigate the effects of market microstructure noise, and a popular choice is to set $\delta$ equal to five minutes. If we normalize the interval of time to be $[0,1]$ and set $\delta = \frac{1}{n}$, then the correlation estimators given above, may be applied to $(x_i, y_i) = (\Delta_{\frac{1}{n}} X_{\frac{i}{n}}, \Delta_{\frac{1}{n}} Y_{\frac{i}{n}})$ for $i = 1, \ldots, n$.

Let $N$ denote the number of intraday returns at the highest possible sampling frequency and suppose, for simplicity, that $N$ is divisible by $n$, such that $S = N/n \in \mathbb{N}$. Then we can create $S$ distinct grids by shifting the initial observation time to be $t_s = s/(Sn)$ for $s = 0, \ldots, S-1$. Each grid will have sparely and non-overlapping returns, and combined we have $N - S + 1$ pairs of sparsely sampled returns, $(\Delta_\delta X_{\frac{j}{N}}, \Delta_\delta Y_{\frac{j}{N}})$, for $j = S, \ldots, N$.[3]

### 2.3.2 Subsampled Quadrant Estimator

We are now ready to introduce the subsampled variant of the Quadrant estimator, defined by

$$Q_S = \sin(\tfrac{\pi}{2}\hat{\tau}_S) \qquad \text{with} \qquad \hat{\tau}_S = \frac{1}{N-S+1} \sum_{j=S}^{N} \operatorname{sgn}(\Delta_{\frac{S}{N}} X_{\frac{j}{N}} \Delta_{\frac{S}{N}} Y_{\frac{j}{N}}).$$

The estimator does not require $N$ to be divisible by $S$, but if $N$ is divisible by $S$, then $\hat{\tau}_S$ can be expressed as a simple average of $S$ $\tau$-estimators based on different grids. This construction is similar to many robust estimators of the long-run variance. Politis et al. (1999) noted that the subsampled sample variance is identical to the moving-blocks estimator and the jackknife variance estimator, and it is almost identical to the Bartlett estimator, Bartlett (1946, 1950), which is often referred to as the Newey-West estimator in the econometrics literature.[4] In the context of volatility estimation with high-frequency data, the subsampling idea was first used in Zhou (1996). The theoretical foundation for subsampled realized variances was established in Zhang et al. (2005) and Zhang (2006), and the close connection between subsampled estimators and kernel estimators is detailed in Barndorff-Nielsen et al. (2011*b*).

The subsampled quadrant correlation estimator has several appealing properties. First, it inherits the robustness of the quadrant estimator while being more precise than $Q$. The robustness is character-

[3]For instance, a 6.5 hour long trading day has $n = 78$ intraday returns when partitioned into 5-minute intervals. By shifting the starting time we obtain partitions with distinct 5-minute returns, each having just 77 returns. By shifting the starting time in one-minute increments we obtain $S = 5$ different partitions and a total of 386 5-minute returns.

[4]Politis et al. (1999, p.98): "[...] the variance estimator $\hat{\sigma}^2_{\text{SUB}}$ is actually asymptotically equivalent to the Bartlett kernel estimator[...]" and Politis et al. (1999, p.98): "In addition, $\hat{\sigma}^2_{\text{SUB}}$ is identical to the moving blocks bootstrap and/or jackknife variance estimator of the variance of the sample mean proposed by Künsch (1989) and Liu and Singh (1992) [...]".

ized by the influence function, which is discussed below. Second, $Q_S$ is consistent under to time-varying volatilities. This is important because time-varying volatility is common in economic time series, especially in high-frequency financial data. Third, another computationally attractive feature of the new $Q_S$ estimator, is that it relies on binary variables. This makes it easier to scale this estimator to large data sets.

### 2.3.3 Implementation at Ultra High Frequencies

One challenge with ultra-high-frequency data is that price increments can be zero over short time intervals, resulting in $\Delta_\delta X \Delta_\delta Y = 0$. This issue may be caused by stale prices and rounding to a grid defined by the minimum tick size. This issue abates quickly with sparse sampling, and zeros are infrequent in our empirical analysis once we sample a frequencies below one minute. Most of our emprical results are based on $\delta = 3$-minutes. Still, we will explore the properties of the estimators at at higher sampling frequencies to gain insight about them and market microstructure noise. For this reason we need to account for zero returns, and we do so by redefining the estimator,

$$\hat{\tau}_S = \frac{1}{N_1 - S + 1} \sum_{j=S}^{N} \operatorname{sgn}(\Delta_\delta X_{\frac{j}{N}} \Delta_\delta Y_{\frac{j}{N}}), \qquad \text{with} \quad N_1 = \sum_j 1_{\{\Delta_\delta X_{\frac{j}{N}} \Delta_\delta Y_{\frac{j}{N}} \neq 0\}},$$

such that we only count non-zero product-pairs.

## 3 Properties of the Estimators

In this section, we establish several properties of the estimators, and we highlight some of the key advantages that are unique to $Q_S$. We first consider the simple case with iid and normally distributed variables. This is the situation that arises when the price process are given from Brownian motions with constant volatility and the observed prices are measured without error. We then proceed with more realistic models with time-varying volatility and discuss robustness by means of the influence function of the estimators. The impact of general types of market microstructure noise will be analyzed in Section 4.

### 3.1 Limit Distributions under Ideal Circumstances

We begin with the simplest possible situation, where logarithmic price processes follow Brownian motions with constant volatilities and constant correlation.

**Assumption 1.** *Suppose that* $(X, Y)$ *is given by a bivariate Brownian motion, such that* $(X_t, Y_t) \sim N_2(0, t\Sigma)$.

In this situation, intraday returns, $(\Delta_\delta X_{i\delta}, \Delta_\delta Y_{i\delta})$, $i = 1, \ldots, n$ are iid and normally distributed. This is the ideal situation for the Pearson estimator, because $P$ is the maximum likelihood estimator of $\rho$ for the sample with the $n$ pairs of observations. We should therefore expect $P$ to compare favorable to $Q$ and $K$. It is less obvious how $P$ will compare with $Q_S$, because the latter utilizes the shifted grids of sparsely sampled returns, $(\Delta_\delta X_{j\delta/S}, \Delta_\delta Y_{j\delta/S}), j = S, \ldots, N$, and is thus computed from a larger data set. The asymptotic distribution of the new estimator is given next.

**Theorem 1.** *Suppose that Assumption 1 holds and let* $S \in \mathbb{N}$ *be fixed. Then the subsampled quadrant correlation estimator is asymptotically normally distributed*

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

*as* $n \to \infty$ *, where*

$$V_S(\rho) = (1-\rho^2)\frac{1}{S}\sum_{s=-S}^{S}\left[\text{asin}^2(w_s) - \text{asin}^2(w_s\rho)\right], \qquad w_s = \tfrac{S-|s|}{S},$$

*which is decreasing in* $S$ *and bounded from below by*

$$\lim_{S\to\infty} V_S(\rho) = (1-\rho^2)2\left[\text{asin}^2(1) - 2\tfrac{\sqrt{1-\rho^2}}{\rho}\text{asin}(\rho) - \text{asin}^2(\rho)\right].$$

The corresponding asymptotic distributions for the estimators, $Q$, $K$, and $P$, are well known, see e.g. Croux and Dehon (2010). For the sake of comparison, these are included below.

**Proposition 1.** *Suppose that Assumption 1 holds, then as* $n \to \infty$ *we have*

$$\begin{aligned}
\sqrt{n}(P-\rho) &\xrightarrow{d} N(0, V_P), \qquad \textit{with} \quad V_P = (1-\rho^2)^2,\\
\sqrt{n}(K-\rho) &\xrightarrow{d} N(0, V_K), \qquad \textit{with} \quad V_K = (1-\rho^2)[(\tfrac{\pi}{3})^2 - 4\text{asin}^2(\tfrac{\rho}{2})],\\
\sqrt{n}(Q-\rho) &\xrightarrow{d} N(0, V_Q), \qquad \textit{with} \quad V_Q = (1-\rho^2)[(\tfrac{\pi}{2})^2 - \text{asin}^2(\rho)].
\end{aligned}$$

We can compare the asymptotic variance of $Q_S$ to those of the other estimators. The asymptotic variances depend on $\rho$ and those for $P$, $K$, and $Q_S$ are shown in Figure 2. With $S = 1$ we obviously

have $V_{Q_S} = V_Q$.[5] For a sufficiently large $S$, the subsampled Quadrant estimator is more accurate than the Kendall estimator. For small values of $\rho$, $Q_S$ is more accurate than $K$ when $S \geq 5$, whereas a larger value of $S$ in required for larger values of $\rho$. The $Q_S$ estimator is similar to $P$ for large values of $S$, with $Q_S$ having the edge for small values of $\rho$, whereas $P$ has the edge for large values of $\rho$.

Realized measures are commonly computed from sparsely sampled returns, such as 5-minute returns, to minimize the impact of market microstructure noise. There will typically be a large number of observations within each 5-minute interval, and this makes it possible to use a relatively large value for $S$.

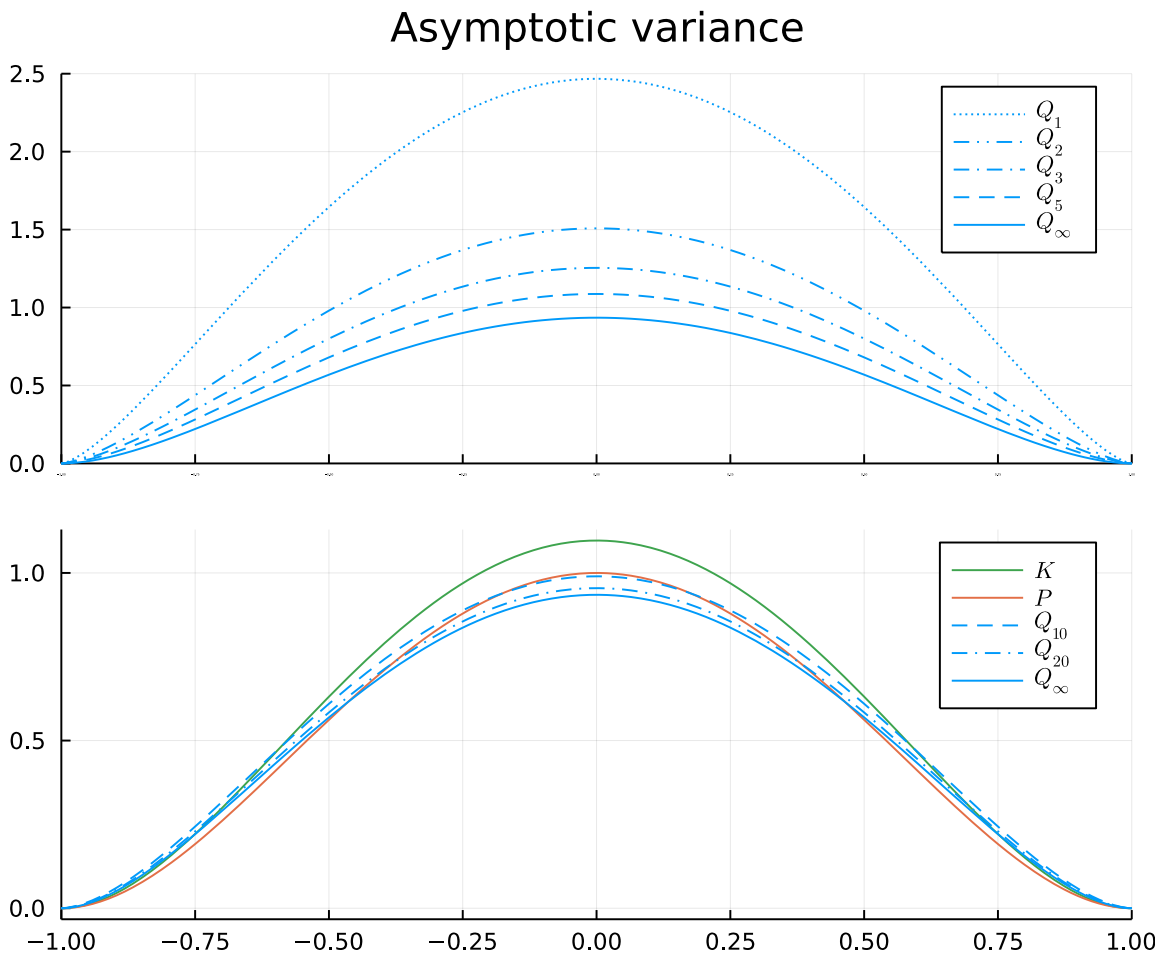



Figure 2: Asymptotic variances as a function of the true correlation, $\rho$, for Quadrant and subsampled Quadrant estimators in the upper panel and subsampled Quadrant, Pearson, and Kendall estimators in the lower panel.

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

### 3.2 Properties with Time-Varying Volatility

Time-varying volatility is an intrinsic feature of financial time-series. For instance, volatility if found to vary substantially in high-frequency financial data, even within a trading day. Next, we relax the assumption that volatility is constant and evaluate the effect this has on the correlation estimators. The asymptotic properties of correlation estimators stated above need not apply in this context, because they were derived under constant volatility.

We can illustrate the issues that arise from time-varying volatility with a simple bivariate Brownian semimartingale.

**Assumption 2.** *Suppose that*

$$\begin{pmatrix} X_t \\ Y_t \end{pmatrix} = \begin{pmatrix} X_0 \\ Y_0 \end{pmatrix} + \int_0^t \sigma(u)\mathrm{d}W(u), \tag{2}$$

*where $W(u)$ is a bivariate Wiener process with $\mathrm{cor}(\mathrm{d}W_x, \mathrm{d}W_y) = \rho$ and*

$$\sigma(u) = \begin{pmatrix} \sigma_x(u) & 0 \\ 0 & \sigma_y(u) \end{pmatrix},$$

*is a squared integrable CADLAG process.*

The assumption can be generalized in many ways, such as having a random drift term,[6] but the simple setup presented here suffices to show that traditional correlation estimators are biased in the presence of time-varying volatility, and establish that quadrant-based estimators are robust to time-varying volatility. We will, initially, take the correlation coefficient to be constant over time. The case with time varying correlation is discussed below in Section 3.3.

**Theorem 2.** *If Assumption 2 holds, then $Q_S \xrightarrow{p} \rho$ as $\delta \to 0$, whereas the probability limits for $P$ and $K$ are given by*

$$\begin{aligned} P &\xrightarrow{p} \lambda\rho, \qquad \text{where} \quad \lambda = \frac{\int_0^1 \sigma_x(u)\sigma_y(u)\mathrm{d}u}{\sqrt{\int_0^1 \sigma_x^2(u)\mathrm{d}u \int_0^1 \sigma_y^2(u)\mathrm{d}u}}, \\ K &\xrightarrow{p} \sin\left(\int_0^1 \int_0^1 \arcsin\left(h(u,v)\rho\right)\mathrm{d}u\mathrm{d}v\right), \end{aligned}$$

[6] Formally, we can let the logarithmic price process be defined on the filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F})_{t\in[0,1]}, \mathbb{P})$ with a locally bounded predictable drift function, $a(u)$, where $a$, $\sigma$, and $W$ are adapted to a $\mathcal{F}_t$.

*where* $h(u,v) = \frac{\sigma_x(u)\sigma_y(u)+\sigma_x(v)\sigma_y(v)}{\sqrt{\sigma_x^2(u)+\sigma_x^2(v)}\sqrt{\sigma_y^2(u)+\sigma_y^2(v)}}$.

The important result from Theorem 2 is that $Q_S$ emerges as the only consistent estimator when volatility is time-varying. Both $P$ and $K$ are generally inconsistent, except in the following special case where the two volatility processes are perfectly collinear.

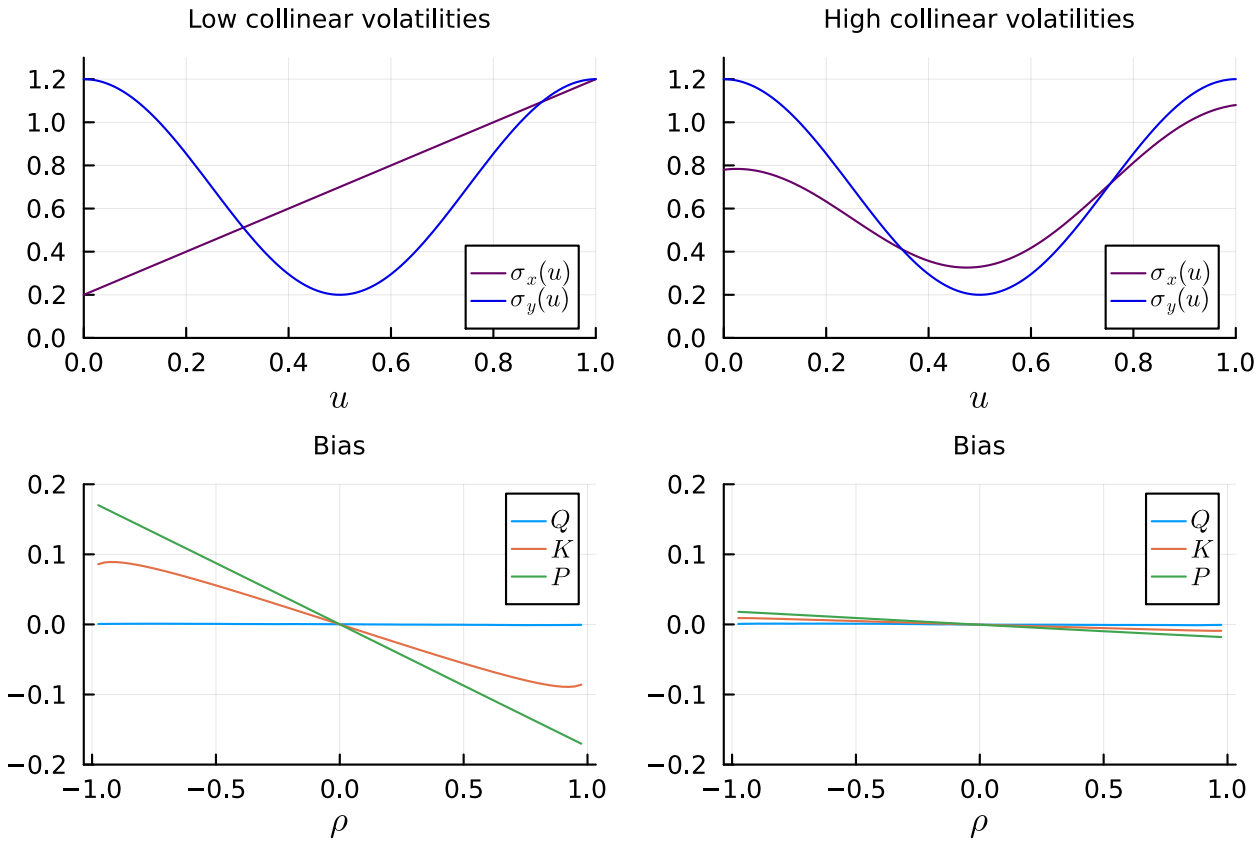



Figure 3: Bias in correlation estimators, $P$, $K$, and $Q_S$, with two different degrees of collinearity between the volatilities.

**Corollary 1.** *Suppose that Assumption 2 holds and that* $\sigma_y(u) = c\sigma_x(u)$ *for some* $c > 0$. *Then* $P \xrightarrow{p} \rho$ *and* $K \xrightarrow{p} \rho$.

The results for $P$ and $K$ in this special case are easy to verify, because perfectly collinearity implies $\lambda = 1$ and that

$$h(u,v) = \frac{c\sigma_x^2(u) + c\sigma_x^2(v)}{\sqrt{\sigma_x^2(u) + \sigma_x^2(v)}\sqrt{c^2(\sigma_x^2(u) + \sigma_x^2(v))}} = 1,$$

for all $u, v$.

We illustrate the inconsistency with a simple example. Consider the functions, $g(u) = \frac{1}{5} + \frac{4}{5}u$

and $h(u) = \frac{1}{2}(\frac{6}{5} + \cos(2\pi u))$, which we will use to construct volatility paths with varying degrees of collinearity. The upper left panel in Figure 3 represents a case with low collinearity, where $\sigma_x(u) = g(u)$ and $\sigma_y(u) = h(u)$, and the upper right panel corresponds to a case with high collinearity, where $\sigma_x(u) = \frac{3}{10}g(u) + \frac{6}{10}h(u)$ and $\sigma_y(u) = h(u)$. The lower panels show the resulting bias of the correlation coefficients, $P$, $K$, and $Q$, as a function of the true correlation coefficient, $\rho$. With low collinearity, the sample correlation, $P$, has a large bias unless $\rho$ is near zero, and the bias in $K$ is about half that in $P$. These estimators, $P$ and $K$, are also biased in the example with high collinearity, but the bias is substantially smaller. The bias of these estimators are also pronounced in standard simulation design with the Heston model, as we document in Section 4.

An important implication of the results in this subsection is that conventional estimates of correlations between assets are systematically influenced by the degree of collinearity in their volatilities.

### 3.3 Time varying correlations

The correlation may be time varying, as is the case for volatility. To accommodate this situation we could modify Assumption 2 and let $\rho(u)$ be a CADLAG process. In this situation, the integrated correlation, $\rho_\bullet = \int_0^1 \rho(u)\mathrm{d}u$, is a natural object of interest. Unfortunately, none of the correlation estimators are consistent for $\rho_\bullet$. For instance, $Q_S$ will estimate $\tilde{\rho}_\bullet = \sin[\int_0^1 \mathrm{asin}\{\rho(u)\}\mathrm{d}u]$, and since $\mathrm{asin}(t)$ is strictly convex for $t > 0$ and concave for $t < 0$, it follows that $\tilde{\rho}_\bullet \geq \rho_\bullet$ if the $\rho(u) \geq 0$ and $\tilde{\rho}_\bullet \leq \rho_\bullet$ if the $\rho(u) \leq 0$.

One way to partially account for time-varying correlations is to apply the correlation estimators over relatively short intervals of time and aggregate these local estimates to an estimate of $\int_0^1 \rho(u)\mathrm{d}u$. This approach was used in jump-robust estimation of the integrated covariance in Boudt et al. (2012*b*). In our empirical analysis we will also use local estimates of $\rho$ to assess time-variation in $\rho(u)$.

### 3.4 Estimating Integrated Covariance

Interestingly, it is not advisable to combine the robust correlation estimator with volatility estimators for the purpose of estimating the integrated covariance, $\mathrm{IC} = \int \sigma_{xy}(u)\mathrm{d}u$, which simplifies to $\rho \int \sigma_x(u)\sigma_y(u)\mathrm{d}u$ when $\rho(u) = \rho$ for all $u$. Now, if we multiply $Q_S$ by consistent estimates of $\sqrt{\int \sigma_x^2(u)\mathrm{d}u}$ and $\sqrt{\int \sigma_y^2(u)\mathrm{d}u}$, we will be estimating $\frac{1}{\lambda}\mathrm{IC}$, instead of IC. Using $K$ is not advisable either, because it leads to another incorrect limit. For this problem, localized estimators of spot volatility and spot correlation can be use, as proposed in Boudt et al. (2012*b*). In order to be robust to jumps, they combine

the MedRV estimator by Andersen et al. (2012) and the Gaussian rank correlation. This is further explored in Vander Elst and Veredas (2016) who employ additional robust correlation estimators, as a component to estimate IC. They also combine non-localized estimates of $\sqrt{\int \sigma_x^2(u)\mathrm{d}u}$ and $\sqrt{\int \sigma_y^2(u)\mathrm{d}u}$ with a range of correlation estimators. Some of these combinations will be inconsistent for the reason stated earlier. This may explain that Vander Elst and Veredas (2016) find the bivariate realized kernel estimator by Barndorff-Nielsen et al. (2011*a*) to be the most accurate estimator of IC in the absence of jumps.

### 3.5 Influence Function

The *influence function* can be used to measure an estimator's sensitivity to data contamination. It measures the sensitivity of a statistical functional, $R$, to data contamination in a baseline distribution, $F$, and is defined by

$$\mathrm{IF}((x_0, y_0), R, F) = \lim_{\eta \searrow 0} \frac{R((1-\eta)F + \eta \boldsymbol{\Delta}_{(x_0,y_0)}) - R(F)}{\eta},$$

where $\boldsymbol{\Delta}_{(x_0,y_0)}$ is the Dirac measure at $(x_0, y_0)$. We have $R_P(F) = R_Q(F) = R_K(F) = \rho$ for $F = \Phi_\rho$, which denotes the standard bivariate normal distribution with correlation equal to $\rho$. From Devlin et al. (1975) and Croux and Dehon (2010) we have their influence functions.

**Proposition 2.** *The influence functions of correlation estimators at $\Phi_\rho$ are given by*

$$\begin{aligned}
\mathrm{IF}((x_0, y_0), R_Q, \Phi_\rho) &= \tfrac{\pi}{2}\sqrt{1-\rho^2}(\mathrm{sgn}(x_0 y_0) - \tau) \\
\mathrm{IF}((x_0, y_0), R_K, \Phi_\rho) &= 2\pi\sqrt{1-\rho^2}(2\Phi(x_0, y_0) + 1 - \Phi(x_0) - \Phi(y_0) - \tfrac{1}{2}(\tau + 1)) \\
\mathrm{IF}((x_0, y_0), R_P, \Phi_\rho) &= x_0 y_0 - (\frac{x_0^2 + y_0^2}{2})\rho
\end{aligned}$$

*where $\Phi(\bullet, \bullet)$ denotes the joint CDF for $\Phi_\rho$ and and $\Phi(\bullet)$ denote the marginal CDF for a standard normal distribution.*

The important message from the influence functions is that $Q$ and $K$ have bounded influence functions whereas $P$ has an unbounded influence function. This difference motivate their labeling as robust and non-robust estimators, respectively. The unbounded influence function of $P$ makes it sensitive to outliers. It is intuitive that $Q$ and $K$ are less sensitive to outliers, since they are computed from signed variable alone. This limits the harm an outlier can cause to merely flipping the sign.

The analogous results for $Q_S$ are qualitative very similarly, and are presented in the Supplementary Material. One way to alleviate the sensitivity that $P$ has to outliers is to use truncation estimators, which is commonly used for estimating realized variances, see e.g. Mancini (2009).

The influence function for the Spearman estimator is also bounded but can be shown to have a larger bound than $Q$ and $K$, whereas the Gaussian rank estimator has an unbounded influence function, see Rousseeuw (1984), Boudt et al. (2012*a*), and Raymaekers and Rousseeuw (2021) for details and additional results on influence functions.

# 4 Simulation Study

We compare the estimators in simulation studies that are designed to emulate the situation we encounter in our empirical analyses with high-frequency data. We generate the two logarithmic price processes, $X_t^*$ and $Y_t^*$, using the Heston model:

$$\begin{aligned} \mathrm{d}X_t^* &= \mu_j \mathrm{d}t + \sigma_{xt} \mathrm{d}W_{xt}, \\ \mathrm{d}\sigma_{x,t}^2 &= \kappa_x(\bar{\sigma}_x^2 - \sigma_{x,t}^2)\mathrm{d}t + s_x \sigma_{x,t} dB_{x,t}, \end{aligned} \tag{3}$$

where $W_{x,t}$ and $B_{x,t}$ are standard Brownian motions with $\mathrm{cov}(\mathrm{d}W_{x,t}, \mathrm{d}B_{xt}) = \varrho_x dt$, and $Y_t^*$ is generated similarly with $\mathrm{cov}(\mathrm{d}W_{y,t}, \mathrm{d}W_{y,t}) = \rho \mathrm{d}t$. The model is calibrated using the simulation design in Table 1, which was previously used in Ait-Sahalia et al. (2010). The initial values for volatility $\sigma_{x,0}^2$ and $\sigma_{y,0}^2$ are drawn from Gamma distributions, $\Gamma(2\kappa_x\bar{\sigma}_x^2/s_x^2, s_x^2/2\kappa_x)$ and $\Gamma(2\kappa_y\bar{\sigma}_y^2/s_y^2, s_y^2/2\kappa_y)$, and the price processes are initialized with $X_0^* = \log(100)$ and $Y_0^* = \log(40)$. The simulated model is a discretized version with $23,400$ increments, which translates to 1 second observations over a 6.5 hours period – the length of a typical trading day.

Table 1: Parameters calibration for the Heston model

| | $\mu$ | $\bar{\sigma}^2$ | $\kappa$ | $s$ | $\varrho$ |
|---|---|---|---|---|---|
| $X$ | 0.05 | 0.16 | 3 | 0.8 | -0.60 |
| $Y$ | 0.03 | 0.09 | 2 | 0.5 | -0.75 |

We present results for two values of the true correlation, $\rho = 1/4$ and $\rho = 2/3$, which are typical levels of the correlation in our empirical analysis. In the Supplementary Material we present the corresponding results for $\rho = 1/2$ and $\rho = 3/4$.

### 4.1 Case without Noise

We first consider the case where prices are observed without measurement error. This defines the limit to which we can apply subsampling. For instance, for sparsely sampled 1-minute returns we can set $S = 60$. In the absence of noise, there is no need to sample sparsely, but we gain valuable insight about the the estimators by studying their properties at lower sampling frequencies.

The Heston model generates prices process with time-varying volatilities. For this reason, we should not expect $P$ and $K$ to be consistent. While $Q$ and $Q_S$ are consistent, they may have a bias in finite samples, because sampling error in $\hat{\tau}$ and the non-linear transformation, $\rho = \sin(\frac{\pi}{2}\tau)$, will induce a finite-sample bias in $Q$ and $Q_S$.

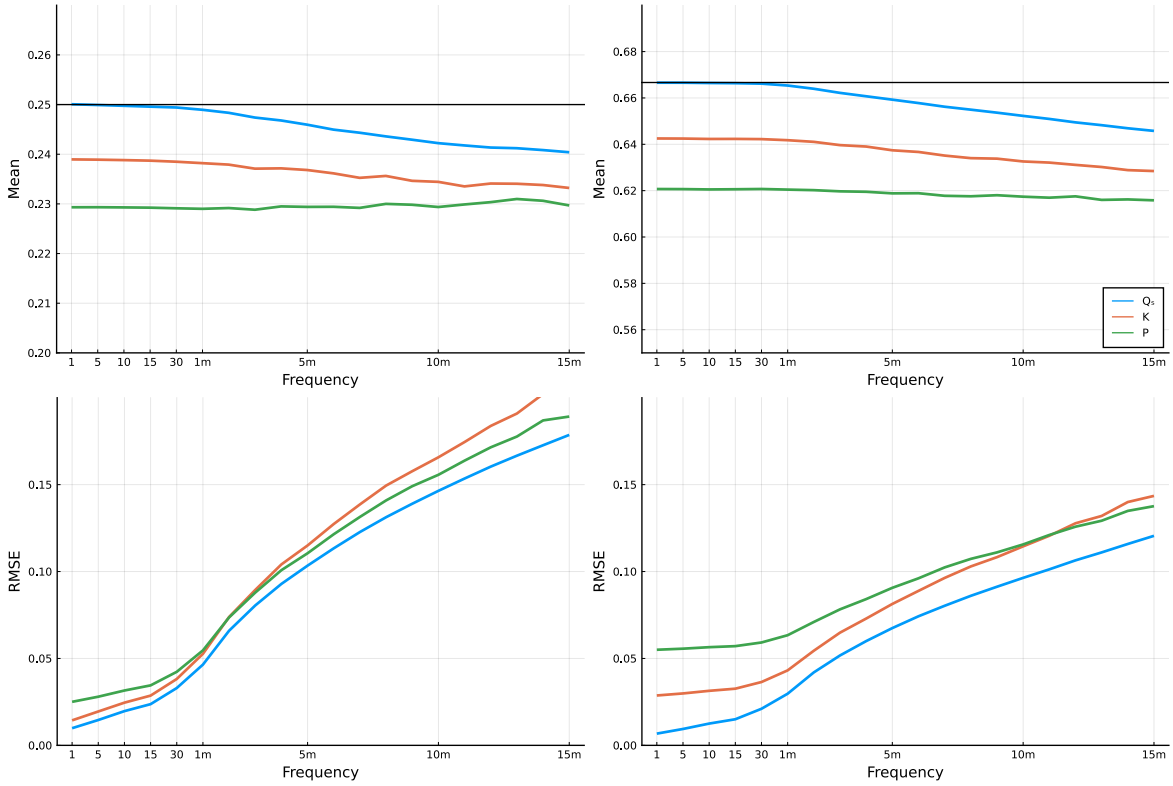



Figure 4: Means and RMSEs for the estimators as a function of sampling frequencies. Prices are generated by the Heston model with the true correlation being $\rho = 0.25$ (left panels) and $\rho = 0.6\overline{6}$ (right panels).

The average values of the estimators are shown in the upper panels of Figure 4 for the case were $\rho = 0.25$ and $\rho = 0.6\overline{6}$. As expected, $P$ and $K$ are biased as expected, since volatility is time-varying in the Heston model, which $P$ being substantially more biased than $K$. At the highest sampling frequency, $P$ and $K$ become very accurate estimates of incorrect quantities, as defined in Theorem 2. The $Q_S$ estimator is largely unbiased when returns are sampled more frequently that every minute. At slower

sampling frequencies a bias begin to emerge in $Q_S$, which is a consequence of Jensen's inequality. The variance of $\hat{\tau}$ increases with the sampling frequency and the concavity of $\tau \mapsto \sin(\frac{\pi}{2}\tau)$ for $\tau > 0$ explains the downwards bias that becomes evident at slow sampling frequencies. However the bias of $Q_S$ is substantially smaller than those of $K$ and $P$.

The corresponding root mean squared errors (RMSEs) are shown in the two lower panels. The new estimator has the smallest RMSE, which is driven by its ability to reduce the bias.

## 4.2 Microstructure Issues

Next, we amend the simulation to mimic features commonly seen in empirical data. We do so, by adding different forms of market microstructure noise. Noise will influence estimators in different ways. Noise that only alters the sign of a small fraction of returns will have minute impact on the robust estimators, but could have a large impact on $P$. Rare outliers provide an example of this scenario, and can be inferred directly from the influence functions for the different estimators.

### 4.2.1 Independent Noise

Independent noise in the price processes can induce the Epps effect. The independent noise reduces the correlation in returns and this downwards bias is increasing in the sampling frequency. We simulate independent noise as follows:

$$X_t = X_t^* + \epsilon_{xt}$$

where $\epsilon_{xt} \sim iidN(0, \omega_x^2)$ and similar for $Y_t$ with $\epsilon_{xt}$ independent of $\epsilon_{yt}$. Following similar simulation designs in this literature, see e.g. Bandi and Russell (2006) and Barndorff-Nielsen et al. (2008), we set $\omega_x^2 = \xi^2 \sqrt{T^{-1} \sum_{i=1}^{T} \sigma_{x,i/T}^4}$ with $\xi^2 = 0.001$, such that variance of the noise is proportional to square root of the integrated quarticity.

### 4.2.2 Prices with tick-size increments

In practice, high-frequency financial prices are restricted to a grid defined by their tick-size. This induces a particular type of market microstructure noise, as analyzed in Delattre and Jacod (1997), Horel (2007), Rosenbaum (2009), Mancini and Gobbi (2012), Hansen (2015), Li and Mykland (2015), Hansen et al. (2016), and Li et al. (2018). We will study this phenomenon by letting observed prices

be given by

$$X_t = \alpha\lfloor X_t^*/\alpha \rfloor \qquad \text{and} \qquad Y_t = \alpha\lfloor Y_t^*/\alpha \rfloor,$$

where $\alpha$ defines the coarseness of the grid.[7] In our simulations we let the coarseness be proportional to the level of volatility, $\alpha = c\sigma$ in order to control the average number of price changes within a given period of time. The true price processes are, as before, define by (3).

#### 4.2.3 Tick size with Noise

Next we add noise to the grid of observed prices. Specifically we now observe

$$X_t = \alpha\lfloor X_t^*/\alpha \rfloor + \epsilon_{xt}, \qquad \text{with} \quad \epsilon_{xt} = \begin{cases} 0 & \text{with probability } 1-p, \\ \pm\alpha & \text{with probability } p/2, \end{cases}$$

and similarly for $Y_t$ with $\epsilon_{xt}$ and $\epsilon_{yt}$ independent.

#### 4.2.4 Stale Prices

We introduce stale pricing using

$$X_{\frac{j}{N}} = \begin{cases} \alpha\lfloor X^*_{\frac{j}{N}}/\alpha \rfloor & \text{with probability } 1-q, \\ X_{\frac{j-1}{N}} & \text{with probability } q. \end{cases} \tag{4}$$

This will generate "flat pricing" and the expected duration between price updates will be $(1-q)^{-1}/N$.

#### 4.2.5 Jumps

Jumps are prevalent in high-frequency prices, and we could generate such with

$$X_t = X_t^* + \sum_{s \le t} J_s^x \qquad Y_t = Y_t^* + \sum_{s \le t} J_s^y,$$

where $J_t^x$ and $J_t^y$ denote jump processes. The impact that jumps have on the estimators is characterized by their influence functions. The robust estimators, $Q_S$ and $K$, are essentially unaffected by jumps, whereas $P$ is highly sensitive. Independent jumps will cause $P$ to be biased towards zero, whereas a

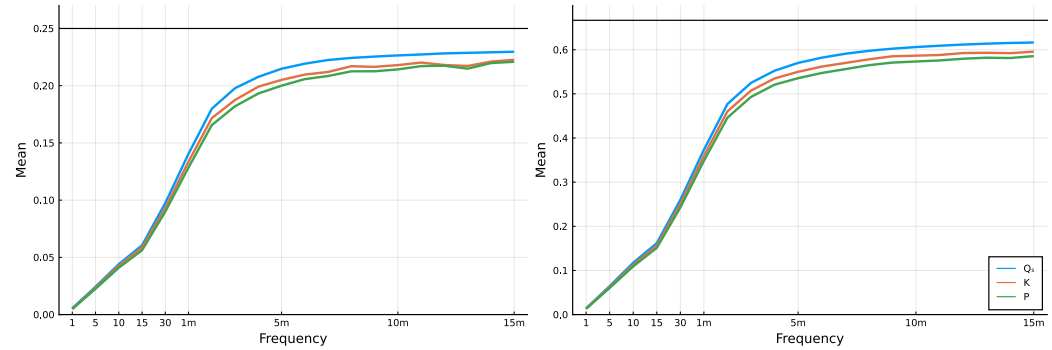
[7] In reality it is nominal prices, $\exp X_t$ and $\exp Y_t$ that are confined to a grid, but it makes no practical difference over trading day.

co-jump (a simultaneous jump in both series) will bias $P$ towards $-1$ or 1, depending on the sign of $J_s^x J_s^y$. Co-jumps in the same direction will cause $P$ to be biased towards one, whereas co-jumps in the opposite direction will cause $P$ to be biased towards $-1$. Jumps can be alleviated by truncation methods, see Mancini (2001, 2009) and Andersen et al. (2012). Simulation results with jumps are presented in the Supplementary Material.

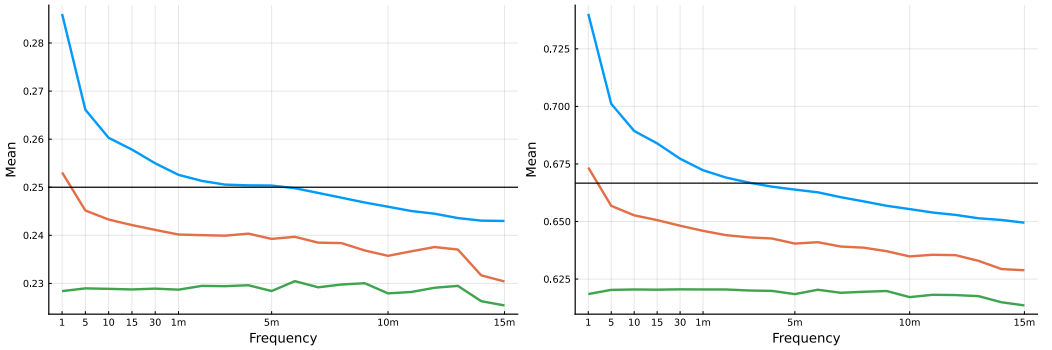
### 4.3 Simulation Results

The bias that different types of noise induce on the correlation estimators are show in Figure 5. The true correlation is $\rho = 1/4$ in the left panels and $\rho = 2/3$ in the right panels. Results for additional levels of correlation and types of noise are presented in the supplementary material. Panel (a) in Figure 5 presents the results when the efficient prices are contaminated with independent Gaussian noise with a variance that is about $10^{-3}$ times the square root of integrated quarticity of the two series. Independent noise is one (of several ways) to bring about the Epps effect. The independent noise reduces the correlation between returns, which induces a downwards bias that increases with sampling frequency, to an extend that all estimators essentially becomes noisy estimates of zero when computed with 1-second intraday returns. Independent noise is a good stating point for studying estimators, but there is overwhelming empirical evidence that contradicts the independent noise assumption in high frequencies data, see Hansen and Lunde (2006), which is also the case in our empirical analysis.

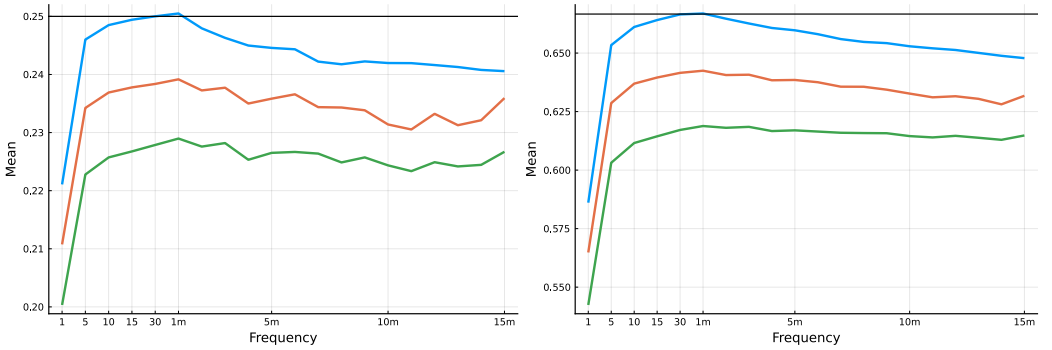
The correlation signature plots in our empirical analysis resemble those in Panel (b) of Figure 5, where the noise is defined by a rounding error ($\alpha = 10^{-4}$), to resemble the tick size in prices. Interestingly, the rounding error causes $Q_S$ to be upwards bias a higher sampling frequencies. This is also true for $K$, but to a much lesser extend, whereas $P$ is largely unaffected, but maintains the downwards bias caused by time-varying volatilities. In Figure 5 (c) we consider the same level of rounding error ($\alpha = 10^{-4}$) and add additional noise by shifting the price up or down by one tick size with equal probability, $p/2$ with $p = 0.75$. This induces a downwards bias, which is most pronounced a high sampling frequencies. Finally, in Figure 5 (d) we add additional staleness to prices on the grid, as defined by (4), where one price series remains stale with probability $q_x = 0.5$ and the other series remains stale with probability $q_y = 0.8$. The combined impact of rounding and staleness is a sizable downwards bias.



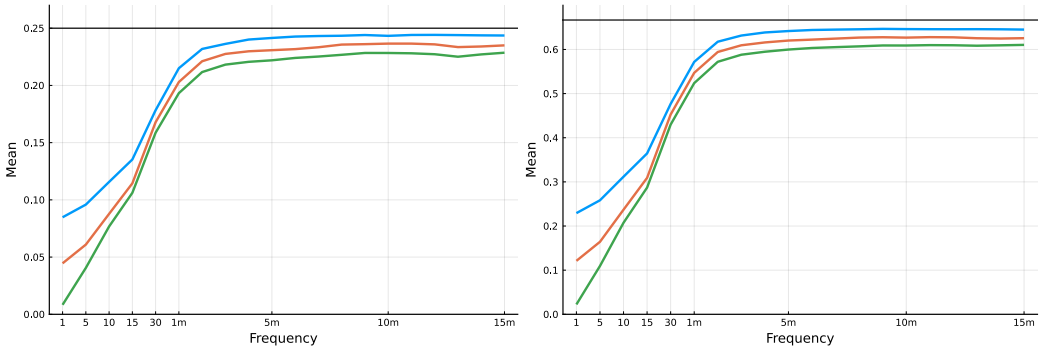

(a) Heston with independent noise




(b) Heston with rounding to a grid

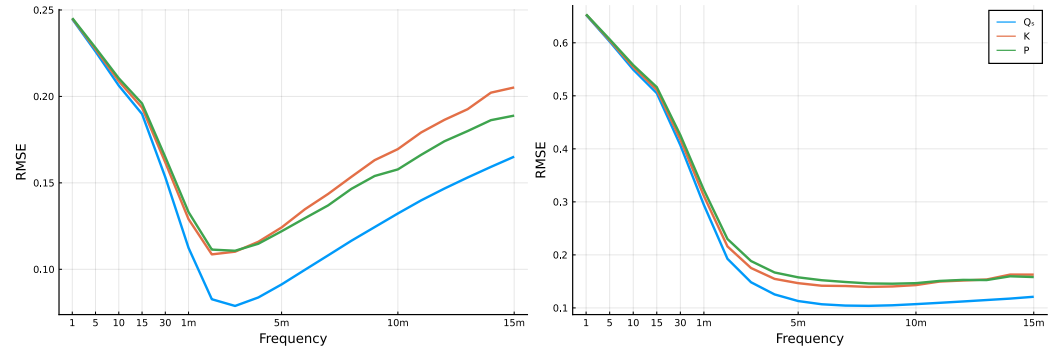



(c) Heston with rounding and grid-noise

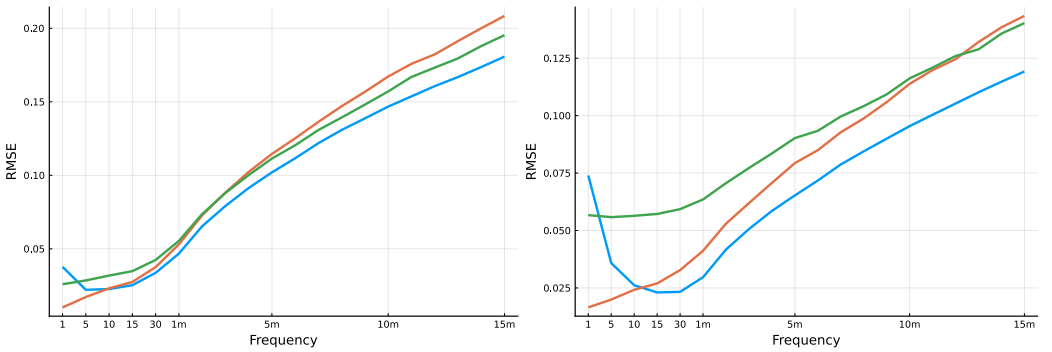



(d) Heston with rounding and staleness

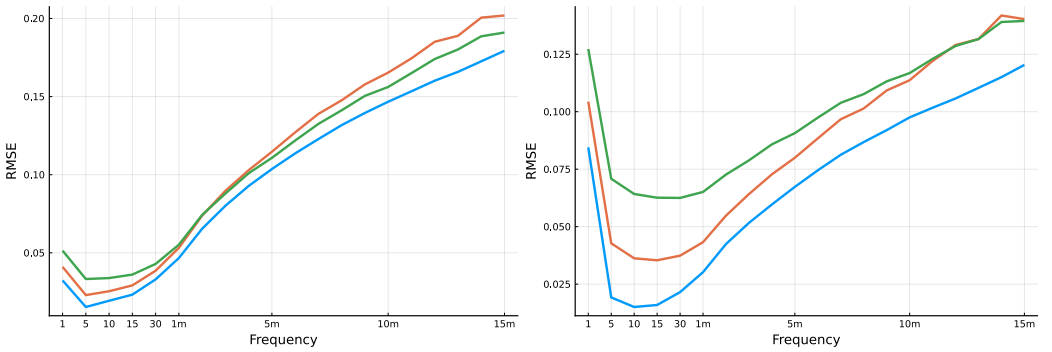
Figure 5: Correlation signature plots, with sampling frequency ranging from 1 second to 15 minutes. Observed prices are generated with Heston models with a layer of noise added. The true correlation is $\rho = 1/4$ in left panels and $\rho = 2/3$ in right panels. The four types of noise are: (a) Independent noise ($\xi^2 = 10^{-3}$); (b) rounding to a grid ($\alpha = 10^{-4}$); (c) rounding with grid-noise ($\alpha = 10^{-4}$ and $p = 0.75$); and (d) rounding with stale prices ($\alpha = 10^{-4}$ and $q_x = 0.50$ and $q_y = 0.80$).



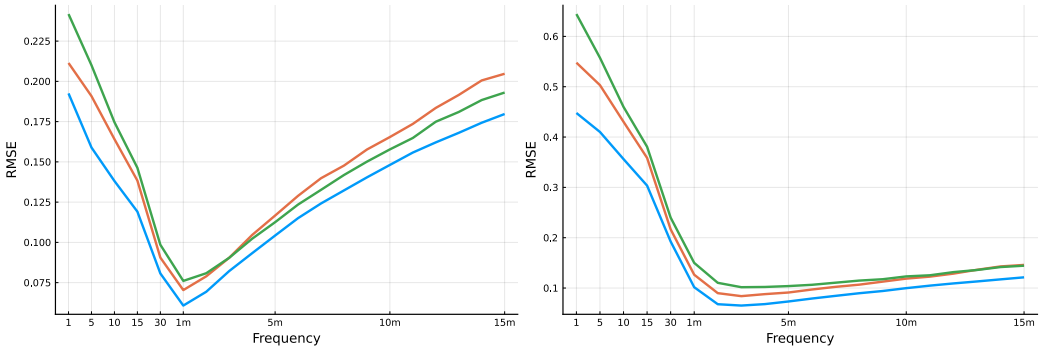

(a) Heston with independent noise




(b) Heston with rounding to a grid




(c) Heston with rounding and grid-noise




(d) Heston with rounding and staleness

Figure 6: RMSE signature plots. Observed prices are generated with Heston models with a layer of noise added. The true correlation is $\rho = 1/4$ in left panels and $\rho = 2/3$ in right panels. The four types of noise are: (a) Independent noise, (b) rounding to a grid ($\alpha = 10^{-4}$) (c) rounding with grid-noise ($\alpha = 10^{-4}$ and $p = 0.75$); and (d) rounding with stale prices ($\alpha = 10^{-4}$ and $q_x = 0.50$ and $q_y = 0.80$).

The corresponding root mean squares errors (RMSEs) are reported in Figure 6. The new correlation estimator, $Q_S$, tends to have the smallest RMSE, which is also true for the additional simulation experiments presented in the Supplementary Material.

# 5 Empirical Application

We apply the correlation estimators to high-frequency data for about 100 assets. We begin by analyzing and comparing their daily correlation estimates. For instance, we use correlation signature plots to study market microstructure noise, and explore how sensitive the estimators are to the choice of sampling frequency, as defined by $\delta$. Then we turn to estimation of intraday correlations, which we find to vary substantially over the hours with active trading. Correlations between stocks and the market are, on average, increasing for all assets in our sample. We obtain estimates of intraday betas, by combining the correlation estimates with estimates of relative volatility. We then proceed to related intraday variation in correlations and betas with asset characteristics, such as low frequency based market beta, market capitalization, and book-to-market valuations. This part of the analysis is done with an expanded set of assets detailed below.

Our sample period covers the period from January 1, 2015 to December 31, 2021 and includes 1,763 trading days. We use NYSE and NASDAQ transaction prices from the TAQ database that were accessed through the Wharton Research Data Services (WRDS) system. The data were cleaned following the guidelines in Barndorff-Nielsen et al. (2011*a*), and prices (when unavailable) were interpolated by the previous-tick methods. We will analyzed 22 stocks and SPY, an exchange traded fund that tracks the S&P 500 index, in great details. We label this data set "Small Universe". The 22 stocks were selected to be the two largest stocks (by market capitalization) within each of the eleven GICS[8] sectors. A larger set of asset of assets, "Large Universe" is used to identify asset-characteristics associated with different patterns in intraday market betas. The Large universe includes the assets in the S&P 100 index, as of [date], we excluded two of these assets from the Large Universe. PYPL (PayPal) was excluded because it only started trading in 2015 after being spun off eBay, and RTX (formerly Raytheon Tech) was excluded because it merged with United Technologies, which was completed in April 2020.

[8]Global Industry Classification Standard.

Table 2: Summary Statistics Small Universe

| Sector (GICS code) | Ticker | Average Price | Duration (seconds) | Zero-returns (%) $\delta = 1$s | Zero-returns (%) $\delta = 3$m |
|---|---|---|---|---|---|
| | SPY | 281.89 | 2.28 | 63.39 | 3.73 |
| Energy | HAL | 33.38 | 5.56 | 89.25 | 8.37 |
| (10) | XOM | 72.44 | 3.83 | 82.73 | 6.71 |
| Materials | LYB | 89.55 | 9.23 | 90.85 | 4.94 |
| (15) | NEM | 40.10 | 6.67 | 89.76 | 8.73 |
| Industrials | AAL | 34.32 | 6.83 | 89.76 | 8.54 |
| (20) | UNP | 144.23 | 6.86 | 87.82 | 4.29 |
| Consumer Discretionary | TSLA | 415.90 | 5.12 | 70.38 | 1.14 |
| (25) | AMZN | 1658.92 | 5.03 | 76.60 | 0.75 |
| Consumer Staples | PG | 101.81 | 4.86 | 85.85 | 6.92 |
| (30) | WMT | 99.02 | 4.46 | 84.87 | 6.75 |
| Health Care | JNJ | 131.77 | 4.57 | 83.74 | 5.89 |
| (35) | MRK | 68.76 | 5.06 | 87.05 | 7.50 |
| Financials | JPM | 101.17 | 3.22 | 78.24 | 5.53 |
| (40) | WFC | 48.11 | 4.50 | 86.83 | 7.84 |
| Information Technology | AAPL | 167.52 | 2.27 | 63.05 | 3.57 |
| (45) | AMD | 33.45 | 14.87 | 82.87 | 19.17 |
| Communication Services | DIS | 121.95 | 4.19 | 80.95 | 5.18 |
| (50) | FB | 181.59 | 2.80 | 68.63 | 3.18 |
| Utilities | D | 75.08 | 8.04 | 91.35 | 7.92 |
| (55) | DUK | 84.94 | 7.47 | 90.61 | 7.26 |
| Real estate | AMT | 168.88 | 10.18 | 91.43 | 4.15 |
| (60) | PLD | 72.88 | 9.77 | 92.62 | 8.70 |

Note: Summary statistics for SPY and 22 assets (two from each of the 11 sectors) for the sampling period from January 1, 2015 to December 31, 2021. Average price, average duration between two consecutive transactions are listed along with the percentage of zero returns when returns are sampled at 1 second and 3 minutes, respectively.

Table 2 presents the summary statistics for the Small Universe with 22 assets. The exchange traded fund, SPY, is the most frequently traded asset, followed by AAPL and FB. On average, these securities have just over 2 seconds between transaction prices. The price range is an interesting statistic, because the tick-size is more likely to induce rounding errors and price staleness for assets trading at low prices. This appears to be relevant for AMD that traded for less than \$3 in all of 2015 and below \$10 during most of the first three years in our sample period. This likely explains the many zero increments. More than 19% of all 3-minute returns are zero in this sample period.

The assets in the Large Universe are listed and organized by sectors in Table 3.

Table 3: Large Universe

| Energy | Materials | Industrials | Consumer Discretionary | Consumer Staples | Healthcare |
|---|---|---|---|---|---|
| COP | DOW | BA | AMZN | CL | ABBV |
| CVX | LIN | CAT | BKNG | COST | ABT |
| XOM | | EMR | F | KHC | AMGN |
| | | FDX | GM | KO | BMY |
| | | GD | HD | MDLZ | CVS |
| | | GE | LOW | MO | DHR |
| | | HON | MCD | PEP | GILD |
| | | LMT | NKE | PG | JNJ |
| | | MMM | SBUX | PM | LLY |
| | | UNP | TGT | WBA | MDT |
| | | UPS | TSLA | WMT | MRK |
| | | | | | PFE |
| | | | | | TMO |
| | | | | | UNH |

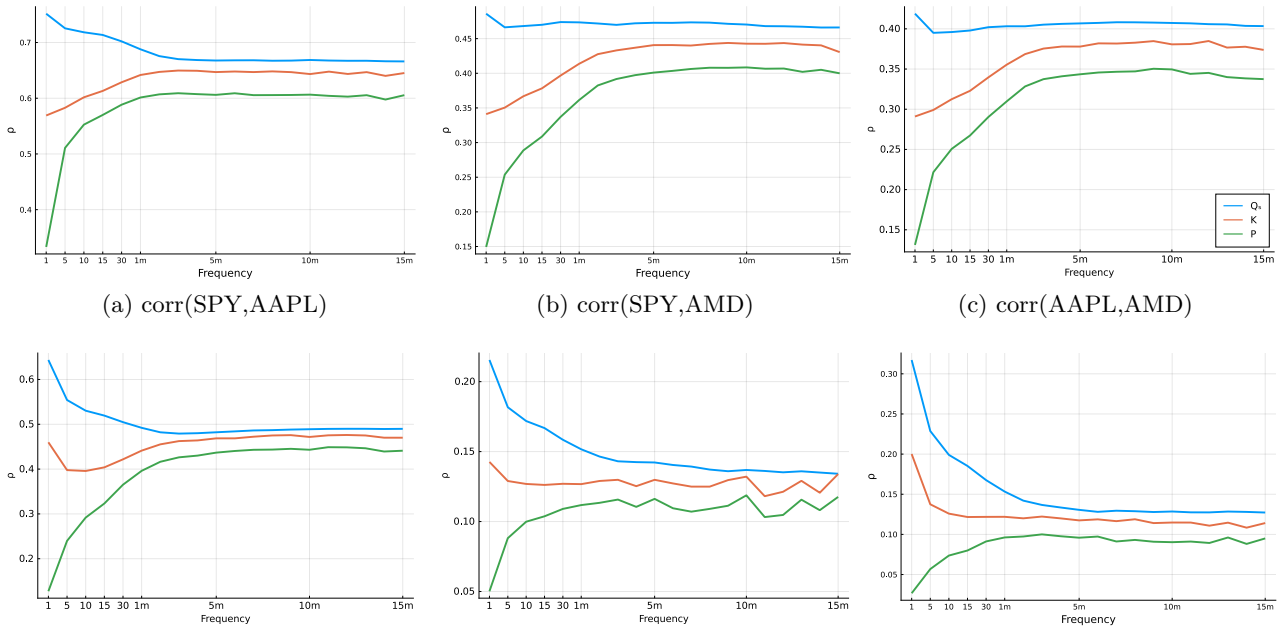
| Financials | Information Technology | Telecom. Services | Utilities | Real Estate |
|---|---|---|---|---|
| AIG | AAPL | CHTR | DUK | AMT |
| AXP | ACN | CMCSA | EXC | SPG |
| BAC | ADBE | DIS | NEE | |
| BK | AMD | GOOGL | SO | |
| BLK | AVGO | FB | | |
| BRKB | CRM | NFLX | | |
| C | CSCO | T | | |
| COF | IBM | TMUS | | |
| GS | INTC | VZ | | |
| JPM | MA | | | |
| MET | MSFT | | | |
| MS | NVDA | | | |
| SCHW | ORCL | | | |
| USB | QCOM | | | |
| WFC | TXN | | | |
| | V | | | |

Note: List of assets in "Large Universe", organized by sectors.

### 5.1 Estimates of Daily Correlations

We apply the correlation estimators to daily high frequency data using calendar-time sampling with frequencies ranging from 1 second to 15 minutes. The resulting correlation signature plots are shown in Figure 7 for a subset of the assets. These are the two most actively traded securities, SPY and AAPL, the stock with most zero returns, AMD, and the two stocks from the Material sector, LYB and NEM, whose liquidity and percentage of zero returns is more typical for assets in the Small Universe.

Signature plots were introduced in Andersen et al. (2000) who plotted the average realized variance against the sampling frequency used to compute the underlying intraday returns. Signature plots help identify bias in the estimators, which tend to be most pronounced at high sampling frequencies. If the estimator is unbiased over a range of sampling frequencies, then the signature plot will be roughly flat over that those sampling frequencies.




(a) corr(SPY,AAPL) (b) corr(SPY,AMD) (c) corr(AAPL,AMD)

(d) corr(SPY,LYB) (e) corr(SPY,NEM) (f) corr(LYB,NEM)

Figure 7: Correlation signature plots, where the average correlation estimate is plotted against sampling frequency for the three estimators, $Q_S$, $P$, and $K$.

The signature plots in Figure 7 are signature plot for correlations, which can be used to visualize biases, such as the Epps effect. Here we observe that many of the plots have patterns that resemble the effect for rounding to a grid, because $Q_S$ often has an upwards bias at high sampling frequencies, while $P$ has a downwards biased. Additional signature plots are presented in the Supplementary Material, see Figure S.1. We adopt 3-minutes as a common sampling frequency for all estimators. This is in part motivated by the signature plots tend to be flat for $\delta \geq$3 minutes, and in part because it makes our results more comparable to those in ATT. Ideally, one would determine an empirical way to select an optimal sampling frequency, because the optimal sampling frequency likely varies over time and across assets. We leave this for future research.

Table 4 presents summary statistics for correlations between stocks in the same sector. For each of the three estimators, we compute the average, median, and interquartile range across the 1,763 daily

Table 4: Correlation estimates.

| Sector code (Asset pair) | Average | | | 25th quantile | | | Median | | | 75th quantile | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $Q_S$ | $K$ | $P$ | $Q_S$ | $K$ | $P$ | $Q_S$ | $K$ | $P$ | $Q_S$ | $K$ | $P$ |
| 10 (HAL,XOM) | 0.59 | 0.55 | 0.53 | 0.49 | 0.45 | 0.42 | 0.60 | 0.56 | 0.54 | 0.70 | 0.66 | 0.65 |
| 15 (LYB,NEM) | 0.14 | 0.11 | 0.09 | 0.02 | 0.00 | -0.03 | 0.14 | 0.12 | 0.10 | 0.25 | 0.23 | 0.22 |
| 20 (AAL,UNP) | 0.31 | 0.28 | 0.25 | 0.19 | 0.17 | 0.13 | 0.29 | 0.27 | 0.24 | 0.41 | 0.38 | 0.37 |
| 25 (AMZN,TSLA) | 0.37 | 0.35 | 0.33 | 0.24 | 0.23 | 0.21 | 0.35 | 0.34 | 0.33 | 0.48 | 0.48 | 0.45 |
| 30 (PG,WMT) | 0.41 | 0.37 | 0.33 | 0.29 | 0.25 | 0.20 | 0.40 | 0.36 | 0.33 | 0.52 | 0.49 | 0.46 |
| 35 (JNJ,MRK) | 0.56 | 0.52 | 0.49 | 0.46 | 0.41 | 0.38 | 0.57 | 0.52 | 0.50 | 0.66 | 0.63 | 0.61 |
| 40 (JPM,WFC) | 0.74 | 0.72 | 0.68 | 0.67 | 0.64 | 0.61 | 0.76 | 0.72 | 0.70 | 0.83 | 0.79 | 0.78 |
| 45 (AAPL,AMD) | 0.40 | 0.37 | 0.33 | 0.26 | 0.21 | 0.16 | 0.40 | 0.36 | 0.33 | 0.55 | 0.52 | 0.49 |
| 50 (DIS,FB) | 0.34 | 0.31 | 0.27 | 0.20 | 0.17 | 0.14 | 0.33 | 0.30 | 0.26 | 0.48 | 0.45 | 0.42 |
| 55 (D,DUK) | 0.76 | 0.72 | 0.70 | 0.70 | 0.66 | 0.63 | 0.78 | 0.73 | 0.71 | 0.83 | 0.79 | 0.78 |
| 60 (AMT,PLD) | 0.52 | 0.48 | 0.44 | 0.43 | 0.38 | 0.32 | 0.53 | 0.48 | 0.46 | 0.63 | 0.59 | 0.57 |

Note: Summary statistics for pairs of assets, within each of the 11 sectors: Energy (10), Materials (15), Industrials (20), Consumer Discretionary (25), Consumer Staples (30), Health Care (35), Financials (40), Information Technology (45), Communication Services (50), Utilities (55), Real Estate (60).

estimates. For all pairs, these quantities are similar for the three estimators. The interquartile range is across days in the sample that predominately is driven by time-variation in the daily correlation. So, a similar width for the interquartile range should not be interpreted as the estimators having similar precision. In the next subsection, we present results that strongly indicate that $Q_S$ is more precise than $K$ an $P$. While the measurements are similar for the three estimators, we always have $Q_S > K > P$. This ordering is in line with our theoretical results, that time-varying volatility induces a bias in $P$ than in $K$, and that $P$ is more biased than $K$.

Next, we estimate daily correlations between each of the 22 stocks and SPY. The average, median, and interquartile range (over the 1,763) estimates are shown in Figure 8. Once again we see that the quantities are similar for the three estimators, as was the case in Table 4, and once again do we have $Q_S > K > P$ uniformly across all assets and across all measurements.

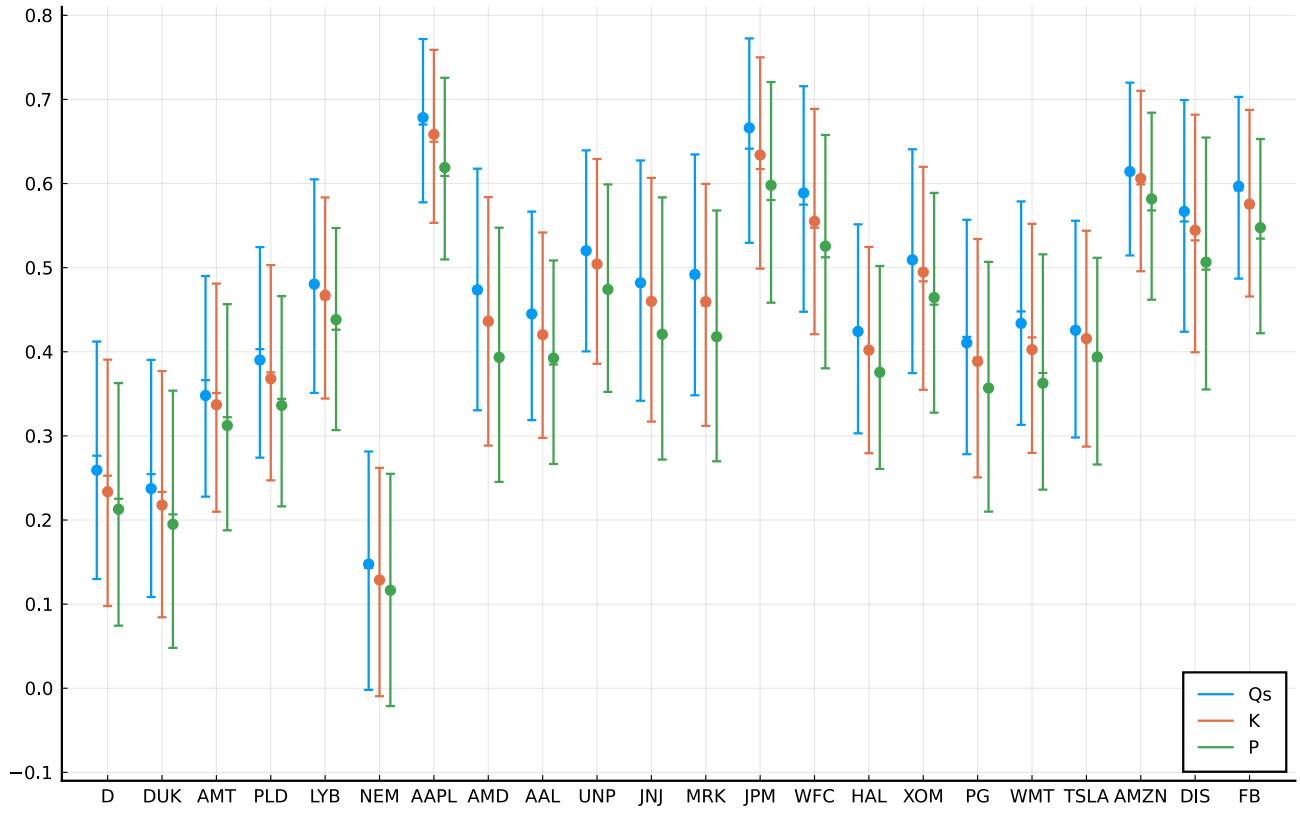



Figure 8: Daily correlations between assets and market returns. The three estimators were applied to 1,763 trading days. The average estimate (dash) and median estimate (bullet) are shown. The vertical lines present the interquartile range over the 1,763 daily estimates.

### 5.2 Intra-day Correlations and Market Beta Estimation

Estimating betas from high frequency data is an active research area, see e.g. Andersen et al. (2005, 2006), Todorov and Bollerslev (2010), Dovonon et al. (2013), Hansen et al. (2014), and Reiß et al. (2015). This literature as also documented substantial time variation in the betas over time. Recently, Andersen et al. (2021) (ATT) documented systematic time-variation in betas within the trading day. Specifically, they estimated betas for rolling windows (spanning two hours) using 3-minute intraday returns. Their local estimate of beta is simply a local estimator of the covariance between asset and market returns divided by a local estimates of the quadratic variation of market returns. The local (time-of-the day) estimates are averaged over the 2,243 days in their sample period (2010-2018). Interestingly, ATT found a great deal of variation in the betas within the day. Some stocks have increasing betas over the trading day while other other assets had decreasing betas over the day.

We can use correlation estimators to cast new light on the patterns in intraday market betas, by

decomposing the market beta into correlation multiplied by relative volatility,

$$\beta_{i,t} = \rho_{i,t} \times \lambda_{i,t}, \qquad \lambda_{i,t} = \frac{\sigma_{i,t}}{\sigma_{0,t}},$$

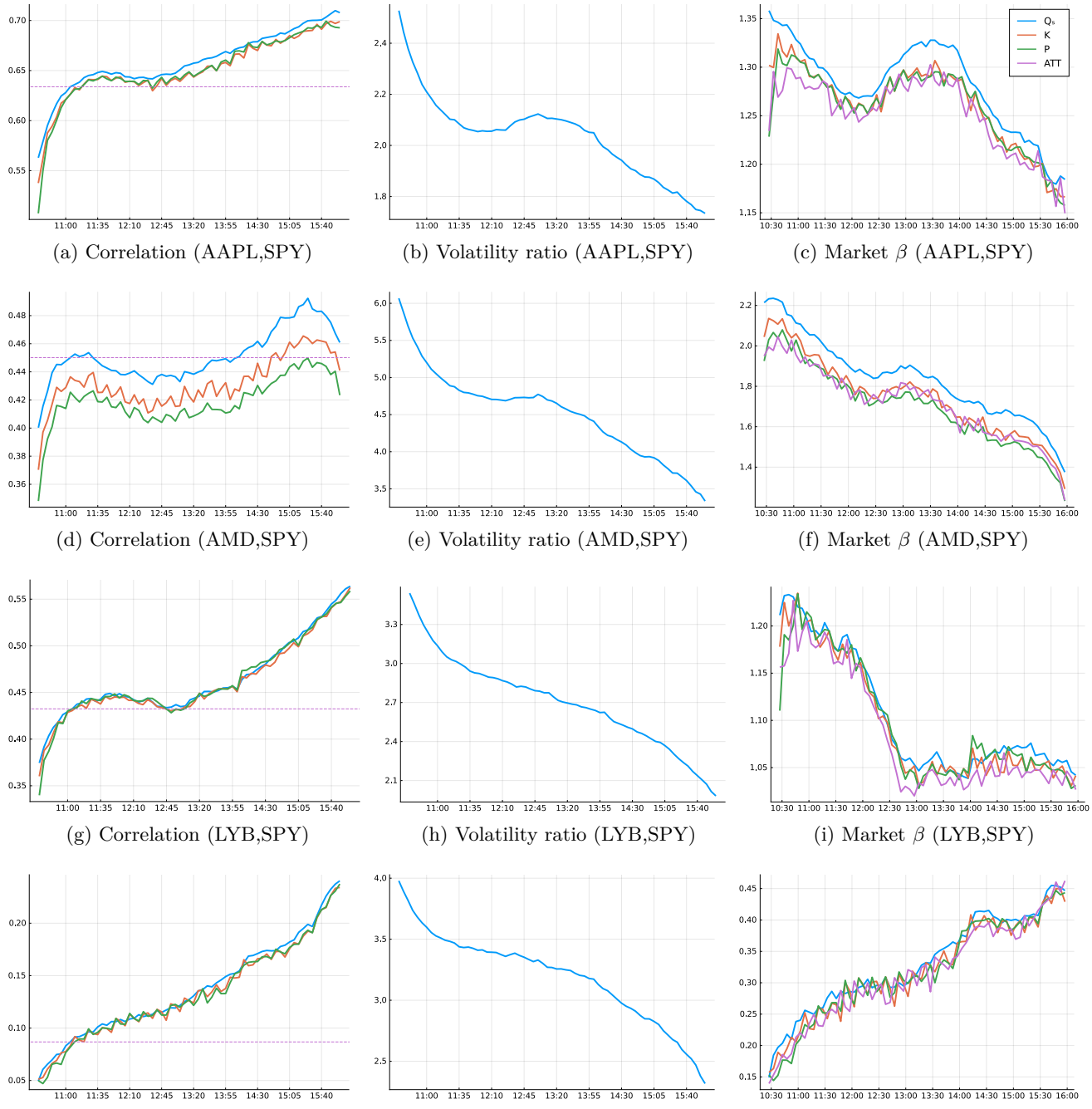
where $\rho_{i,t}$ is the correlation between the $i$-th asset and the market and $\sigma_{i,t}$ and $\sigma_{0,t}$ are the volatilities for the $i$-th asset and the market, respectively. We will estimated local market betas using local correlation estimators combined with estimators of relative volatility. Specifically, we compute $P$, $K$, and $Q_S$ and $\lambda$ using a rolling window with 60 minutes of high frequency data. We will the investigate how much of the intraday variation in betas is explained by into intraday variation in correlations and how much can be ascribed to intraday variation in relative volatility.

We estimate $\rho_{i,t}$ with each of the correlation estimators, $P$, $K$, and $Q_S$, using a rolling window that spans 60 minutes. Similarly, we estimate the relative volatility, $\lambda_{i,t}$, with subsampled range-based estimators with truncations, as defined by

$$\hat{\lambda}_{i,t} = \frac{\sum_{j \in I_t} \left[\!\left[ \Delta_{\frac{S}{N}} Y_{\frac{j}{N}} \right]\!\right]_{\nu_y}}{\sum_{j \in I_t} \left[\!\left[ \Delta_{\frac{S}{N}} X_{\frac{j}{N}} \right]\!\right]_{\nu_x}}, \qquad [\![x]\!]_{\nu} = \begin{cases} x & \text{if } |x| < \nu \\ 0 & \text{otherwise} \end{cases} \qquad I_t = [\lfloor tN \rfloor - W + S, \lfloor tN \rfloor],$$

where $\nu_x$ and $\nu_y$ are adaptive thresholds for jump truncation. These are defined by $\nu_x = 4\sqrt{\mathrm{BV}_x}/n^{0.49}$ and $\nu_y = 4\sqrt{\mathrm{BV}_y}/n^{0.49}$, where $\mathrm{BV}_x = \frac{\pi}{2} \sum_{j=2}^{n} |\Delta_\delta X_{j\delta}||\Delta_\delta X_{(j-1)\delta}|$ is the jump robust bipower variation estimator of daily integrated volatility, see Barndorff-Nielsen and Shephard (2004*b*), and $\mathrm{BV}_y$ is defined analogously. We have explored estimation of relative volatility using the bipower variation measures, with the same thresholds. These estimated we virtually identical to those of $\hat{\lambda}_{i,t}$.

In our implementation, we have $N = 23,400$, $W = 3,600$, and $S = 180$. This results in 3,421 overlapping 3-minute returns within each hour we use to compute the subsampled quantities.




(a) Correlation (AAPL,SPY) (b) Volatility ratio (AAPL,SPY) (c) Market $\beta$ (AAPL,SPY)

(d) Correlation (AMD,SPY) (e) Volatility ratio (AMD,SPY) (f) Market $\beta$ (AMD,SPY)

(g) Correlation (LYB,SPY) (h) Volatility ratio (LYB,SPY) (i) Market $\beta$ (LYB,SPY)

(j) Correlation (NEM,SPY) (k) Volatility ratio (NEM,SPY) (l) Market $\beta$ (NEM,SPY)

Figure 9: Intraday correlations (left panels), relative volatilities (middle panels), and market $\beta$ (right panels) for four assets, AAPL, AMD, LYB, and NEM, relative to SPY that tracks the S&P 500 index. Quantities are estimated with a rolling window of data that spans the 60 minutes leading up to the indicated timestamp. Estimates are averaged over the days in the sample period. The estimates of market betas using the methodology in ATT are included in the right panels.

Intraday estimates of correlations, relative volatilities, and betas are show for four assets in Figure 9. All quantities are estimated using a rolling window that spans 60 minutes return. The time-stamp

used along the x-axis refers to the end of the 60 minute period. The estimates are averaged over the 1,763 trading days in the sample. The left panels report the intraday correlation estimates for $P$, $K$, and $Q_S$. A horizontal dashed line indicate the average correlations over the trading hours. The middle panels report the relative volatility as defined by $\hat{\lambda}_{i,t}$ above, which is multiplied by the three correlation estimates to obtain estimates of market betas. These three intraday market betas are shown in the right panels along with the regression based estimate, based on the same methodology as ATT. The corresponding results for all assets in the Small Universe is presented in the Supplementary Material, Figure S.3.

We note that the time variation in the estimates of $Q_S$ tends to be smoother than those of $K$ and $P$. This strongly suggests that $Q_S$ is a more accurate than $K$ and $P$. We also note that $Q_S$ tends to be slightly larger than $K$ and $P$, which may be related to them having a larger variance causing another source of bias in $K$ and $P$. These smoother lines for $Q_S$ and slightly larger values carries over to the intraday estimates of market betas. The lines for the regression based estimates of intraday betas are also less smooth that those based on $Q_S$.

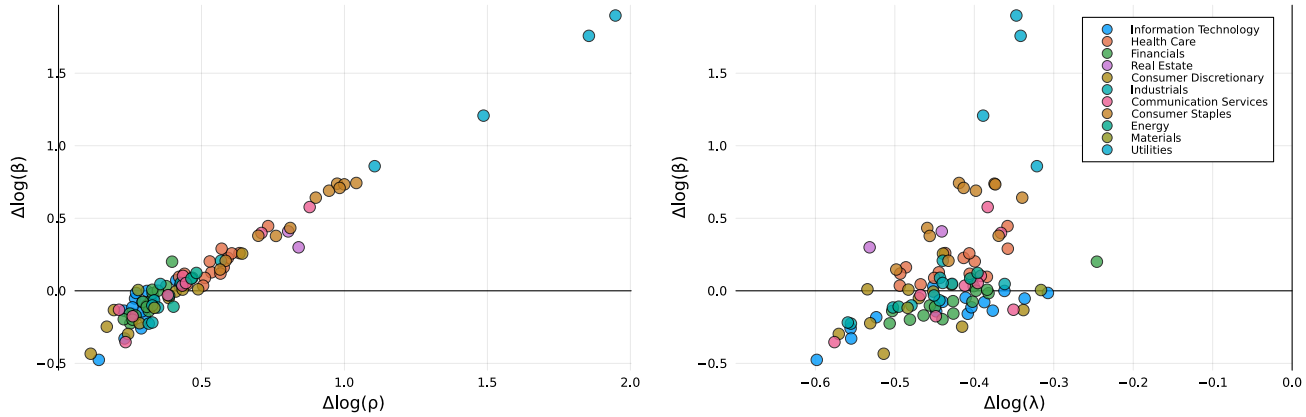
We find correlations to be generally increasing over the day, while relative volatilities are decreasing. Whether their product, the market beta, is increasing or decreasing will depend on which of the terms changes the most. Unlike correlations and relative volatility, the paths for intraday market betas take many different shapes. Some assets have clearly increasing market beta over the day (e.g. NEM), others have decreasing market betas (e.g. AMD), and a third group of assets have market betas that goes both up and down (e.g. AAPL), or stay relatively flat for a large part of the day (e.g. LYB). It is interesting to compare the market betas for LYB and NEM, which are both Materials sector stocks, with trading intensity below the average for stocks in the Small Universe. Despite these commonalities the intraday beta patterns for LYB and NEM are very different. The reason can be found in their intraday correlations. For LYB the correlation only increases by about 50% over the trading hours (from about 0.37 to 0.55), whereas the correlation for NEM increases by nearly 500% (from about 0.05 to 0.25). A great variety of shapes for time-varying betas are shown in Figure S.3 in the Supplementary Material.

### 5.3 Decomposing Intraday Variation in Market Betas

All estimated correlations are positive, we can therefore factorize the logarithm of intraday market betas, as

$$\log \beta_{i,t} = \log \rho_{i,t} + \log \lambda_{i,t}.$$

We use this decomposition to investigate who much of the intraday variation in market betas can be ascribed to changes in correlations and changes in relative volatilities. For this purpose, we expand this part of our analysis to include all assets in the Large Universe.




(a) (b)

Figure 10: Scatterplot of changes in intraday betas plotted against changes in correlations (a) and changes in relative volatility (b). Changes are defined as the difference between the last hourly estimate of the trading day and the first hourly estimate of the trading day.

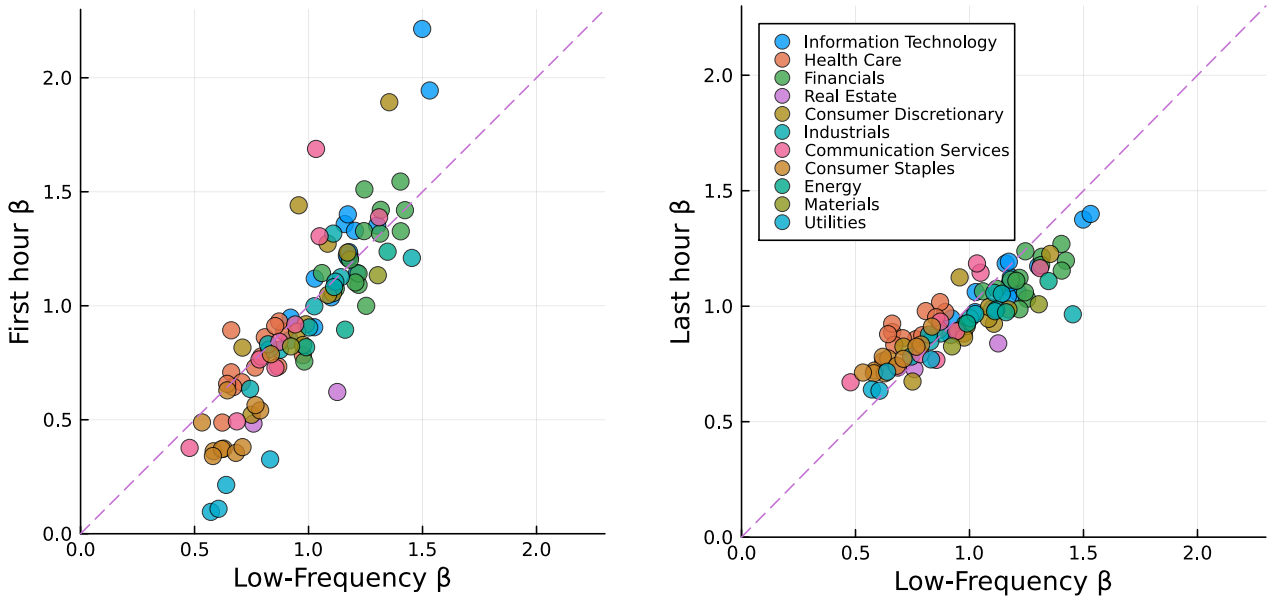
In Figure 10 we have plotted changes in intraday market betas against intraday changes in intraday correlations and against changes in relative volatility, for all assets in the Large Universe. We use color codes to indicate the GICS industry sectors for each of the assets. The increments (changes) are defined by the logarithmic difference between the estimate from the first hour of trading and the analogous quantity from the last hour of trading.

When changes in intraday betas are plotted against intraday changes in correlations, it reveals a strong linear relationship between the two, see left panel of Figure 10. In contrast there is only a weak relationship between changes in market betas and changes in relatively volatilities. In fact, as can be seen from the range of the x-axis in the right panel of Figure 10, there is far less cross-sectional variation in the changes in relative volatility. Overall, we can see that most of the cross sectional variation in market betas can be ascribed to variation in intraday correlations.

Interestingly, there is a great deal of clustering by sectors in Figure 10. In terms of changes in intraday correlations, there is a large degree of sector-specific separation, whereas in terms of changes in relative volatilities there are notable variation within all sectors, as evident by asset dispersion along the x-axis.

In the Supplementary material, Figure S.7, we have explored the intraday variation in greater details. For instance, we plot changes in intraday market correlations, relative volatility, and market betas against Fama-French type variables. We do not detect a strong association with key characteristics such as market capitalization and book-to-market ratios.

In Figure 11, we present scatterplots of the intraday market betas against conventional market betas, which are computed from daily returns. The left panel has the market beta for the first hour of trading and the right panel has the market beta for the last trading hour plotted against the low frequency market beta. Not surprisingly, do we find a strong relationship between the low-frequency market betas and the intraday market betas. The scatterplots in Figure 11 corroborates findings in ATT, who found market betas to be less disperse at the end of the day, than the beginning of the day.




(a) First hour

(b) Last hour

Figure 11: Market beta for the first and last hour plotted against the conventional market beta, which is computed from daily returns.

# 6 Concluding Remarks

The correlation coefficient is a fundamental measure of linear dependence with broad applications across various fields of empirical analysis. For instance, in modern finance, it has a central role in risk management, portfolio selection, and the pricing of derivatives.

In this paper, we have introduced a novel robust correlation estimator that is particularly well-suited for high-frequency financial data analysis. We have shown that the sample correlation, $P$, and Kendall's tau, $K$, are inconsistent under time-varying volatility, while the quadrant estimator is robust to time-varying volatilities. The subsampled quadrant estimator, $Q_S$, inherits the consistency of the quadrant and is far more efficient, because it leverages additional high-frequency data. The theoretical properties we established for the estimators are supported by simulation-based evidence and an extensive empirical analysis spanning seven years of high-frequency return data for about 100 securities.

The empirical analysis also offers valuable insights into the time-varying nature of market betas within a trading day. Market betas can be expressed as the product of the correlation (with market returns) and relative volatility. We have documented that the time-variation in market betas within the day is mainly driven by time-variation in intra-day correlations.

While the estimator, $Q_S$, is particularly well-suited for high-frequency financial data analysis, it may also be useful for other time-series with time-varying volatility, or time series that are prone to outliers and noise. The $Q_S$ estimator might also be useful for nonparametric estimation of the leverage leverage effect, as analyzed in Kalnina and Xiu (2017). There are several ways the $Q_S$ estimator could be extended and possibly improved. For instance, there might be more efficient ways to handle zero returns, such as distinguishing between cases were both returns are zero and cases were just one of the returns is zero. A multivariate version of the $Q_S$ estimator would be interesting to explore. Constructing a correlation matrix from univariate correlation estimates, need not result in a positive definite matrix. So, a subsequent matrix projection to the set of positive definite correlation matrices might be needed, such as those proposed by Higham (2002) and Qi and Sun (2006).

# A Appendix of Proofs

**Proposition A.1** (Greiner)**.** *Suppose that $(X, Y)$ is elliptically distributed with location parameter $\mu$ and dispersion matrix $\Sigma$. Then Greiner's identity (1) holds with $\rho = \Sigma_{12}/\sqrt{\Sigma_{11}\Sigma_{22}}$.*

*Proof.* An elliptical distribution has the stochastic representation,

$$\begin{pmatrix} X \\ Y \end{pmatrix} = \mu + VAU, \qquad \text{where} \quad A = \begin{pmatrix} a & b \\ c & d \end{pmatrix},$$

with $A'A = \Sigma$, and where $V > 0$ is independent of $U$, and $U = (\cos\theta, \sin\theta)'$, with $\theta \sim \text{Uniform}[-\pi, \pi]$, see Cambanis et al. (1981). From $A'A$ we have

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

which is well-defined if $\mathbb{E}[V^2] < \infty$. Since $V > 0$ it follows that

$$X > 0 \Leftrightarrow (a\cos\theta + b\sin\theta) > 0 \Leftrightarrow \sin(\theta + \phi_1) > 0 \Leftrightarrow \theta \in (\phi_1, \pi - \phi_1),$$

where $\sin\phi_1 = a/\sqrt{a^2 + b^2}$. Similarly, $Y > 0 \Leftrightarrow \theta \in (\phi_2, \pi - \phi_2)$ where $\sin\phi_2 = \frac{c}{\sqrt{c^2+d^2}}$. Thus, the quadrant probability is

$$\Pr[X > 0, Y > 0] = \Pr[\theta \in (\phi_1, \pi - \phi_1) \cap (\phi_2, \pi - \phi_2)] = \frac{\pi - |\phi_2 - \phi_1|}{2\pi}.$$

Next, observe that $\rho$ can be expressed as:

$$\rho = \sin\phi_1 \sin\phi_2 + \cos\phi_1 \cos\phi_2 = \cos(\phi_c - \phi_a) = \sin(\pi/2 - |\phi_c - \phi_a|),$$

such that

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

□

**Lemma A.1.** *Let $(X, Y, \tilde{X}, \tilde{Y})$ be normally distributed with mean zero and correlation matrix*

$$C(\rho, \omega) = \begin{bmatrix} 1 & \rho & \omega & \omega\rho \\ \rho & 1 & \omega\rho & \omega \\ \omega & \omega\rho & 1 & \rho \\ \omega\rho & \omega & \rho & 1 \end{bmatrix},$$

*and define* $Z = 1_{\{XY>0\}}$ *and* $\tilde{Z} = 1_{\{\tilde{X}\tilde{Y}>0\}}$. *Then*

$$\text{cov}(Z, \tilde{Z}) = \Gamma(\rho, \omega) = \frac{\arcsin^2(\omega) - \arcsin^2(\omega\rho)}{\pi^2}.$$

*Proof.* Let $(\oplus \oplus \ominus\ominus)$ denote the orthant event, $(X > 0, Y > 0, -\tilde{X} > 0, -\tilde{Y} > 0)$, and define events with other combinations of the signs similarly. For a quadrivariate variable with covariance $C(\rho, \omega)$, we define the orthant probability

$$G(\rho, \omega) = \Pr(\oplus \oplus \oplus\oplus),$$

and note that $G(\rho, \omega) = \Pr(\ominus \ominus \ominus\ominus)$, and since $\text{corr}((X, Y, -\tilde{X}, -\tilde{Y})') = C(\rho, -\omega)$ we have $\Pr(\oplus \oplus \ominus\ominus) = \Pr(\ominus \ominus \oplus\oplus) = G(\rho, -\omega)$.

Consider the two terms of $\text{cov}(Z, Z') = \mathbb{E}(Z\tilde{Z}) - \mathbb{E}(Z)\mathbb{E}(\tilde{Z})$. For the first term we have

$$\begin{aligned} \mathbb{E}(Z\tilde{Z}) = \Pr(XY > 0, \tilde{X}\tilde{Y} > 0) &= \Pr((\oplus \oplus \oplus\oplus) \cup (\oplus \oplus \ominus\ominus) \cup (\ominus \ominus \oplus\oplus) \cup (\ominus \ominus \ominus\ominus)) \\ &= 2\left[G(\rho, \omega) + G(\rho, -\omega)\right], \end{aligned}$$

and for the second term we use that $\mathbb{E}(Z) = \mathbb{E}(\tilde{Z})$ can be obtained by setting $\omega = 1$, such that

$$\text{cov}(Z, \tilde{Z}) = 2\left[G(\rho, \omega) + G(\rho, -\omega)\right] - 4\left[G(\rho, 1) + G(\rho, -1)\right]^2.$$

From Chang (1969), we have the expression

$$G(\rho, \omega) = \tfrac{1}{16} + \tfrac{\arcsin\rho + \arcsin\omega + \arcsin(\omega\rho)}{4\pi} + \tfrac{\arcsin^2\rho + \arcsin^2\omega - \arcsin^2(\omega\rho)}{4\pi^2},$$

such that

$$2\left[G(\rho, \omega) + G(\rho, -\omega)\right] = \tfrac{1}{4} + \tfrac{\arcsin\rho}{\pi} + \tfrac{\arcsin^2\rho + \arcsin^2\omega - \arcsin^2(\omega\rho)}{\pi^2},$$

and

$$\begin{aligned} \left(2[G(\rho, 1) + G(\rho, -1)]\right)^2 &= \left(\tfrac{1}{4} + \tfrac{\arcsin\rho}{\pi} + \tfrac{\arcsin^2(1)}{\pi^2}\right)^2 = \left(\tfrac{1}{4} + \tfrac{\arcsin\rho}{\pi} + \tfrac{\pi^2/4}{\pi^2}\right)^2 \\ &= \left(\tfrac{1}{2} + \tfrac{\arcsin\rho}{\pi}\right)^2 = \tfrac{1}{4} + \tfrac{\arcsin\rho}{\pi} + \tfrac{\arcsin^2\rho}{\pi^2}. \end{aligned}$$

Finally, we arrive at

$$\begin{aligned}\operatorname{cov}(Z,\tilde{Z}) &= \frac{1}{4}+\frac{\arcsin(\rho)}{\pi}+\frac{\arcsin^2(\rho)+\arcsin^2(\omega)-\arcsin^2(\omega\rho)}{\pi^2}-\frac{1}{4}-\frac{\arcsin(\rho)}{\pi}-\frac{\arcsin^2(\rho)}{\pi^2}\\ &= \frac{\arcsin^2(\omega)-\arcsin^2(\omega\rho)}{\pi^2}.\end{aligned}$$

□

**Proof of Theorem 1.** Define $Z_{S,j} = 1_{\{\Delta_{\frac{S}{N}} X_{\frac{j}{N}} \Delta_{\frac{S}{N}} Y_{\frac{j}{N}}\}}$ and $\hat{q}_S = \frac{1}{N-S+1}\sum_{j=S}^{N} Z_{S,j}$ and note that $\hat{\tau}_S = 2\hat{q}_S - 1$. Thus, the the asymptotic variance of $Q_S = \sin(\frac{\pi}{2}\hat{\tau}_S)$ is given by that of $\hat{q}_S$ and the delta-method. For the latter we need to multiply the asymptotic variance of $\hat{q}_S$ with the square of:

$$\frac{\partial \sin(\frac{\pi}{2}(2q-1))}{\partial q} = \pi\cos(\tfrac{\pi}{2}(2q-1)) = \pi\cos(\arcsin\rho) = \pi\sqrt{1-\rho^2}.$$

We have

$$\begin{aligned}\operatorname{var}\left(\sum_{j=S}^{N} Z_{S,j}\right) &= \sum_{i=S}^{N}\sum_{j=S}^{N}\operatorname{cov}\left(Z_{S,i},Z_{S,j}\right)\\ &= \sum_{h=-N+1}^{N-1}(N-S+1-|h|)\operatorname{cov}\left(Z_{S,i},Z_{S,i+h}\right)\\ &= \sum_{h=-S+1}^{S-1}(N-S+1-|h|)\tfrac{\arcsin^2(\omega_h)-\arcsin^2(\omega_h\rho)}{\pi^2},\end{aligned}$$

where we have used Lemma A.1 with $\omega = \frac{S-|h|}{S}$ and $\omega = 0$ for $|h| \geq S$. Next, as $N \to \infty$ we have $\frac{n}{N-S+1} \to \frac{1}{S}$ and $\frac{N-S+1-|h|}{N-S+1} \to 1$ for all $|h| \leq S$, such that

$$\operatorname{var}\left(\sqrt{n}\hat{q}_S\right) = \operatorname{var}\left(\sqrt{n}\frac{1}{N-S+1}\sum_{j=S}^{N} Z_{S,j}\right) \to \frac{1}{S}\sum_{h=-S+1}^{S-1}\tfrac{\arcsin^2(\omega_h)-\arcsin^2(\omega_h\rho)}{\pi^2}.$$

The asymptotic behavior of the sum, as $S \to \infty$, can be inferred from

$$\begin{aligned}I(\rho) = \int_0^1 \arcsin^2(x\rho)\mathrm{d}x &= \left[\tfrac{2\sqrt{1-\rho^2x^2}\arcsin(\rho x)}{\rho} + x\arcsin^2(\rho x) - 2x\right]_0^1\\ &= \left(2\tfrac{\sqrt{1-\rho^2}}{\rho}\arcsin(\rho) + \arcsin^2(\rho) - 2\right) - 0,\end{aligned}$$

such that avar $(\hat{q}_S)$ is $2/\pi^2$ times

$$\begin{aligned}\int_0^1 \arcsin^2(x) - \arcsin^2(x\rho)\mathrm{d}x &= I(1) - I(\rho) \\ &= 0\arcsin(1) + \arcsin^2(1) - 2 - \tfrac{2\sqrt{1-\rho^2}\arcsin(\rho)}{\rho} - \arcsin^2(\rho) + 2 \\ &= (\tfrac{\pi}{2})^2 - 2\tfrac{\sqrt{1-\rho^2}}{\rho}\arcsin(\rho) - \arcsin^2(\rho).\end{aligned}$$

Thus, multiplying by $2/\pi^2$ and applying the delta-method we have

$$\begin{aligned}\mathrm{var}(\sqrt{n}\hat{Q}_S) &\to \pi^2(1-\rho^2)\frac{2}{\pi^2}\left[(\tfrac{\pi}{2})^2 - 2\tfrac{\sqrt{1-\rho^2}}{\rho}\arcsin(\rho) - \arcsin^2(\rho)\right] \\ &= (1-\rho^2)2\left[\arcsin^2(1) - 2\tfrac{\sqrt{1-\rho^2}}{\rho}\arcsin(\rho) - \arcsin^2(\rho)\right].\end{aligned}$$

□

**Proof of Theorem 2.** In the proof we rely on the local-constancy approximation by Mykland and Zhang (2009). Their assumptions 1 and 2 are satisfied by our equidistant sampling and $\sigma(u)$ being bounded away from zero. Under the approximating measure we have that $(\Delta_{\frac{S}{N}} X_{\frac{j}{N}}, \Delta_{\frac{S}{N}} Y_{\frac{j}{N}}) = (\sigma_{x,j} Z_{x,j}, \sigma_{y,j} Z_{y,j})$ where $\sigma_{x,j} = \sigma_x(\frac{j-S}{N})$ and $\sigma_{y,j} = \sigma_y(\frac{j-S}{N})$, and $(Z_{x,j}, Z_{y,j}) \sim N_2(0, C)$ where $C$ is a correlation matrix with correlation coefficient equal to $\rho$ . Hence, under the approximate measure we have

$$\mathbb{E}[\mathrm{sgn}(\Delta_{\frac{S}{N}} X_{\frac{j}{N}} \Delta_{\frac{S}{N}} Y_{\frac{j}{N}})] = \tau = \tfrac{2}{\pi}\arcsin\rho,$$

and the statistic

$$T_{S,s} = \frac{1}{\lfloor (N-s+1)/S \rfloor} \sum_{k=1}^{\lfloor (N-s+1)/S \rfloor} \mathrm{sgn}(\Delta_{\frac{S}{N}} X_{\frac{kS+s-1}{N}} \Delta_{\frac{S}{N}} Y_{\frac{kS+s-1}{N}}),$$

is based on non-overlapping returns for each $s = 1, \ldots, S$, such that $T_{S,s} \xrightarrow{p} \tau$ as $N/S \to \infty$ by the standard law of large numbers. Since $Q_S = \sin(\frac{\pi}{2}T)$ where $T$ is a (nearly evenly) weighted average of $T_{S,1}, \ldots, T_{S,S}$ it also follows that $Q_S$ is consistent under the local-constancy approximation measure, and since consistency is not affected by the change of measure, $Q_S$ is consistent for $\rho$.

For $P$ we simply note that $\sum_{k=1}^{\lfloor \delta^{-1} \rfloor} \Delta_\delta X_{k\delta} \Delta_\delta Y_{\delta k} - \rho \int_0^1 \sigma_x(u)\sigma_y(u)\mathrm{d}u = o_p(1)$ with a similar result for the denominator, such that $P - \rho\lambda = o_p(1)$ where $|\lambda| \leq 1$ is given in the theorem. For $K$ we sketch

the proof, using the same local-constancy approximation method as above. We have

$$\begin{aligned}\operatorname{cov}(\Delta_\delta X_{i\delta} - \Delta_\delta X_{j\delta}, \Delta_\delta Y_{i\delta} - \Delta_\delta Y_{j\delta}) &= \operatorname{cov}(\sigma_{x,i}Z_{x,i} - \sigma_{x,j}Z_{x,j}, \sigma_{y,i}Z_{y,i} - \sigma_{y,j}Z_{y,j}) \\ &= \operatorname{cov}(\sigma_{x,i}Z_{x,i}, \sigma_{y,i}Z_{y,i}) + \operatorname{cor}(\sigma_{x,j}Z_{x,j}, \sigma_{y,j}Z_{y,j}) \\ &= (\sigma_{x,i}\sigma_{y,i} + \sigma_{x,j}\sigma_{y,j})\rho,\end{aligned}$$

such that

$$\mathbb{E}[\operatorname{sgn}([\Delta_\delta X_{i\delta} - \Delta_\delta X_{j\delta}][\Delta_\delta Y_{i\delta} - \Delta_\delta Y_{j\delta}])] = \tfrac{2}{\pi}\arcsin\left(\rho\frac{\sigma_{x,i}\sigma_{y,i} + \sigma_{x,j}\sigma_{y,j}}{\sqrt{\sigma_{x,i}^2 + \sigma_{x,j}^2}\sqrt{\sigma_{y,i}^2 + \sigma_{y,j}^2}}\right).$$

Now define

$$h_\delta(u,v) = \frac{\sigma_{x,\lceil u/\delta\rceil}\sigma_{y,\lceil u/\delta\rceil} + \sigma_{x,\lceil v/\delta\rceil}\sigma_{y,\lceil v/\delta\rceil}}{\sqrt{\sigma_{x,\lceil u/\delta\rceil}^2 + \sigma_{x,\lceil v/\delta\rceil}^2}\sqrt{\sigma_{y,\lceil u/\delta\rceil}^2 + \sigma_{y,\lceil v/\delta\rceil}^2}},$$

which converges uniformly to

$$h(u,v) = \frac{\sigma_x(u)\sigma_y(u) + \sigma_x(v)\sigma_y(v)}{\sqrt{\sigma_x^2(u) + \sigma_x^2(v)}\sqrt{\sigma_y^2(u) + \sigma_y^2(v)}},$$

for $(u,v) \in (0,1] \times (0,1]$. The Kendall estimator

$$\hat{\tau}_K = \frac{1}{\lfloor 1/\delta\rfloor^2}\sum_{i,j=1}^{\lfloor 1/\delta\rfloor}\operatorname{sgn}([\Delta_\delta X_{i\delta} - \Delta_\delta X_{j\delta}][\Delta_\delta Y_{i\delta} - \Delta_\delta Y_{j\delta}]) \xrightarrow{p} \tfrac{2}{\pi}\int_0^1\int_0^1 \arcsin[\rho h(u,v)]\mathrm{d}u\mathrm{d}v.$$

For the Spearman and Gaussian rank correlation estimators their inconsistency follows from a simple example. First note that under constant volatility these estimators will be consistent. Now suppose $\rho$ is large, $\rho = 0.8$ say, but the volatility process are time-varying such that $(\sigma_x(u), \sigma_y(u)) = (10, 1)$ for $u < \frac{1}{2}$ and $(\sigma_x(u), \sigma_y(u)) = (1, 10)$ for $u \geq 0.5$. Then the extreme ranks for $\Delta X$ will be concentrated in the first half of the data where as those $\Delta Y$ will be concentrated on the second half of the observations. This implies a low rank correlation, such that the estimators will be inconsistent and biased towards zero in this example.

□

# References


Ait-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**, 1504–1517.

Andersen, T., Dobrev, D. and Schaumburg, E. (2012), 'Jump-robust volatility estimation using nearest neighbor truncation', *Journal of Econometrics* **169**, 75–93.

Andersen, T. G. and Bollerslev, T. (1998*a*), 'Answering the skeptics: Yes, standard volatility models do provide accurate forecasts', *International Economic Review* **39**, 885–905.

Andersen, T. G. and Bollerslev, T. (1998*b*), 'Deutsche Mark-Dollar Volatility: Intraday Activity Patterns, Macroeconomic Announcements, and Longer Run Dependencies', *Journal of Finance* **53**, 219–265.

Andersen, T. G., Bollerslev, T., Diebold, F. X. and Labys, P. (2000), 'Great realizations', *Risk* **13**(3), 105–108.

Andersen, T. G., Bollerslev, T., Diebold, F. X. and Wu, J. (2005), 'A framework for exploring the macroeconomic determinants of systematic risk', *American Economic Review* **95**, 398–404.

Andersen, T. G., Bollerslev, T., Diebold, F. X. and Wu, J. (2006), Realized beta: Persistence and predictability, *in* 'Advances in Econometrics: Econometric Analysis of Economic and Financial Time Series in Honor of R.F. Engle, and C.W.J. Granger', pp. 1–39.

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

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

Bandi, F. M. and Russell, J. R. (2008), 'Microstructure Noise, Realized Variance, and Optimal Sampling', *Review of Economic Studies* **75**, 339–69.

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A. and Shephard, N. (2008), 'Designing realised kernels to measure the ex-post variation of equity prices in the presence of noise', *Econometrica* **76**, 1481–536.

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

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A. and Shephard, N. (2011*b*), 'Subsampling realised kernels', *Journal of Econometrics* **160**, 204–219.

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**, 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**, 1–48.

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**, 27–41.

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

Blomqvist, N. (1950), 'On a measure of dependence between two random variables', *The Annals of Mathematical Statistics* **21**, 593–600.

Boudt, K., Cornelissen, J. and Croux, C. (2012*a*), 'The Gaussian rank correlation estimator: robustness properties', *Statistics and Computing* **22**, 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**, 2993–3005.

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

Chang, M. C. (1969), 'The orthant probabilities of four gaussian variates', *Annals of Mathematical Statistics* **40**, 152–161.

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**, 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.

Croux, C. and Dehon, C. (2010), 'Influence functions of the spearman and kendall correlation measures', *Statistical methods & applications* **19**, 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* **3**, 1–28.

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

Dovonon, P., Gonçalves, S. and Meddahi, N. (2013), 'Bootstrapping realized multivariate volatility measures', *Journal of Econometrics* **172**, 49–65.

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

Esscher, F. (1924), 'On a method of determining correlation from the ranks of the variates', *Scandinavian Actuarial Journal* **1924**, 201–219.

Greiner, R. (1909), 'Über das fehlersystem kollektivmaßlehre', *Zeitschrift für Mathematik und Physik* **57**, 121–158,225–260,337–373.

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**, 58–68.

Hansen, P. R. (2015), 'A martingale decomposition of discrete Markov chains', *Economics Letters* **133**, 14–18.

Hansen, P. R., Horel, G., Lunde, A. and Archakov, I. (2016), 'A Markov Chain Estimator of Multivariate Volatility from High Frequency Data', *The Fascination of Probability, Statistics, and their Applications. In Honour of Ole E. Barndorff-Nielsen* .

Hansen, P. R. and Lunde, A. (2006), 'Realized Variance and Market Microstructure Noise', *Journal of Business and Economic Statistics* **24**, 127–161.

Hansen, P. R., Lunde, A. and Voev, V. (2014), 'Realized beta GARCH: A multivariate GARCH model with realized measures of volatility', *Journal of Applied Econometrics* **29**, 774–799.

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

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

Horel, G. (2007), Estimating Integrated Volatility with Markov Chains, PhD thesis, Stanford University.

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**, 2249–2276.

Kalnina, I. and Xiu, D. (2017), 'Nonparametric estimation of the leverage effect: A trade-off between robustness and efficiency', *Journal of the American Statistical Association* **112**, 384–396.

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

Kendall, M. G. (1949), 'Rank and product-moment correlation', *Biometrika* **36**, 177–193.

Kruskal, W. H. (1958), 'Ordinal measures of association', *Journal of the American Statistical Association* **53**, 814–861.

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

Li, Y. and Mykland, P. A. (2015), 'Rounding errors and volatility estimation', *Journal of Financial Econometrics* **13**, 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**, 187–222.

Liu, L. Y., Patton, A. J. and Sheppard, K. (2015), 'Does anything beat 5-minute RV? A comparison of realized measures across multiple asset classes', *Journal of Econometrics* **187**, 293–311.

Liu, R. Y. and Singh, K. (1992), Moving blocks jackknife and bootstrap capture weak dependence., *in* R. LePage and L. Billard, eds, 'Exploring the Limits of Bootstrap', John Wiley, New York, pp. 225–248.

Malliavin, P. and Mancino, M. (2002), 'Fourier series method for measurement of multivariate volatilities', *Finance and Stochastics* **6**, 49–56.

Mancini, C. (2001), 'Disentangling the jumps of the diffusion in a geometric jumping Brownian motion', *Giornale dell'Istituto Italiano degli Attuari* **64**, 19–47.

Mancini, C. (2009), 'Non-parametric threshold estimation for models with stochastic diffusion coefficient and jumps', *Scandinavian Journal of Statistics* **36**, 270–296.

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

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**, 1231–1246.

Mykland, P. A. and Zhang, L. (2009), 'Inference for continuous semimartingales observed at high frequency: A general approach', *Econometrica* **77**, 1403–1445.

Politis, D. N., Romano, J. P. and Wolf, M. (1999), *Subsampling*, Springer, New York.

Precup, O. V. and Iori, G. (2007), 'Cross-correlation measures in the high-frequency domain', *European Journal of Finance* **13**, 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**, 360–385.

Raymaekers, J. and Rousseeuw, P. J. (2021), 'Fast robust correlation for high-dimensional data', *Technometrics* **63**, 184–198.

Reiß, M., Todorov, V. and Tauchen, G. (2015), 'Nonparametric test for a constant beta between itô semi-martingales based on high-frequency data', *Stochastic Processes and their Applications* **125**, 2955–2988.

Renò, R. (2003), 'A closer look at the epps effect', *International Journal of Theoretical and Applied Finance* **6**, 87–102.

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

Rousseeuw, P. J. (1984), 'Least median of squares regression', *Journal of the American statistical association* **79**, 871–880.

Sheppard, W. F. (1899), 'On the application of the theory of error to cases of normal distribution and normal correlation', *Philosophical Transactions of the Royal Society of London. Series A* **192**, 101–167.

Todorov, V. and Bollerslev, T. (2010), 'Jumps and betas: A new framework for disentangling and estimating systematic risks', *Journal of Econometrics* **157**, 220–235.

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**, 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**, 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**, 68–104.

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

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

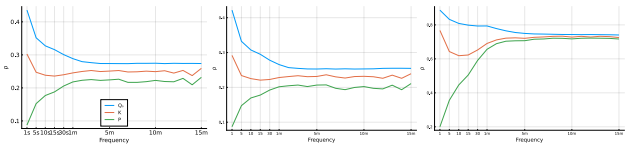
Zhou, B. (1996), 'High-frequency data and volatility in foreign exchange rates', *Journal of Business and Economic Statistics* **14**, 45–52.

Zhou, B. (1998), Parametric and nonparametric volatility measurement, *in* C. L. Dunis and B. Zhou, eds, 'Nonlinear Modelling of High Frequency Financial Time Series', John Wiley Sons Ltd, chapter 6, pp. 109–123.

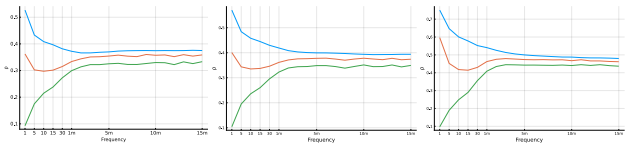
# Supplementary Material

## S.1 Supplementary Empirical Results

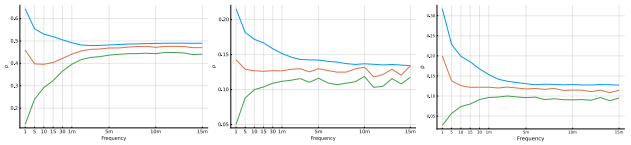
We presented correlation signature plots for some selected pair of assets in Figure 7. Here we present the complete set of signature plots for all assets in the Small Universe, see Figure S.1.



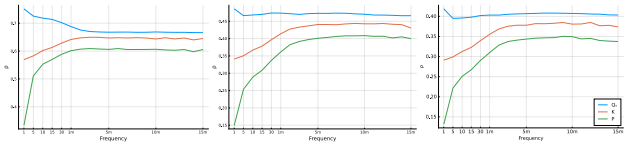
(a) corr(SPY,D), corr(SPY,DUK), corr(D,DUK)



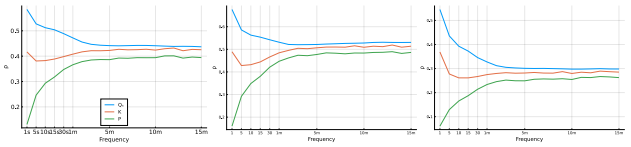
(b) corr(SPY,AMT), corr(SPY,PLD), corr(AMT,PLD)



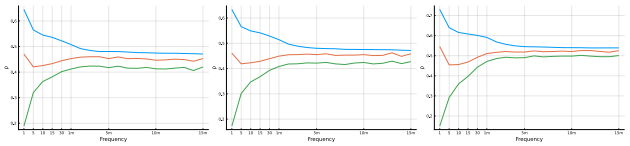
(c) corr(SPY,LYB), corr(SPY,NEM), corr(LYB,NEM)



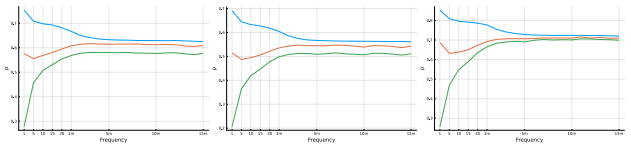
(d) corr(SPY,AAPL), corr(SPY,AMD), corr(AAPL,AMD)



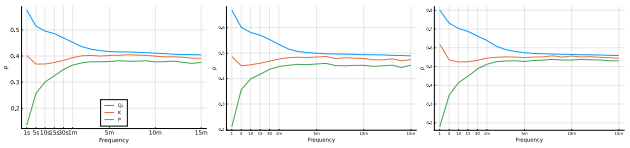
(e) corr(SPY,AAL), corr(SPY,UNP), corr(AAL,UNP)



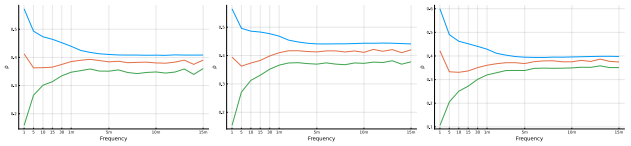
(f) corr(SPY,JNJ), corr(SPY,MRK), corr(JNJ,MRK)



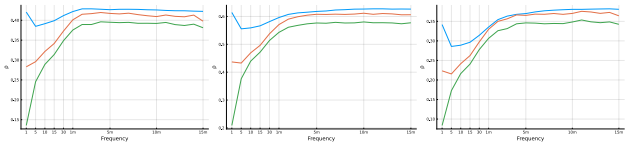
(g) corr(SPY,JPM), corr(SPY,WFC), corr(JPM,WFC)



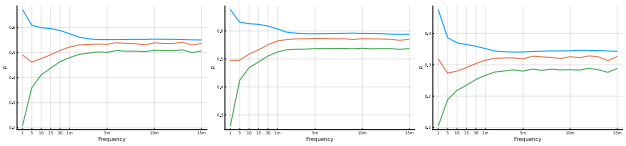
(h) corr(SPY,HAL), corr(SPY,XOM), corr(HAL,XOM)



(i) corr(SPY,PG), corr(SPY,WMT), corr(PG,WMT)



(j) corr(SPY,TSLA), corr(SPY,AMZN), corr(TSLA,AMZN)



(k) corr(SPY,DIS), corr(SPY,FB), corr(DIS,FB)

Figure S.1: Correlations signature plots for assets in Small Universe.

In Figure S.2 we present the range of correlation estimates for the pairs of assets within the same section. This figure is analogous to Figure 8, where we reported correlations between each asset and

SPY.

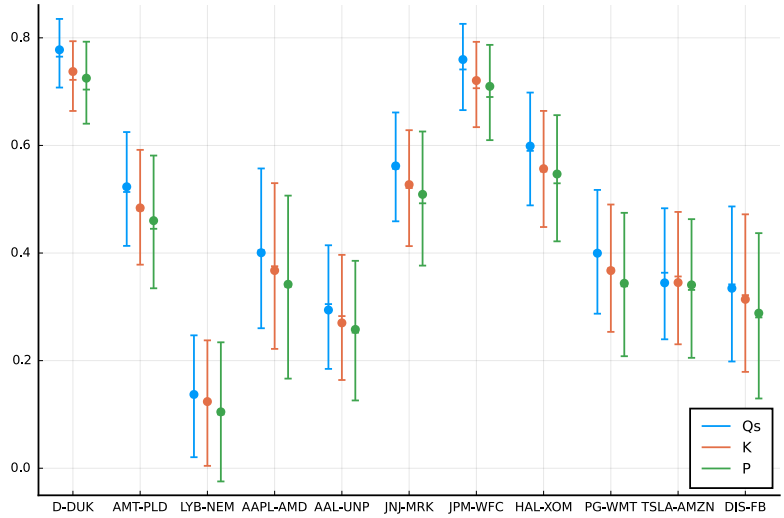
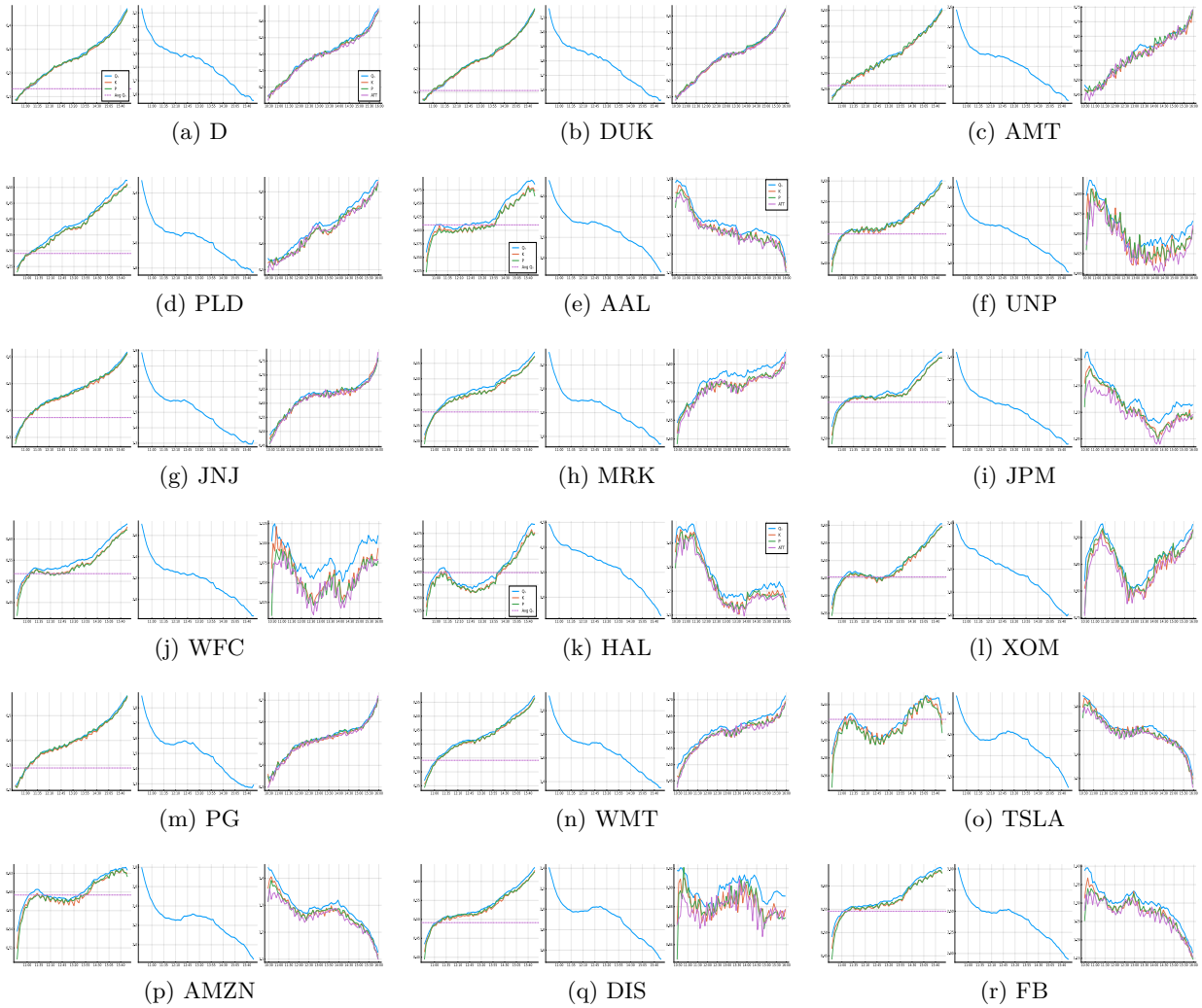



Figure S.2: Daily correlations between pair of assets in each of the 11 sectors. The three estimators were applied to 1,763 trading days. The average estimate (dash) and median estimate (bullet) are shown. The vertical lines present the interquartile range over the 1,763 daily estimates.

Rolling window estimates of correlations, relative volatilities, and market betas are presented for all assets in the Small Universe in Figure S.3.



(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 S.3: Average intraday correlations (left), relative volatility (middle), and market $\beta$ (right) for 18 stocks.

The correlation increases for all assets over the trading day, and the relative volatility decreases. The effect on the market beta (the product of the two) is therefore determined by which of the two increases/decreases the most.

The correlation signature plots are based on averages over many trading days. It is therefore important to explore if the findings are robust to choice of sample period. Correlation signature plots for each sub-period are shown in Figure S.4 for each stock in the Small Universe and the SPY, and for each pair of assets within the same sector in Figure S.5.

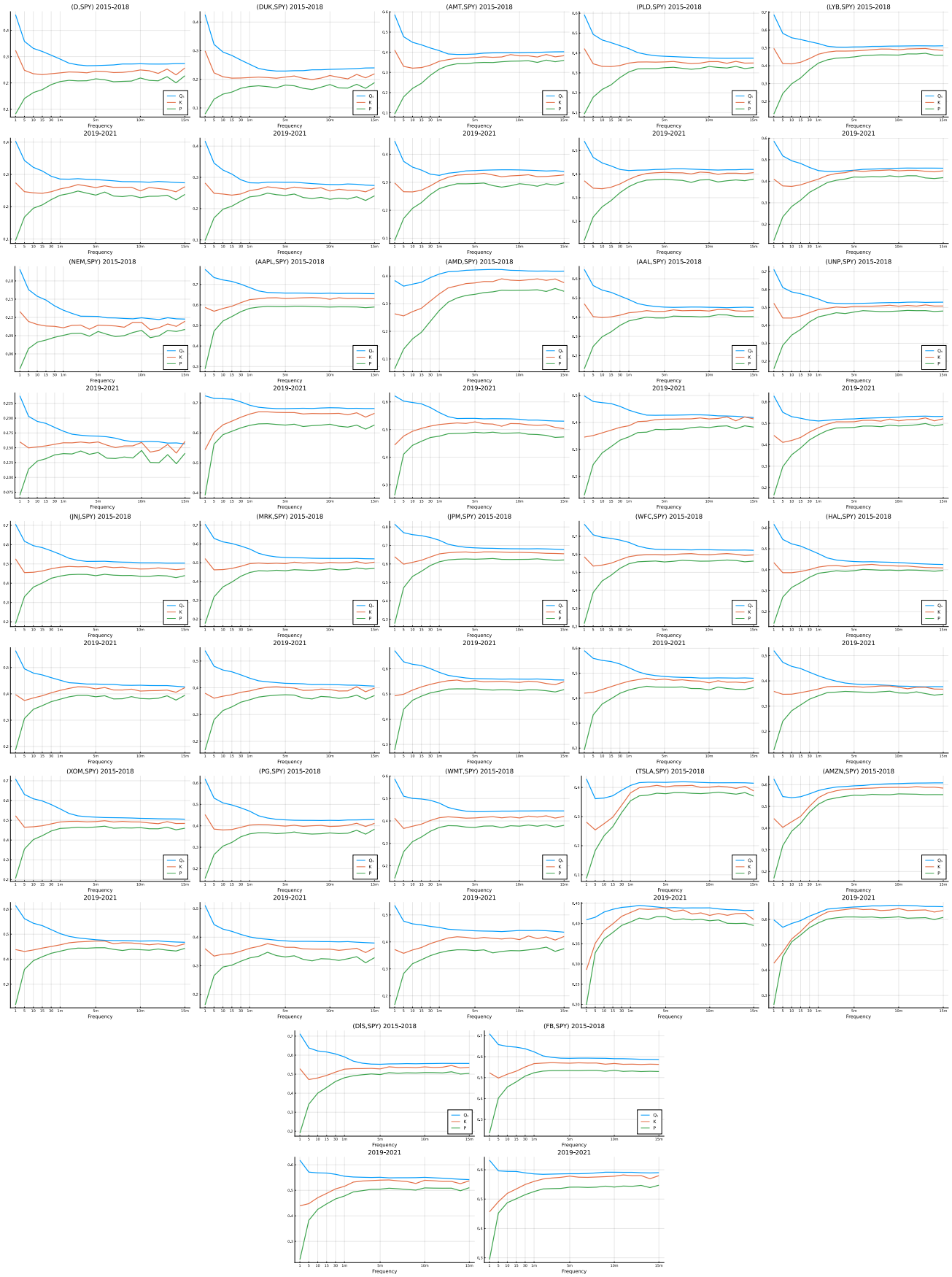



Figure S.4: Correlation signature plots for 22 assets and SPY.

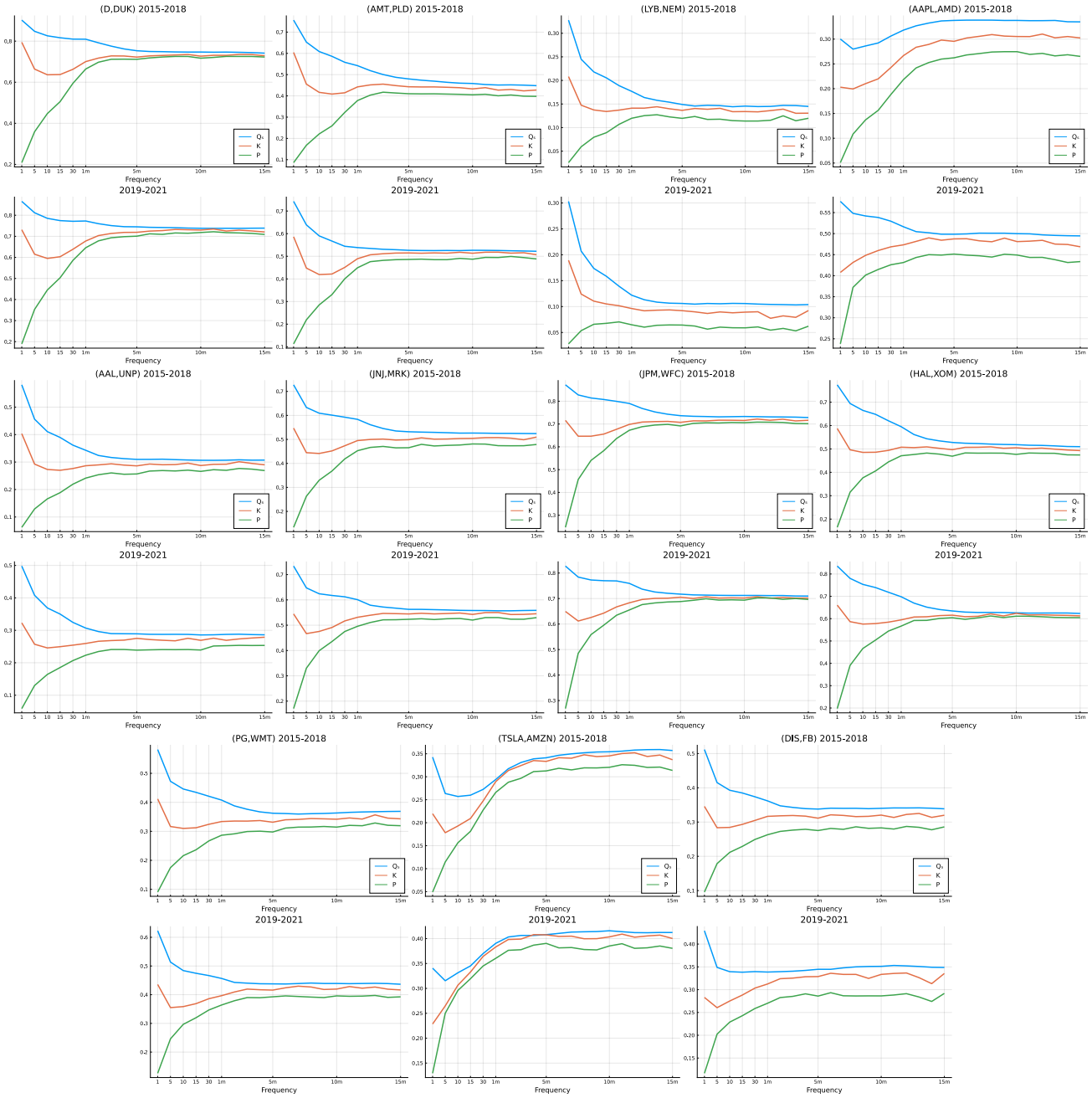



Figure S.5: Correlation signature plots for 11 pairs of stocks.

The average level of correlation can obviously be different in the two sample periods, but the shapes of the signature plots are nevertheless remarkably similar for the two sub-period for all pairs of assets, see Figures S.4 and S.5.

Similarly it is important to investigate if the patterns we observed in intraday correlations, relative volatilities, and market betas are robust features or merely specific to the sample period. We would expect these results to be robust, because ATT reported the same results (for market betas) and they

used a different sample period. To explore this a bit further, we split the sample in two and estimated intraday correlations, relative volatilities, and market betas for each sub-period and each assets in the Small Universe.

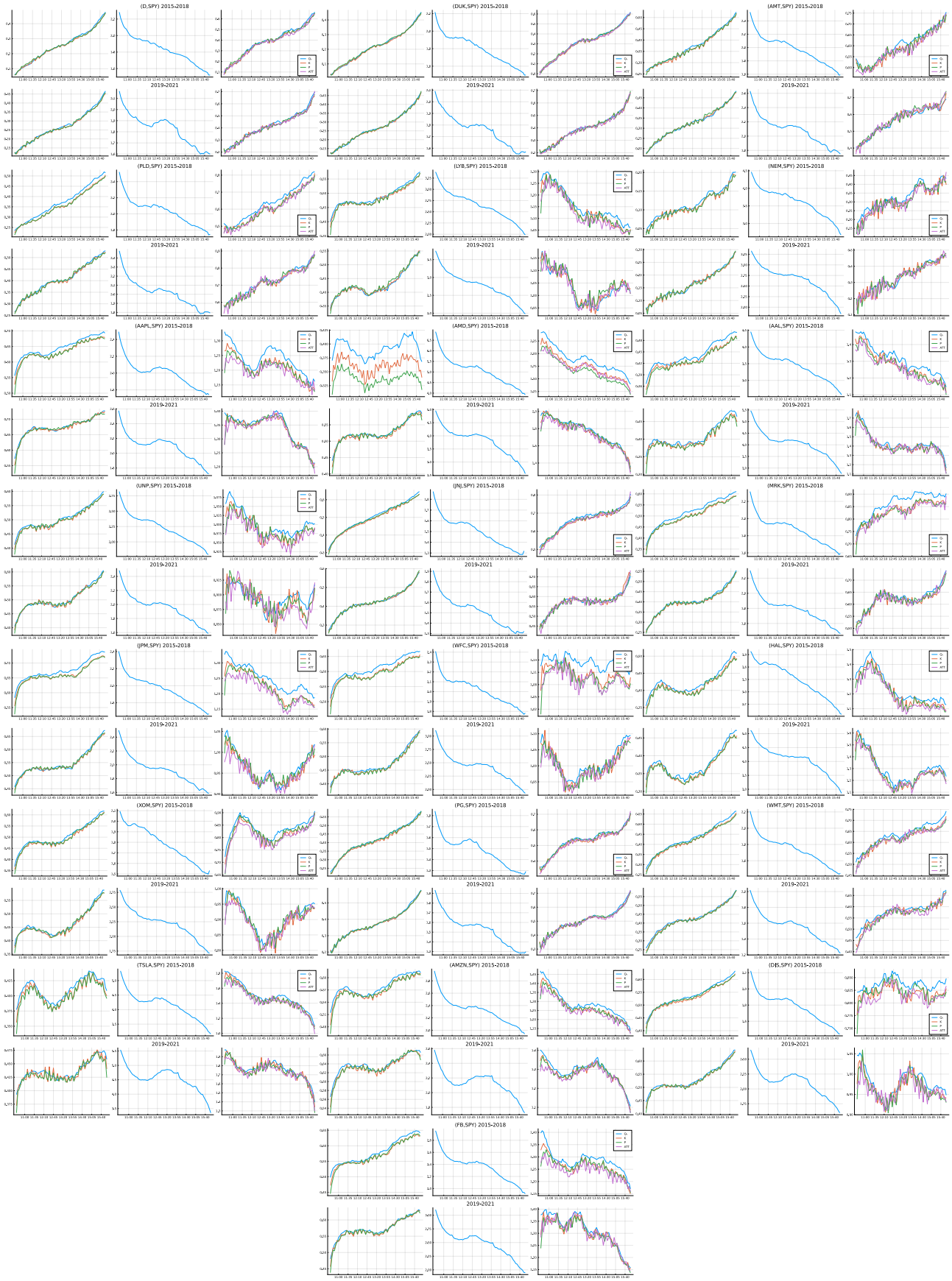



Figure S.6: Intraday estimates for 22 assets for two sample periods.

The results are show in Figure S.6 and the results for the two sample periods are again very similar.

Additional results for the Large Universe of assets are reported in Figures S.7 and S.8. We consider intraday changes in market correlations, relative volatility, and intraday betas, as defined by the difference between the estimate from the first hour of trading and the estimate from the last hour of trading. The changes are measured for the average estimates over the days in the sample period, and the increments are plotted against low-frequency market betas, size, and book-to-market.

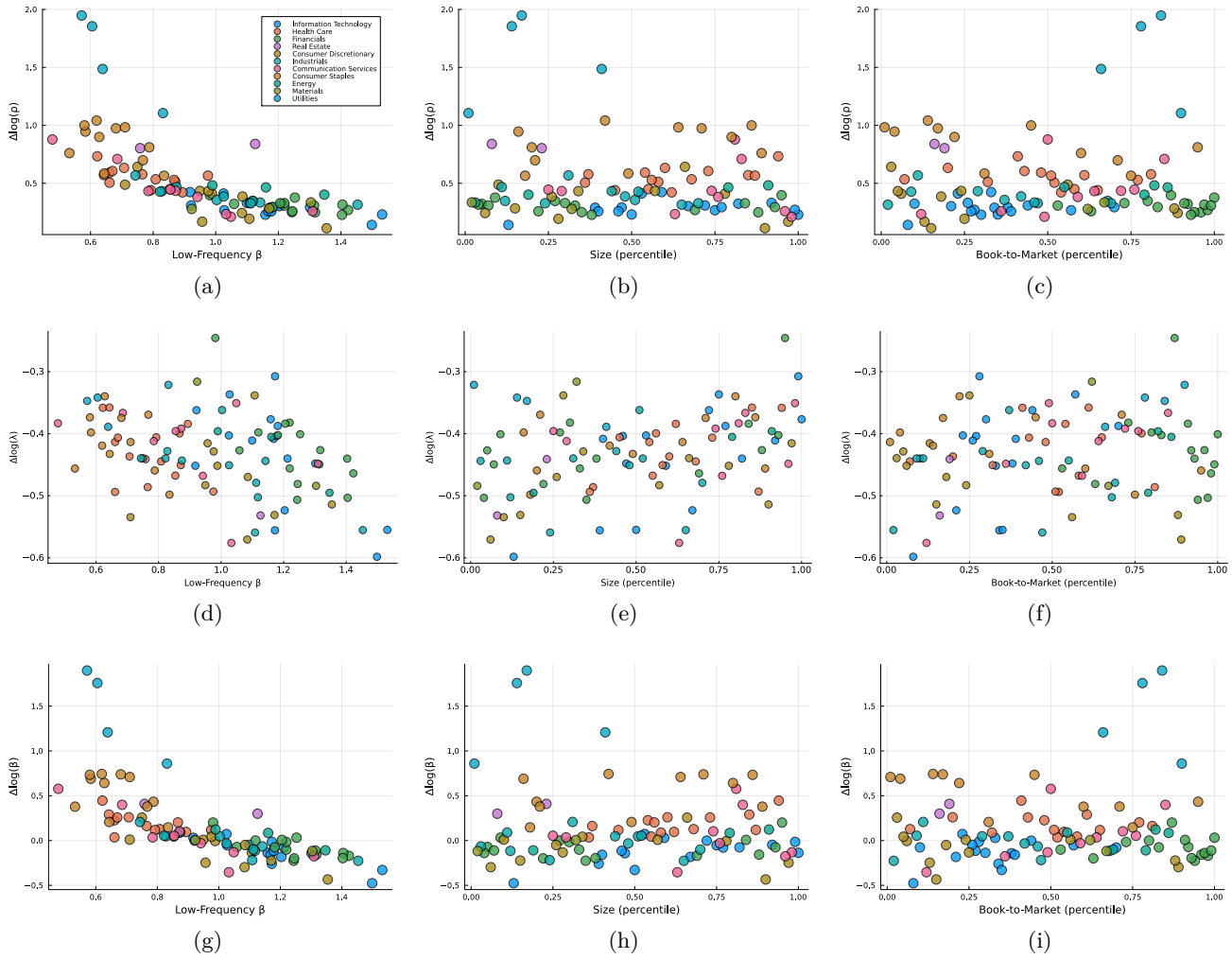



Figure S.7: The average changes in market correlations and market beta, from the first hour to the last hour of the active trading day are plotted against three variables. In the left panels the quantities are plotted against the (low frequency) market beta, which is computed from daily returns. In the middle panels they are plotted against (percentiles of) market capitalization size, and in the right panels they are plotted against the (percentiles of) book-to-market values.

The low-frequency market beta is the conventional estimate, which is based on daily returns in our sample period. We have also sorted assets by "Size" and "Book-to-Market" where the former is the

market value (market cap) of the company and book-to-market is defined by the company's book value relative to its market value.

Figure S.8 plots the intraday correlations and market betas, estimated for the first and last hour of the trading day, against the low-frequency beta and the percentiles for Size and Book-to-Market. The association is clearly strongest with the low-frequency betas in the left panels.

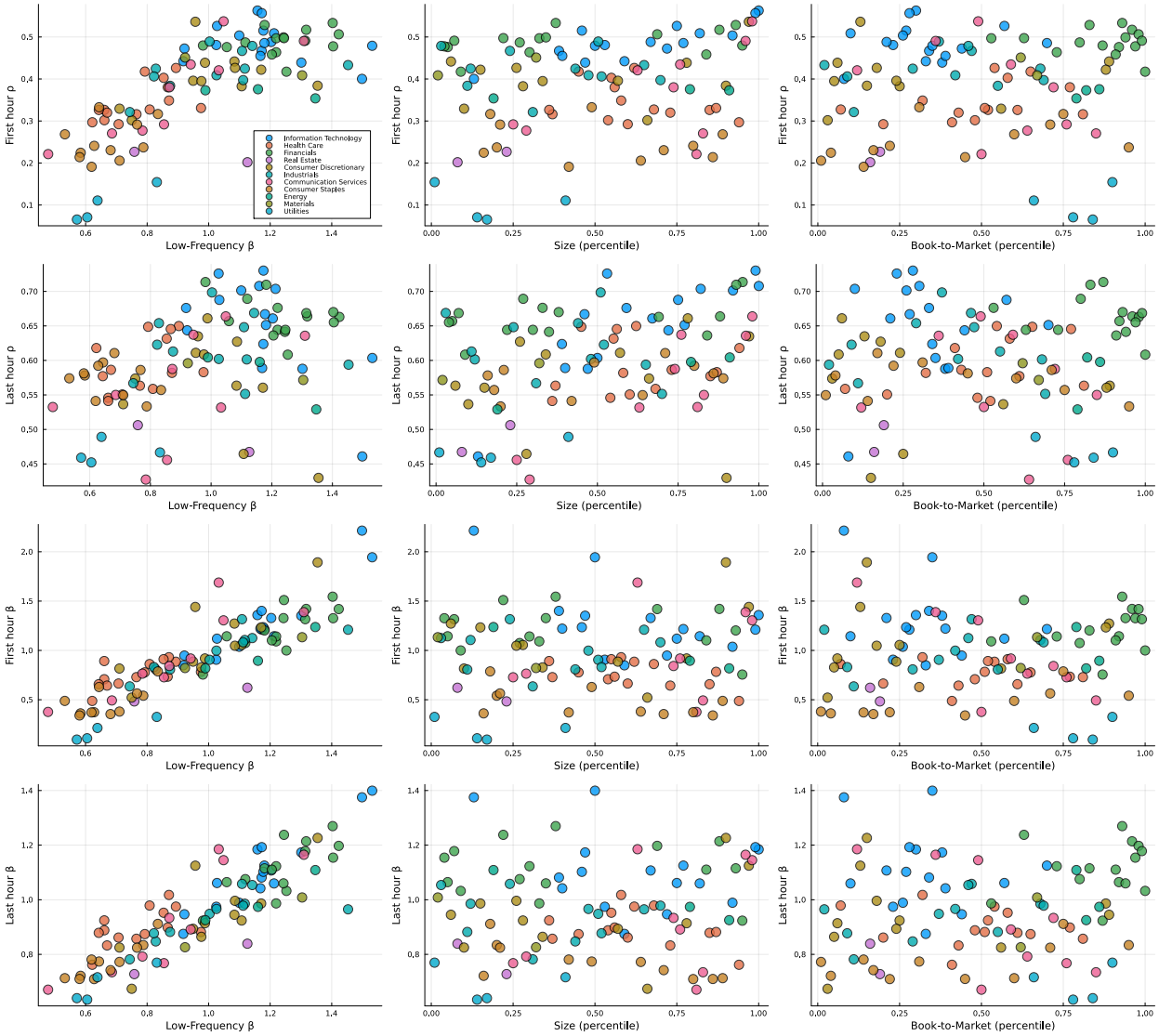



Figure S.8: The average market correlations and market beta for the first and last hours of the trading day are plotted against three variables. In the left panels the quantities are plotted against the (low frequency) market beta, which is computed from daily returns. In the middle panels they are plotted against (percentiles of) market capitalization size, and in the right panels they are plotted against the (percentiles of) book-to-market values.

## S.2 Supplementary Theoretical Results

For high-frequency financial data, estimators are typically applied to sparsely sampled prices, and we consider the influence functions of the estimators in this context. We derive the influence functions of these estimators for the sparse-sampled observations.

**Proposition S.1.** *The influence function of Pearson estimator at $\Phi_\rho$ is given by*

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

*The influence function of Kendall estimator at $\Phi_\rho$ is given by*

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

*The influence function of (subsampled) Quadrant estimator at $\Phi_\rho$ is given by*

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

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

**Proof of Proposition S.1.** 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 $\varepsilon$ of they are exactly at point $(x_0, y_0)$. Define a function $G$ in terms of $\varepsilon$ 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

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

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

$$\begin{aligned}
F_K(h_1,h_2) &= \mathrm{Prob}[(\sum_{i=1}^{S-h_1}X_i - h_1x_0 - \sum_{i=1}^{S-h_2}X_{S+i} + h_2x_0)(\sum_{i=1}^{S-h_1}Y_i - h_1y_0 - \sum_{i=1}^{S-h_2}Y_{S+i} + h_2y_0) > 0] \\
&= \mathrm{Prob}[(\sum_{i=1}^{S-h_1}X_i - \sum_{i=1}^{S-h_2}X_{S+i} - (h_1-h_2)x_0)(\sum_{i=1}^{S-h_1}Y_i - \sum_{i=1}^{S-h_2}Y_{S+i} - (h_1-h_2)y_0) > 0] \\
&= 2\Phi(\frac{(h_1-h_2)x_0}{\sqrt{2S-h_1-h_2}}, \frac{(h_1-h_2)y_0}{\sqrt{2S-h_1-h_2}}) - \Phi(\frac{(h_1-h_2)x_0}{\sqrt{2S-h_1-h_2}}) - \Phi(\frac{(h_1-h_2)y_0}{\sqrt{2S-h_1-h_2}}) + 1
\end{aligned}$$

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

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

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

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

with

$$F_Q(h) = \text{Prob}[(\sum_{i=1}^{S-h} X_i + hx_0)(\sum_{i=1}^{S-h} Y_i + hy_0)]$$
$$= 2\Phi(\frac{hx_0}{\sqrt{S-h}}, \frac{hy_0}{\sqrt{S-h}}) - \Phi(\frac{hx_0}{\sqrt{S-h}}) - \Phi(\frac{hy_0}{\sqrt{S-h}}) + 1$$

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

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

□

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.

### S.2.1 Decomposition

$$\begin{aligned}\bar{\beta}(\tfrac{i}{N}) &= \bar{\rho}(\tfrac{i}{N})\bar{\lambda}(\tfrac{i}{N}) + \text{cov}(\rho(\tfrac{i}{N}), \lambda(\tfrac{i}{N})) \\ &= \bar{\rho}(\tfrac{i}{N})\bar{\lambda}(\tfrac{i}{N})\left[1 + \frac{\text{cov}(\rho(\frac{i}{N}), \lambda(\frac{i}{N}))}{\bar{\rho}(\frac{i}{N})\bar{\lambda}(\frac{i}{N})}\right] = \bar{\rho}(\tfrac{i}{N})\bar{\lambda}(\tfrac{i}{N})\bar{\gamma}(\tfrac{i}{N}).\end{aligned}$$

$$\log\bar{\beta}(\tfrac{i}{N}) = \log\bar{\rho}(\tfrac{i}{N}) + \log\bar{\lambda}(\tfrac{i}{N}) + \log\bar{\gamma}(\tfrac{i}{N})$$

First and last hour betas for asset $j$

$$\Delta\beta_j = \log\frac{\bar{\beta}_j(close)}{\bar{\beta}_j(open)} = \log\bar{\beta}_j(close) - \log\bar{\beta}_j(open).$$

Similarly

$$\Delta\rho_j = \log\bar{\rho}_j(close) - \log\bar{\rho}_j(open)$$

Scatterplots of $\Delta\beta_j$ against $\Delta\rho_j$, $\Delta\lambda_j$, and $\Delta\gamma_j$.

## S.3 Additional Simulation Results

Figure S.9 presents the results for the Heston model without noise for two different levels of the correlation: $\rho = 0.5$ and $\rho = 0.75$. The analogous results for levels 0.25 and 0.66 are in Figure 4.

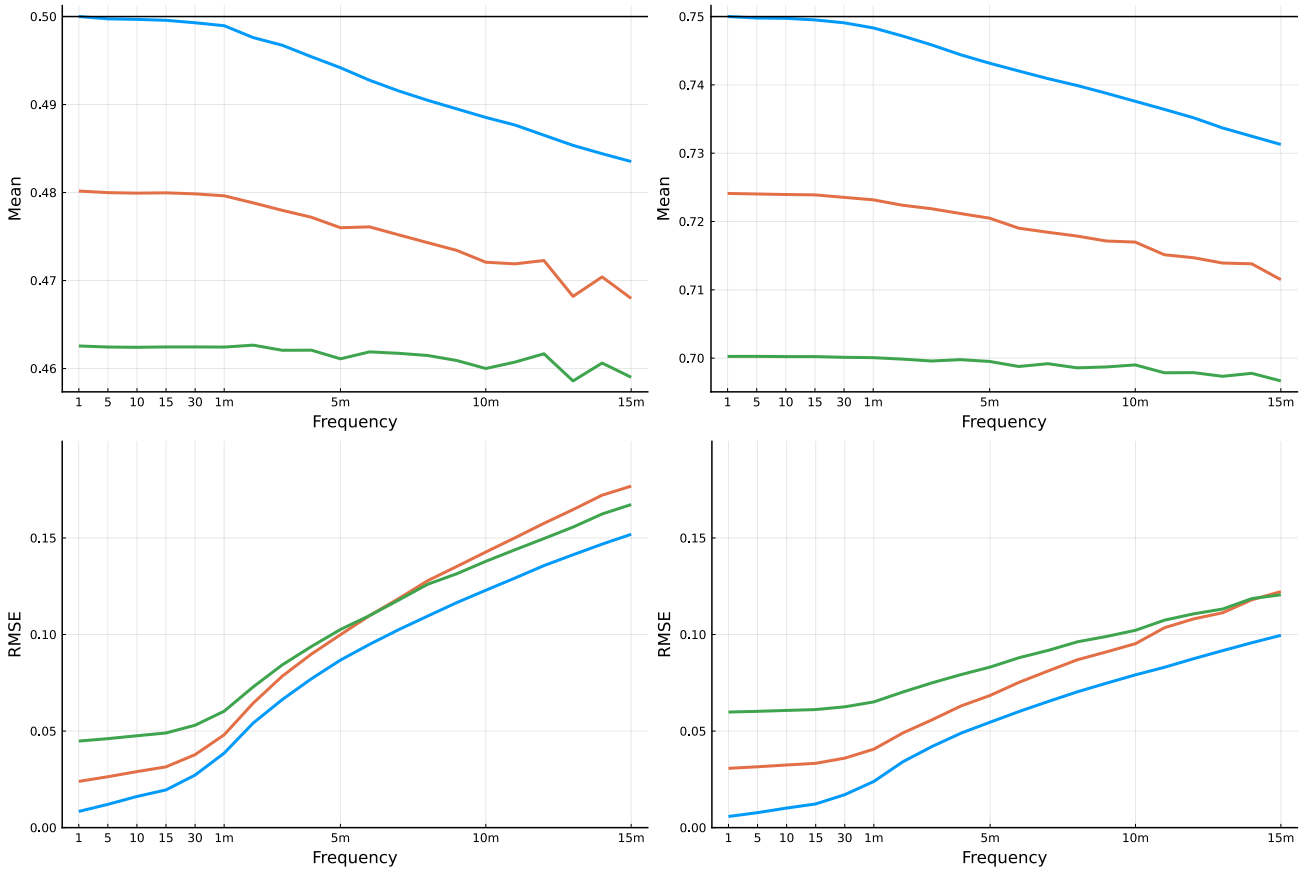



Figure S.9: Means and RMSEs for the estimators as a function of sampling frequencies. Prices are generated by the Heston model with the true correlation being $\rho = 0.50$ (left panels) and $\rho = 0.75$ (right panels).

### S.3.1 Jumps

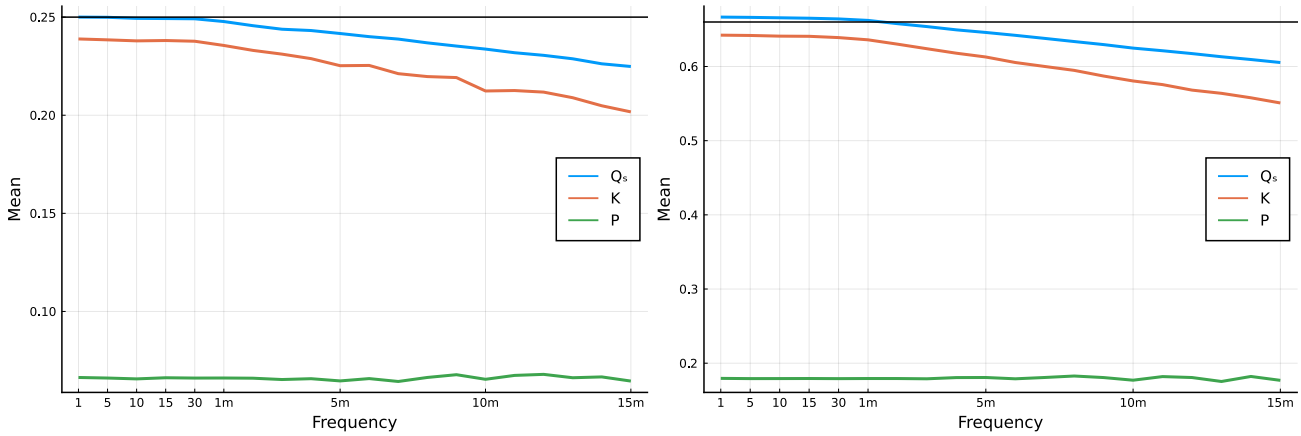
Jumps are prevalent in high-frequency prices. We consider the following jump processes,

$$X_t = X_t^* + \sum_{s \le t} J_s^x \qquad Y_t = Y_t^* + \sum_{s \le t} J_s^y,$$

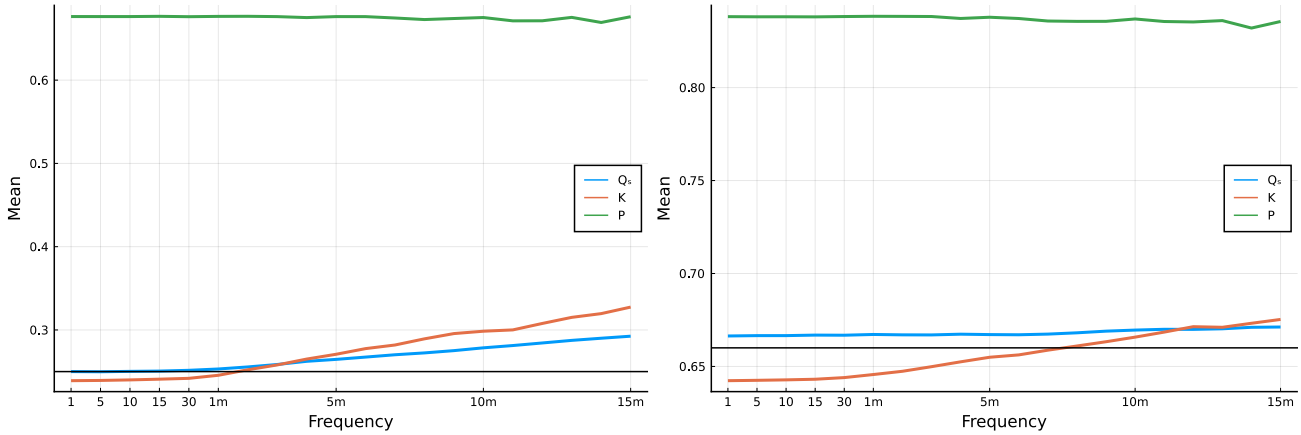
where $J_t^x$ and $J_t^y$ are Poisson jump processes with intensities $\lambda_x$ and $\lambda_y$, respectively, and a jump size that is uniformly distributed on $[-\frac{2}{\sqrt{2\lambda}}, -\frac{1}{\sqrt{2\lambda}}] \bigcup [\frac{1}{\sqrt{2\lambda}}, \frac{2}{\sqrt{2\lambda}}]$.

We focus on two types of jumps: Independent jumps, $\sum J_s^x \perp\!\!\!\perp \sum J_s^y$, and co-jumps, $\sum J_s^x = \sum J_s^y$. In all cases we use the intensities $\lambda_x = \lambda_y = 1$.

The bias properties of the estimators with independent jumps are in the upper panels of Figure S.10 and the analogous results for the case with co-jumps are in the lower panels. We present results for $\rho = 0.25$ in the left panels and $\rho = 2/3$ in the right panels. Both the $Q_S$ and $K$ estimators are largely unaffected by jumps. This is to be expected, given their construction and influence functions. The situation is quite different for the Pearson estimator. Independent jumps induce a strong bias towards zero in $P$, while co-jumps induce a strong positive bias towards one. That $P$ is very sensitive to jumps is an implication of the influence function for $P$. This sensitivity to jumps motivates the commonly used truncation methods for computing realized variances, see Mancini (2001, 2009) and Andersen et al. (2012), and truncation methods are clearly also very useful if $P$ is to be use.




(a) Heston with independent jumps




(b) Heston with co-jumps

Figure S.10: The average correlation estimates of the correlation estimators, $P$, $K$, and $Q_S$ plotted against sampling frequency. The underlying model is the Heston model with independent jumps in the upper panels and co-jumps in the lower panels. The correlation coefficient is $\rho = 0.25$ in the left panels and and $\rho = 2/3$ in the right panels.