Estimation of Stock-Bond Correlation using High Frequency Data with Application to a Risk-Targeted Risk Parity Portfolio

Emil F. Andersen[a]

[a]*MSc. Mathematics-Economics at the Institute of Mathematics - Science & Technology, Aarhus University.*

Abstract

We study the empirical accuracy of a variety of estimators of asset price variation (realized measures) constructed from high frequency data and compare them using forecast comparison analysis which include the Superior Predictive Ability (SPA) of Hansen (2005) and the Equal Predictive Ability (EPA) testing approach of Diebold & Mariano (1995) and West (1996). We find the best estimator which have the lowest empirical loss under the "literature-preferred" proxy (a Realized Covariance sampled on a five minute frequency, RCov_{5min}) utilizing Patton (2011a) data-based ranking method in conjunction with the best results obtained from the comparison analyses. This estimator is then used in a bivariate Risk-Targeted Equally-weighted Risk Contribution (RTERC) portfolio containing ETF's for U.S. stocks and bonds where it is compared to a benchmark portfolio. The conclusion is threefold. First, our results determine that no estimator could vastly outperform the Threshold Realized Covariance on a one minute sampling scheme (ThreshCov_{1min}) as noted from the main analysis together with the average empirical loss. Furthermore we find little evidence that any of the estimators significantly outperform the "literature-preferred" RCov_{5min}, with the exception of ThreshCov_{1min} and ThreshCov_{5min} together with the Modulated Realized Covariance on a one second sampling scheme (that is, MRC_{1sec}). Second, we observe that non noise-robust estimators sampled at minute frequencies seems to provide much of the benefits of high frequency data without exposing the estimators to microstructure noise, and the empirical accuracy only slightly increases (on average) when considering noise-robust estimators on second frequencies. Finally, we find evidence that constructing the RTERC portfolio using the best performing realized measure, ThreshCov_{1min} (high frequency portfolio) yields better overall results in contrast to the "industry standard" being a portfolio constructed from a daily covariance estimator, RCov_{daily} (benchmark portfolio). The high frequency portfolio reduces the vol of vol and is closer to the risk-target while also reducing the likelihood of very negative returns. In conclusion the high frequency portfolio is much more stable against strong fluctuations in the correlation estimates contrary to the benchmark portfolio.

1. Introduction

One of the most important shifts in the investment landscape over the past two decades has been the emergence of a negative stock-bond correlation. This transformed basic hedging properties of bonds giving them a more sustainable role in the construction of efficient and diversified portfolios. Negative correlation imposes better diversification, in the sense that the stock-bond correlation to some degree reflects how effectively the bond position will hedge against a significant equity market sell-off. Despite a fundamental importance of the stock-bond correlation for hedging properties and for portfolio construction, its underlying macroeconomic drivers are not well understood. Yet, as argued by Campbell et al. (2019)[1] (and indirectly by Yang et al. (2009)) it is the correlation between economic growth shocks and inflation shocks that contributes most to changes in the stock-bond correlation often leading to a regime shift. If the correlation between these macroeconomic drivers are positive, then the stock-bond correlation is negative and vice versa. This is due to the fact that economic growth shocks tend to induce a negative stock-bond correlation, as equities tend to rally following a positive growth surprise, while bonds tend to underperform due to expectations of tighter monetary policy (as seen via a Taylor Rule equation).

$$i_t = \pi_t + r_t^* + a_\pi(\pi_t - \pi_t^*) + a_y(y_t - \bar{y}_t). \qquad (1)$$

Inflation shocks, on the other hand, tend to induce a positive stock-bond correlation, as both equities and bonds tend to weaken following a positive inflation surprise. Bonds weaken due to expectations of tighter monetary policy (again via a Taylor rule), and the higher bond yields will tend to weaken equities as the NPV of future earnings/dividends fall. Note that this effect also is in place following a growth shock, but will typically be

[1] see p. 22 and *figure 2* at p. 50. The forthcoming sentence is based on the authors intuitive framework.

dominated by higher earnings projections.

The 1970-1980's were a period with large inflation surprises, (and stagflation with rising inflation and weak growth) and this helps explain the positive stock-bond correlation in the late 1900's. However, the Fed and other central banks have since the inflation blow up, managed to stabilize inflation expectations, and long-term inflation expectations have been very ancored around 2% since 2000 (see OECD (2018) pp. 206 - 209). Hence, since 2000 there have been relatively modest inflation surprises, while there have been more frequent and sizable growth surprises (eg. the recession in early 2000 and following the global financial crisis), which explain the fall in the stock-bond correlation.

The negative stock-bond correlation since 2000 have increased the diversification benefits in a risk-parity portfolio, and effectively increased the risk-adjusted expected return, (all else equal) which might explain the popularity of this type of strategies over the past decade.

It has often been the case that asset managers tend to estimate the covariation between stocks and bonds using daily data with an approximately one month estimation window in order to get a stable covariance estimate. This could be an imperfect estimation leading to suboptimal allocations in the portfolio setup and thus deviating from the volatility target. Moreover using intra-day data leads to a quicker "stable" covariance estimate under a smaller estimation window, making it easier for the portfolio manager to reduce the size of the portfolio before losing to much value due to exogenous market events.

We use realized measures constructed from intra-day data to acquire a better estimation of the volatilities and especially the stock-bond correlation using 1 year (250 days) of daily data on ETF's for stocks and bonds[2]. We will compare the realized measures from a high- to low frequency calendar time sampling scheme using the multivariate QLIKE loss function (called MQLIKE or also referred to as Steins loss function), where we impose a volatility proxy being the 5 minute Realized Covariance for comparison, since the quadratic covariation is non-observable in the market. Moreover we will be using jump robust realized estimators which will be compared with a volatility proxy being the 5 minute Bipower Covariance estimator[3]. We will inspect the differences in the forecasting accuracy of the realized measures using the SPA testing method of Hansen (2005) and the results are further strengthened by the EPA pair-wise comparison test of Diebold & Mariano (1995) and West (1996) using the data-based ranking method of Patton (2011), in order to see if we can hypothetically reject the difference in forecasting accuracy. Moreover we perform a small study of the Risk-Targeted Equally-weighted Risk Contribution (RTERC) portfolio using the best realized measure found in the earlier section. We impose different weight-updating schemes and compare them with the RTERC portfolio constructed on a daily sampled realized covariance estimator, called the benchmark portfolio. We end the section by performing a sensitivity analysis of the portfolio volatility by changing the correlation estimates from low to high.

The remainder of the paper is organized as follows. Section 2 provides a brief theoretical discussion of the theoretical model for the price processes including the price process under microstructure noise. Section 3 describes the realized measures used in the analysis together with the sampling frequencies, schemes and a thorough description of the data and the cleaning procedure. Section 4 describes the comparison of the realized measures together with the bootstrap approach and the forecast comparison analysis. Our empirical analysis is presented in Section 5 which also includes some preliminary results.The application to the RTERC portfolio framework is described in Section 6 which includes the theoretical setup together with the main results. The conclusion and further directions of research are presented in Section 7.

2. The Theoretical Setup

We develop a theoretical model that captures the empirical properties of financial asset prices which includes: stochastic volatility, jumps in the price and market microstructure noise. The first two features are modeled using a latent continuous-time stochastic process that represents the unobserved efficient price process. The last feature are modeled on top of the discretely sampled efficient price representing the observed tick data process.

2.1. A model for the price process

We consider a d-dimensional vector of log-prices X, defined on a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$, where the log-price process X is adapted to the information filtration $(\mathcal{F}_t)_{t\geq 0}$, and d denotes the number of assets under consideration. We assume that X is defined in an arbitrage-free frictionless market, such that X belongs to the class of semi-martingale processes (for further detailed information see eg. Delbaen and Schacher-

[2] Asset return volatility estimators constructed using high frequency data, which generally aim to estimate the quadratic variation/covariation are throughout the article referred to as "realized measures".

[3] In the study we found minimal effects on the losses for the jump robust estimators when changing the proxy as described.

mayer (2006)). The log-price process X are following an Îto jump-diffusion model with finite activity jumps, on the form:

$$X_t = X_0 + \int_0^t \mu_s ds + \int_0^t \sigma_s dW_s + \sum_{i=1}^{N_t^J} J_i, \quad t \geq 0, \quad (2)$$

where X_t is the log-price at time t, $\mu = (\mu_t)_{t\geq 0}$ is a d-dimensional predictable locally bounded drift vector, $\sigma = (\sigma_t)_{t\geq 0}$ is an adapted cádlág $d \times d$ covolatility matrix, $W = (W_t)_{t\geq 0}$ is a d-dimensional standard Brownian motion, $N^J = (N_t^J)_{t\geq 0}$ is a counting process, and $J = (J_{1i}, \ldots, J_{di})_{i=1,\ldots,N_t^J} = (J_i)_{i=1,\ldots,N_t^J}$ is a sequence of nonzero d-dimensional independently and identically distributed random variables, which are also independent of the Counting process[4]. Intuitively, N^J represents the total number of jumps that occurs in X up to time t, and J denotes the size of the jumps. The above Îto jump-diffusion model is a subclass of general diffusion processes and depending on the parameter choice of μ, σ and the distribution of J, different particular models arises, which are nested within the jump-diffusion model. As such, we will not specify the driving processes of the jump-diffusion process in order to allow for a less restrictive modeling setup.

2.2. *The noisy tick data model*

In practice we do not observe the efficient price process in continuous time, but instead a discretely sampled counterpart subject to microstructure noise, arising from market imperfections such as bid-ask bounce and the discreteness of price change (see Roll (1984), Harris (1991) & Black (1986)). Another important source of measurement error are rounding effects, since transaction prices are multiples of a tick size, and therefore contributes to the microstructure noise. (Harris (1990) & Gottlieb & Kalay (1985)). These market imperfections often yields observable noisy tick data which can be mathematically formulated as a stochastic process modeled on top of the efficient price process:

$$Y_{i/n} = X_{i/n} + \varepsilon_{i/n}, \quad \text{for } i = 1, \ldots, n \quad (3)$$

where $\frac{i}{n}$, $i = 1, \ldots, n$ are the observed equidistant intra-day time points in an assumed restrictive time interval $t \in [0, 1]$ and n is the number of intra-day points. Here, ε is a multivariate independently and identical distributed microstructure noise process with $\mathbb{E}[\varepsilon_t] = 0$ and covariance $\mathbb{E}[\varepsilon_t \varepsilon_t'] = \Psi$, where Ψ is a positive definite $d \times d$ matrix. Furthermore the noise process ε is independent of the efficient price process X. This implies that Y is a noisy d-dimensional vector of log-prices. It is a well known fact that the above additive noise structure is standard in the literature of high frequency data, because it captures the empirically observed negative first-order autocorrelation in tick returns as described in Hansen & Lunde (2006). Furthermore they show that the i.i.d. assumption and the independence between X and ε can be called into question when sampling at very high frequency eg. below the 1 minute mark.

3. Realized measures

We consider a variety of different realized measures, that have been well documented and widely used in high-frequency financial literature. The first class of realized measures is the standard realized covariance estimator (RCov), which is the sum of outer products of intra-day returns. This simple estimator is the realized analog of the quadratic covariation and is efficient in absence of noisy data (see Barndorff-Nielsen & Shephard (2002, 2004)). Therefore the mathematical formulation of the estimator can be presented as:

$$\text{RCov}_t = \sum_{i=1}^{n} r_{i,t} r_{i,t}' \xrightarrow{\mathbb{P}} [Y]_t, \quad \text{as } n \to \infty, \quad (4)$$

where $r_i = Y_{i/n} - Y_{(i-1)/n}$ are logarithmic returns of the noisy price process Y[5]. One way of countering the microstructure noise presented in the data, is to use sparse sampling and therefore sample with a lower frequency. Hansen & Lunde (2006) shows that the univariate counterpart diverges at high frequences (eg. less than 1 minute) and concludes that with a calender time sampling frequency at or above the 1 minute mark, the estimator becomes stable.

The second class of realized measures are the pre-averaged realized covariance estimator first introduced by Podolskij & Vetter (2009a) in the univariate form and later Christensen, Kinnebrock & Podolskij (2010) extended the work to a multivariate framework, naming the estimator: The Modulated Realized Covariance (MRC). The general idea behind the pre-averaging technique is to average out the noise in the high frequency data. This is done by applying a weighting function to the observed returns to construct the pre-averaged returns. The most common weighting function has a Bartlett kernel-like tent shape

[4]It is a reasonable assumption only to consider a finite activity model like the above jump diffusion model, since it appears in conjunction with an infinite activity diffusion component, and thus the presence of an infinite activity jump component synthesizing small and large moves (with the use of a diffusion component) is both theoretically and practically redundant (see eg. Carr et al. (2002)).

[5]Consistency of this estimator comes as a consequence from the definition of quadratic covariation provided in equation (12).

and thus is identical to locally averaging prices and then taking the difference of the adjacent pre-averaged prices. As described by Podolskij & Vetter (2009) any continuously differentiable function (contained in the set of C^1), $g : [0, 1] \to \mathbb{R}$ with with a piecewise Lipschitz derivative g' such that $g(0) = g(1) = 0$ and $\int_0^1 g(s)^2\, ds > 0$, can be used as a weighting function for the pre-averaging. However we will be following the authors default choice of, $g(x) = \min(x, 1 - x)$. Under the assumption of finite fourth (absolute) moment of the noise process and letting $n \to \infty$ it holds that,

$$\mathrm{MRC}_t \xrightarrow{\mathbb{P}} \int_0^1 \Sigma_s\, ds + \frac{\Psi_1}{\theta^2 \Psi_2} \Psi, \tag{5}$$

with Ψ being the covariance matrix of the noise process and Ψ_1, Ψ_2 being the integrals of the Riemann approximations described in equation (8). Following the original paper the consistent bias-corrected MRC can be defined as:

$$\begin{aligned}\mathrm{MRC}_t = & \frac{n}{n - k_n + 2} \frac{1}{\Psi_2^{k_n} k_n} \sum_{i=0}^{n-k_n-1} \bar{r}_{i,t} (\bar{r}_{i,t})' \\ & - \frac{\Psi_1^{kn}}{\theta^2 \Psi_2^{k_n}} \frac{1}{2n} \sum_{i=1}^{n} r_{i,t} (r_{i,t})' \xrightarrow{\mathbb{P}} \int_0^1 \Sigma_s\, ds,\end{aligned} \tag{6}$$

where the deducted term denotes the bias-correction, which compensates for the residual microstructure noise that remains after pre-averaging. Here k_n is the pre-averaging window length and is given $k_n = \lfloor \theta \sqrt{n} \rfloor$ with theta being a user-defined choice $\theta \in (0, \infty)$ i.e. the pre-averaging horizon grows as the sampling frequency increases, but at a slower rate. $\bar{r}_{i,t}$ denotes the pre-averaged observed returns in a neighbourhood of k_n observations[6]:

$$\bar{r}_{i,t} = \frac{1}{k_n} \left(\sum_{j=k_n/2}^{k_n-1} Y_{\frac{i+j}{n}} - \sum_{j=0}^{k_n/2-1} Y_{\frac{i+j}{n}} \right), \tag{7}$$

where, $k_n \geq 2$ and needs to be even. Furthermore to avoid biases in small samples, we define the Riemann sum-approximations of the corresponding asymptotic numbers:

$$\begin{aligned}\Psi_1^{k_n} &= k_n \sum_{i=1}^{k_n} \left(g\left(\frac{i}{k_n}\right) - g\left(\frac{i-1}{n}\right) \right)^2 \\ \Psi_2^{k_n} &= \frac{1}{k_n} \sum_{i=1}^{k_n-1} g^2 \left(\frac{i}{k_n}\right).\end{aligned} \tag{8}$$

The choice of the noise variance estimator in the bias-correction term in (6) is linked to the estimator proposed by Aït-Sahalia et al. (2005), which converges asymptotically to the true noise process when $n \to \infty$. However under finite sample properties the estimator introduces some bias in the sense that it estimates,

$$\sum_{i=1}^{n} r_{i,t} (r_{i,t})' = 2n\Psi + \int_0^1 \Sigma_s\, ds + o_p(n^{-1}) \tag{9}$$

where the error of this approximation has expectation zero. We quickly note from equation (5 & 6) the noise variance estimator under finite samples becomes:

$$\frac{\Psi_1^{kn}}{\theta^2 \Psi_2^{k_n}} \frac{1}{2n} \sum_{i=1}^{n} r_{i,t} (r_{i,t})' = \Psi \frac{\Psi_1^{kn}}{\theta^2 \Psi_2^{k_n}} + \frac{\Psi_1^{kn}}{\theta^2 \Psi_2^{k_n}} \frac{1}{2n} \int_0^1 \Sigma_s\, ds, \tag{10}$$

which in turn implies that the consistent bias-corrected MRC in equation (6) under finite samples actually estimates,

$$\mathrm{MRC}_t = \left(1 - \frac{\Psi_1^{k_n}}{\theta^2 \Psi_2^{k_n}} \frac{1}{2n}\right) \int_0^1 \Sigma_s\, ds, \tag{11}$$

and conclusively we have to scale the estimator with $1/\left(1 - \frac{\Psi_1^{kn}}{\theta^2 \Psi_2^{kn}} \frac{1}{2n}\right)$. The same construction is described in Christensen, Kinnebrock & Podolskij (2010).

The noise variance can be estimated in a number of ways (see eg. Gatheral & Oomen (2010) for a comparison of estimators), where a good alternative is the first order autocorrelation estimator proposed by Oomen (2006). However any use of alternative noise measures implies a correction of the scaling in the MRC. Using any form of bias-corrected estimator we cannot ensure that the estimate is positive semidefinite in finite samples. Nevertheless we verify the positive semi-definiteness of the estimates by checking that the eigenvalues are positive.

From the work of Andersen et al. (2007) and Barndorff-Nielsen & Shephard (2005) they have shown that you can decompose the quadratic covariation into the integrated

[6]The formulation in (7) corresponds to choosing the authors default choice of weighting function $g(x)$, and is thus a particular coincidence. Another formulation can be considered as:

$$\bar{Y}_{\frac{i}{n}} = \sum_{j=1}^{k_n-1} g\left(\frac{j}{k_n}\right) \left(Y_{\frac{i+j}{n}} - Y_{\frac{i+j-1}{n}}\right),$$

for $Y_{\frac{i}{n}}$ being defined in (3).

covariation and the multivariate jump component:

$$[Y]_t = \operatorname*{plim}_{n\to\infty} \sum_{i=1}^{n} \left(Y_{t_i} - Y_{t_{i-1}}\right)\left(Y_{t_i} - Y_{t_{i-1}}\right)' \qquad (12)$$
$$= \int_0^t \Sigma_s \, ds + \sum_{i=1}^{N_t^J} J_i J_i'$$

where $0 = t_0 < \cdots < t_n = t$ is a sequence with deterministic partitions, and a decreasing mesh $\sup_i (t_i - t_{i-1}) \to 0$ for $n \to \infty$. When using high frequency financial data over a longer time horizon jumps become inevitable. Hence we consider two varieties of jump robust realized measures. The first one is the Bipower realized Covariance estimator (BPCov) of Barndorff-Nielsen & Shephard (2004), who naturally extended the univariate version using the interpolation identity. The estimator is defined as the process whose value at time t is the d-dimensional square matrix with k-, l'th element equal to:

$$\mathrm{BPCov}_t = \frac{\pi}{8} \sum_{i=2}^{n} \Bigg(\left| r_{(k)t,i} + r_{(l)t,i} \right| \cdot \left| r_{(k)t,i-1} + r_{(l)t,i-1} \right| \qquad (13)$$
$$- \left| r_{(k)t,i} - r_{(l)t,i} \right| \cdot \left| r_{(k)t,i-1} - r_{(l)t,i-1} \right| \Bigg).$$

The estimator can be consistently estimated using an equally spaced discretization of high frequency financial data. The scaling $\pi/8$ ensures that the estimator converges to the integrated covariance:

$$\operatorname*{plim}_{n\to\infty} \mathrm{BPCov}_t = \int_0^1 \Sigma_s \, ds. \qquad (14)$$

The intuition behind the jump-robustness of the estimator stems from the fact that the multi-plum in both terms within the sum contains a jump has probability measure zero (we are working under finite active jumps). Therefore multiplying these terms downscales the jump component and thus provides us with a consistent jump-robust estimator. Furthermore we make use of the pre-averaging approach implemented on the Bipower realized Covariance estimator. This yields an estimator that is both jump-robust and noise-robust. The pre-averaging setup follows the structure above with the same bias-correction term and noise variance estimator. Therefore the pre-averaged counterpart can be formulated as (see eg. Christensen et al. (2014) for a slightly different univariate formulation of the same estimator):

$$\mathrm{BPCov}_t^* = \frac{n}{n - 2k_n + 2} \frac{1}{k_n \Psi_2^{k_n}} \mathrm{BPCov}_t^{mod} \qquad (15)$$
$$- \frac{\Psi_1^{kn}}{\theta^2 \Psi_2^{k_n}} \frac{1}{2n} \sum_{i=1}^{n} r_{i,t} (r_{i,t})',$$

with BPCov_t^{mod} being equivalent to BPCov_t just summing over $n - 2k_n + 2$ components[7,8]. The last jump-robust realized measure that we consider is the Threshold Realized Covariance estimator (ThreshCov) originally proposed by Gobbi & Mancini (2012). The estimator uses univariate jump detection rules to truncate the effect of jumps on the covariance estimate, which makes the estimator viable in high dimensions. The estimator is defined as the value at time t, in the d-dimensional square matrix with k-, l'th element equal to:

$$\mathrm{ThreshCov}_t = \sum_{i=1}^{n} r_{(k)t,i} \mathbb{1}_{\{|r_{(k)t,i}| \le TL_n\}} r_{(l)t,i} \mathbb{1}_{\{|r_{(l)t,i}| \le TL_n\}}, \qquad (16)$$

where the truncation level TL_n follows the paper of Jacod & Todorov (2009) and is equal to:

$$TL_n = 3 \cdot \sqrt{\frac{\mathrm{BPVar}_T}{T}} \cdot \left(\frac{1}{n}\right)^{0.49}, \qquad (17)$$

where BPVar_T is the univariate counterpart of BPCov_t and $T = 1$ in our theoretical setup (see eg. Barndorff-Nielsen & Shephard (2006) for a formulation of the univariate bipower estimator). Intuitively, the truncation level just discards any returns that are above/below three standard deviations away from the mean in the standard Gaussian distribution. This stems from the assumption of asymptotic normalized returns on the form[9]:

$$\sqrt{\Delta_n} \cdot r_{t,i} | \mathcal{F}_{i/n} \overset{a}{\sim} N\left(0, \frac{\sigma^2}{T}\right) \qquad (18)$$

which can be standardized to equation (17) for $T = 1$, and $\Delta_n = 1/n$. The threshold convergence is chosen to be 0.49 so that it converges slower than the Brownian increments (decreases approximately of $\sqrt{\Delta_n}$) imposed on the variability of our returns, when $n \to \infty$. This ensures that we do not truncate away returns in a highly volatile period. However, the increments containing jumps stays the same and consequently fall above/below three standard deviations away from the mean, and is thus truncated away. While BPCov_t and $\mathrm{ThreshCov}_t$ are only jump-robust, we can use sparse sampling in order to eliminate the microstructure noise presented in the financial data.

3.1. Sampling frequency and sampling scheme

As hinted throughout section 3, we use the Calender Time Sampling (CTS) scheme, which selects transactions

[7] In order to avoid confusion, BPCov_t^{mod} just has the summation index going from $i = 1$ to $n - 2k_n + 2$, instead of $i = 1$ to n.

[8] This estimator has also been scaled in the empirical work as described above for the Modulated Realized Covariance or as in Christensen, Kinnebrock & Podolskij (2010).

[9] This follows from the Brownian increments in the diffusion process defined in (2), in a non-noisy setup.

at an regularly spaced time interval. This implies that we will get the same amount of observations each day for each asset, but distributed irregularly throughout the day[10]. We consider using one-, five-, fifteen-minute and daily (close-to-close) sampling frequencies for the realized measures as these are widely used in the high frequency literature and provide a natural reference point. As described in Hautsch & Podolskij (2013) we ought to use a high sampling frequency for the pre-averaging realized measures, since pre-averaging based on sparse sampling implies a downward bias of volatility. Therefore we further consider sampling frequencies of one- and five-seconds in order to aid the pre-averaging realized measures. However using high frequency calendar time sampling, induces the problem of sampling mainly zero return intervals especially when dealing with illiquid assets. As seen in the next section, both of our assets are quite liquid, with a high proportion of non-zero returns in the original data. Moreover the Bipower realized (co)variation (and power variations in general) suffer from a high proportion of zero returns, and will thus be biased towards zero. In total, across sampling frequencies we include 28 different realized measures in our analysis[11].

3.2. Data description

We have a large set of transaction data[12] covering two Exchange Traded Funds (ETFs) in the period of 02-01-2018 to 28-12-2018 amounting to 250 trading days of intra-day data. The first asset is a highly liquid index traded as an ETF titled SPY which tracks the S&P 500 index. The second asset, iShares 20+ Year Treasury Bond ETF (ticker: TLT) securitized by BlackRock seeks to track the investment results of an index composed of U.S. Treasury Bonds with remaining maturities greater than 20 years. TLT can be seen as a semi-liquid asset in contrast to SPY. Furthermore the intra-day data are with second precision timestamps allowing for a detailed evolution of the price data. We will restrict attention to the official trading hours from 9:30 to 16:00 Eastern Standard Time (EST). Table 1 reports the descriptive statistics for our two-asset universe. As noted these two assets display a varying degree of liquidity with SPY being the most liquid asset. Furthermore we have reported the univariate noise statistic ξ, which gives us information on the magnitude of microstructure noise as represented by its variance $\hat{\mu}_t^2 = \int_0^t \mu_s^2\, ds$. The estimation is given as:

$$\hat{\mu}_t^2 = \begin{cases} -\frac{1}{n-1}\sum_{i=2}^{n}(\Delta_i^t r)(\Delta_{i-1}^t r). & \text{if } \sum_{i=2}^{n}(\Delta_i^t r)(\Delta_{i-1}^t r) < 0, \\ \frac{1}{2n}\sum_{i=1}^{n}(\Delta_i^t r)^2. & \text{otherwise,} \end{cases} \tag{19}$$

where $(\Delta_i^t r)$ denotes the noisy univariate returns for each asset at a specific day t. Then as suggested by Oomen (2006) and Hautsch & Podolskij (2013), we can compute the noise-to-signal ratio per trade as $\hat{\xi} = \hat{\mu}_t^2/(\mathrm{IV}_t/n)$, where IV_t is computed using the maximum likelihood estimator proposed by Aït-Sahalia et al. (2005)[13]. We find that the noise ratio of SPY is three times as high as the noise ratio for TLT in the period of 2018.

As a preliminary step, we subdued our data to the cleaning procedure described in Barndorff-Nielsen et al. (2008), and considered only some minor changes. (1) We retained data from two primary exchanges for SPY and the three biggest exchanges in terms of daily transactions executed for TLT (see Table 1). Instead of retaining data from a single exchange we used multiple in order to obtain more data (especially for TLT). (2) We deleted entries outside the official trading hours EST, (3) we deleted data points containing abnormal sales condition, corrected trades and/or a transaction price of zero, and (4) we aggregated data with the same time stamp using the median price. Furthermore we removed outliers by deleting data points above (below) the ask (bid) plus (minus) the bid-ask-spread.

Assets	SPY	TLT
Exchanges	N & T	Z, D & P
Raw trades	108693.73	15371.27
Avg. time aggregation	72110	5497.54
Avg. number of outliers	96.52	7.12
Avg. number of trades	11775.72	2610.54
Avg. proportion of nonzero trade returns	0.73	0.71
Noise Ratio, ξ	0.63	0.20

Table 1: Summary statistic for SPY and TLT. Raw Trades denotes the average amount of transaction data available from these exchanges during the trading session. Average number of trades constitutes the average amount of daily data after cleaning the prices. The two primary exchanges used for SPY are N: NYSE and T: NASDAQ. Using the total number of transactions executed each day, the three biggest exchanges for TLT are Z: BATS, D: NASD ADF and P: ARCA. Furthermore BATS constitutes for the biggest amount of volume for TLT.

[10] New information gets build into the asset prices at varying intensities.

[11] MRC and BPCov* could not be calculated using close-to-close returns.

[12] We will only be considering transaction data throughout the paper, due to the fact that using mid-quotes from quote data often would lead to the same outcome of the core analysis. Conclusively our realized measures are calendar-time sampled, transaction price estimators.

[13] A compact and clear description of this estimator can be found in the appendix of Hautsch & Podolskij (2013).

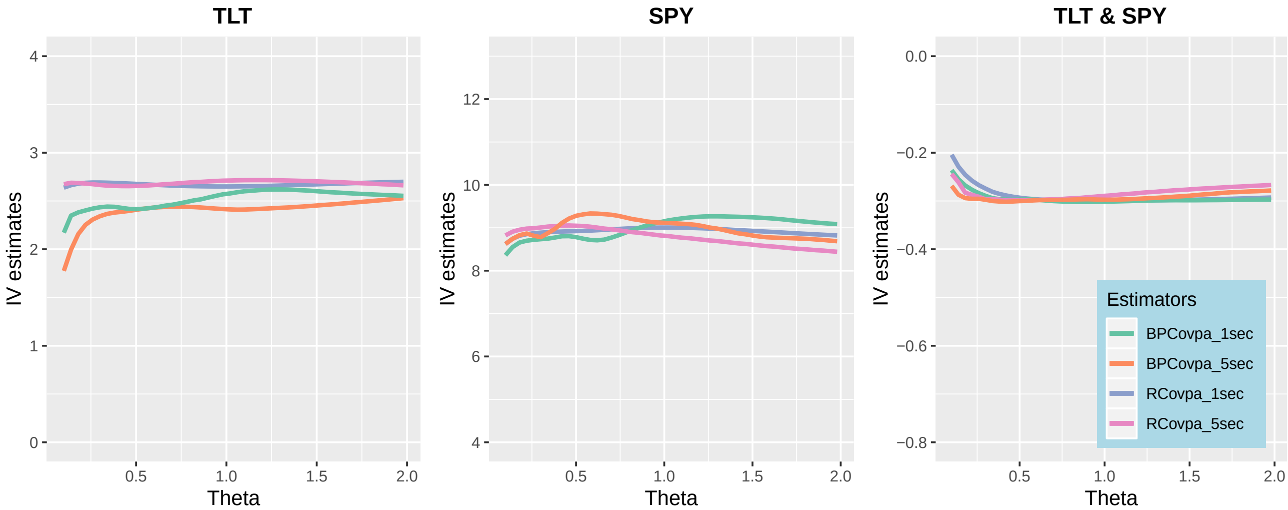

Figure 1: The daily average of the pre-averaged realized measures (multiplied by $1e5$) depending on θ. From left to right are the two variance plots for TLT and SPY together with their correlations. The estimators in question are BPCov_{1sec} (BPCovpa_1sec), BPCov_{5sec} (BPCovpa_5sec), MRC_{1sec} (RCovpa_1sec) and RCov_{5sec} (RCovpa_5sec).

3.3. Choosing the optimal θ

We propose a small study of our own in order to find the best choice of θ for our pre-averaging estimators. From Figure 1, we observe the daily averages (multiplied by $1e5$) of the pre-averaging measures depending on θ. Here, we have separated the covariance matrix into the univariate variances for each asset together with the correlation between the two assets. From the figure we observe that the estimators sensitivity to the choice of θ, and therefore the local pre-averaging window k_n, is highest if θ is small, since each different sampling frequency diverges for small thetas. Furthermore when θ is large, the impact of the underlying sampling frequencies diminishes and the estimates tend to stabilize. As described in Hautsch & Podolskij (2013) for larger values of θ, the pre-averaging interval might become too high and conclusively we suffer a downward trend of the estimates. However they argue that an oversmoothing (implied by a high θ) is less harmful for the estimator's bias than undersmoothing (implied by a low θ). Therefore from Figure 1 we can conclude that a reasonable choice is $\theta = 1$, which also corresponds to the results of the in-depth study done in Christensen et al. (2010).

4. Comparing the accuracy of the realized measures

We examine the realized measures by comparing their estimation errors for a given day's quadratic covariation. Since quadratic covariation is not directly observable in the market, we can not directly use a metric like the mean-squared error for our comparison analysis without the use of a proxy. Therefore we overcome this obstacle using the data-based ranking method of Patton (2011a), which require some proxy $\hat{\theta}$ for the quadratic covariation, that is assumed to be unbiased but may be noisy[14]. In this sense, we will choose a proxy that is unlikely to be affected by market microstructure noise. However, as argued by Patton (2011b) using proxies that are more noisy will reduce the ability to discriminate between the estimators, but will not affect the consistency of the procedure. We will be using the RCov_{5min} as a covolatility proxy since it has been a keen choice in earlier high frequency volatility literature (eg. see Lily, Patton & Sheppard (2015) for an extensive study on the RV_{5min}) and Hansen & Lunde (2006) specifically proposed to sample the univariate counterpart with a 5 minute sampling scheme. Patton (2011a) (*p. 287*) describes the use of a one day lead for the proxy in order to "break" the dependence between the estimation error in the realized measure under analysis and the estimation error in the proxy. This is heavy induced from the fact that the estimators using the same daily high frequency data are correlated[15]. The data-based ranking system allows for a variety of loss functions, where we will be focusing on the multivariate QLIKE loss function (MQLIKE) (also referred to as Stein's loss function from James and Stein (1961)). This is based on earlier literature showing that the univariate counterpart leads to more power to reject inferior estimators (see the empirical section in Hansen & Lunde (2005) and Patton (2011a) for the argumentation).

[14] This method requires the realized estimator to be unbiased under finite sampling schemes and not just asymptotically unbiased when the mesh goes to zero.

[15] As described in Patton (2011a), the use of a lead of the proxy relies on the daily quadratic variation following a random walk. The same can be applied to the multivariate setup.

Stein's loss function is given:

$$\text{MQLIKE:} \quad L(\hat{\theta}, \Sigma) = \text{tr}\left(\hat{\theta}^{-1}\Sigma\right) - \log|\hat{\theta}^{-1}\Sigma| - d, \quad (20)$$

where the quasi-likelihood (MQLIKE) measure is based on the negative of the log-likelihood of a multivariate normal. Here, $\hat{\theta}$ is a proxy for the quadratic covariation, Σ is the realized measure under analysis, $|\cdot|$ denotes the determinant and d is the asset dimension (throughout the empirical analysis we only consider two assets, thus $d = 2$)[16]. As described in Patton (2011a) we obtain a consistent (as $T \to \infty$) estimate of the difference in accuracy between any two realized measures:

$$\bar{d}_{ij} = \frac{1}{T}\sum_{t=1}^{T} \Delta\hat{L}_{ij,t} \xrightarrow{\mathbb{P}} \mathbb{E}\left[\Delta L_{ij,t}\right], \quad (21)$$

where $d_{ij} = \Delta\hat{L}_{ij,t} = L(\hat{\theta}_t, \Sigma_{it}) - L(\hat{\theta}_t, \Sigma_{jt})$ and $\Delta L_{ij,t}$ being the loss difference under the true quadratic covariation, θ. Then under the same regularity conditions defined in Patton (2011a) (*p. 287*), we can use block bootstrap to conduct forecast comparison analysis on the realized measures.

Block Bootstrap & Pair-wise comparison

We will improve the quality of statistical inference in our forecast comparison analysis in the noisy-jump diffusion setting by relying on the block bootstrap when computing critical values for the t-statistics. We will use the stationary block bootstrap approach from Politis & Romano (1994) in order to resample our sample population $B = 10000$ times. Furthermore the optimal mean block length has been chosen to be 9 days following the paper of Politis & White (2004). Under standard regularity conditions (see the introduction of Politis & White (2004)), they minimize the mean squared error of the stationary block bootstrap variance by optimally choosing the block length equal to:

$$\hat{b}_{\text{opt},SB} = N^{1/3}\left(\frac{2\hat{G}^2}{\hat{D}_{SB}}\right)^{1/3} \quad (22)$$

which solves the minimization problem of:

$$\text{MSE}(\hat{\sigma}^2_{b,SB}) = \frac{G^2}{b^2} + D_{SB}\frac{b}{n} + o(b^{-2}) + o(b/n). \quad (23)$$

They estimate the bootstrap variance based on the notion of spectral estimation via flat top lag-windows of Politis & Romano (1995). Therefore G and D_{SB} omits closed form solutions as seen in Politis & White (2004), but are estimated using:

$$\hat{G} = \sum_{k=-M}^{M} \lambda\left(\frac{k}{M}\right)|k|\hat{R}(k) \quad (24)$$

$$\hat{R}(k) = N^{-1}\sum_{i=1}^{n-|k|}\left(Z_i - \bar{Z}_n\right)\left(Z_{i+|k|} - \bar{Z}_n\right)$$

and,

$$\lambda(t) = \begin{cases} 1 & \text{if } |t| \in [0, 1/2] \\ 2(1-|t|) & \text{if } |t| \in [1/2, 1] \\ 0 & \text{otherwise} \end{cases}. \quad (25)$$

Here, Z denotes an arbitrary stationary time-series, $\bar{Z}$ denotes the empirical mean and the function $\lambda(t)$ has a trapezoidal shape around zero, which the "flat-top" lag window proposed in Politis & Romano (1995). Furthermore we have:

$$\hat{D}_{SB} = \left(4\hat{g}^2(0)\frac{2}{\pi}\int_{-\pi}^{\pi}(1+\cos(w))\hat{g}^2(w)\,dw\right) \quad (26)$$

with,

$$\hat{g}(w) = \sum_{k=-M}^{M} \lambda\left(\frac{k}{M}\right)\hat{R}(k)\cos(wk), \quad (27)$$

where $\hat{g}(w)$ is an accurate estimation of the spectral density, $g(w) = \sum_{k=-\infty}^{\infty} R(k)\cos(wk)$. As argued by the authors, the choice of the flat top lag-window $\lambda(t)$ is to ensure the best rate of convergence, since the smoothing is highly accurate and taking advantage of the fast decay in the autocovariance function $R(k)$.

In a time-series framework, the block bootstrap appears natural since it has been consistently shown that pricing processes of financial products exhibit serial dependence and heteroskedasticity, which needs to be considered when implementing the bootstrap.

We utilize the stationary bootstrap on Hansen (2005) testing approach for SPA, which uses a studentized test statistic that reduces the influence of erratic forecasts and invoke a sample-dependent null distribution. The test is favourable in contrast to White (2000) Reality Check in terms of robustness for poor forecasts, but the test is not entirely immune to the inclusion of bad forecasts, so some thoughts about the relevancy of the forecasts should be considered (see the conclusion in Hansen (2005)). When testing for Superior Predictive Ability (SPA), the question of interest is whether any alternative forecast is better than the benchmark forecast. Therefore we test the

[16] Furthermore the loss function is scale invariant which was very beneficial when encountering numerically singular covariance estimates in the empirical analysis. The solution was to scale the returns by a constant $C > 1$, and recalculate the loss function.

null hypothesis that "the benchmark is not inferior to any alternative forecasts".

Let the relative performance variables be defined as below equation (21) with the benchmark being model "0", ie $i = 0$ such that, we can redefine the relative performance as $d_{k,t}$[17], then for $k = 1, \ldots, m$ alternative forecasts we vectorize the relative performance as $d_t = (d_{1,t}, \ldots, d_{k,t})^T$. Provided that $\mathbb{E}[d_t] = \mu$ is well-defined, we can formulate the composite null hypothesis of interest as:

$$\mathrm{H}_0 : \quad \mu \leq 0. \tag{28}$$

It is important to note that Hansen (2005) works under the assumption that model k is better than the benchmark if and only if $\mathbb{E}[d_{k,t}] > 0$. So the focus is on the properties of d_t for $t = 1, \ldots, n$. One important assumption of the test is that d_t imposes strict stationarity and second finite moment[18], and an immediate consequence of the assumptions together with the application of the central limit theorem gives us the asymptotic null distribution:

$$n^{1/2}(\bar{d} - \mu) \xrightarrow{d} N_m(0, \Omega_{SPA}), \tag{29}$$

where $\bar{d}$ is the vectorized version of equation (21)[19] and $\Omega_{SPA} \equiv \text{avar}(n^{1/2}(\bar{d} - \mu))$, see Hansen (2005). He considers a simpler test statistic, T_n^{SPA} (defined later) that requires only the diagonal elements of Ω_{SPA} to be estimated. However, as many other studentized test statistics we need a consistent estimate $\hat{\Omega}_{SPA}$ of Ω_{SPA} which in general does not omit an analytical expression. Therefore the motivation for the bootstrap is not driven by higher-order refinements, but merely to handle the nuisance parameter problem (estimation of $\hat{\Omega}_{SPA}$)[20].

The current setup described above is identical to the Reality Check however the differences can be spelled out between Hansen (2005) SPA test and the Reality Check in two simple points:

1. Use the studentized test statistic:

$$T_n^{SPA} = \max\left[\max_{k=1,\ldots,m} \frac{n^{1/2}\bar{d}_k}{\hat{\omega}_k}, 0\right] \tag{30}$$

where $\hat{\omega}_k^2$ is some consistent estimator of $\omega_k^2 \equiv \text{var}(n^{1/2}\bar{d}_k)$.

2. Invoke a null distribution that is based on $N_m(\hat{\mu}^c, \hat{\Omega}_{SPA})$, where $\hat{\mu}^c$ is a carefully chosen estimator for μ, which is defined as:

$$\hat{\mu}^c = \bar{d}_k \cdot 1_{\{n^{1/2}\bar{d}_k/\hat{\omega}_k \leq \sqrt{2\log\log n}\}}, \tag{31}$$

where a detailed theoretical description of the choice of $\hat{\mu}^c$ is found in Hansen (2005) p. 370 - 371.

The bootstrap approach is now centred around the pseudo time series $\{d^*_{b,t}\}$ for $b = 1, \ldots, B$ are resamples of d_t with $B = 10000$ as recommended by Hansen (2005). From the pseudo time-series, we calculate their sample averages:

$$\bar{d}^*_b = n^{-1}\sum_{t=1}^{n} d^*_{b,t}, \quad b = 1, \ldots, B, \tag{32}$$

that can be viewed as asymptotically independent draws of $\bar{d}$ under the bootstrap distribution. This provides an intermediate step to estimate the distribution of the studentized test statistic, where we further consider the bootstrap covariance estimator of Politis & Romano (1994) as the estimator for $\hat{\omega}^2$:

$$\hat{\omega}_k^2 = \hat{R}(0) + 2\sum_{k=1}^{n-1}\kappa(n,k)\hat{R}(k), \tag{33}$$

with

$$\kappa(n,k) = \frac{n-k}{n}(1-q)^k + \frac{k}{n}(1-q)^{n-k}, \tag{34}$$

being the kernel weights under the stationary bootstrap with $q \in [0,1]$ and $\hat{R}(k)$ being defined in equation (24). We seek the distribution of the test statistics under the null hypothesis, so Hansen (2005) imposes the null by recentering the bootstrap variables about $\hat{\mu}^l$, $\hat{\mu}^c$ or $\hat{\mu}^u$. This is done by defining a centered matrix equal to:

$$\begin{aligned} Z^*_{k,b,t} &= d^*_{k,b,t} - g_i(\bar{d}_k), \\ &i = l, c, u, \ b = 1, \ldots, B, \ t = 1, \ldots, n, \end{aligned} \tag{35}$$

where the function $g_i(\cdot)$ is defined differently depending on the choice of mean estimator:

1. $g_l(x) = \max(0, x)$.
2. $g_c(x) = x \cdot 1_{\{z \geq -\sqrt{(\hat{\omega}^2/n)2\log\log n}\}}$.
3. $g_u(x) = x$.

The different means lead to different interpretations of the p-values. The upper bound $\hat{\mu}^u$ (also called SPA_U) produces the p-value of a conservative test that assumes that all of the k models are precisely as good as the benchmark in terms of expected loss ($\mu_1 = \ldots = \mu_k$). The lower bound $\hat{\mu}^l$ (SPA_L) produces the p-value of a liberal

[17] The benchmark model becomes fixed and only the alternative forecasts are varying together with the time dimension

[18] So we can apply Central Limit Theorem.

[19] That is, $\bar{d} = n^{-1}\sum_{t=1}^{n} d_t$.

[20] White (2000) provides a thorough theoretical explanation of the bootstrap approach to the Reality Check, which is identical to the steps described above (but Hansens SPA test will differ later).

test and assumes that the models with worse performance than the benchmark are poor models for the asymptotic distribution, hence if its mean is less than zero. Therefore as stated in Hansen (2005) these can be viewed as upper and lower bounds for the consistent p-value, respectively. The test $\hat{\mu}^c$ producing the consistent p-value (SPA$_C$) will determine which model performs worse than the benchmark and asymptotically prevent them from affecting the distribution of the test statistic. While the upper bound is sensitive to the inclusion of poor forecast models, the consistent and liberal tests are not. Furthermore Hansen (2005) argues that we can approximate the distribution of our test statistics under the null hypothesis by the empirical bootstrap distribution obtained from $Z^*_{b,t}$. The p-value of the three tests for SPA are now simple to obtain. We calculate:

$$T^{SPA*}_{b,n} = \max\left[\max_{k=1,\dots,m}\left(n^{1/2}\bar{Z}^*_{k,b}/\hat{\omega}_k\right),\, 0\right], \tag{36}$$

for $b = 1, \dots, B$ and the bootstrap p-value is given by:

$$\hat{p}_{SPA} = \sum_{b=1}^{B}\frac{1_{\{T^{SPA*}_{b,n} > T^{SPA}_n\}}}{B}, \tag{37}$$

where the null hypothesis should be rejected for a given confidence level α. As a concluding remark Hansen (2005) argues that the SPA test is still valid even if $\hat{\omega}^2_k$ is inconsistent with ω^2_k.

We further consider an Equal Predictive Ability (EPA) test for the covariation forecasts of two estimators, which does not utilize any bootstrap methodology. We can define the 1-step ahead Diebold-Mariano and West statistic from Diebold & Mariano (1995) and West (1996) as:

$$\text{DM} = \frac{\bar{d}}{\sqrt{\frac{2\pi f_{\bar{d}}(0)}{T}}} \xrightarrow{d} N(0,1) \tag{38}$$

where $f_{\bar{d}}(0)$ is the spectral density for the loss differential at frequency 0. This stems from the assumption of covariance stationarity and short memory on the mean loss differential which implies that it is asymptotically normal distributed:

$$\sqrt{T}(\bar{d}_t - \mu) \xrightarrow{d} N(0, 2\pi f_{\bar{d}}(0)). \tag{39}$$

The Diebold-Mariano West test just tests whether two 1-step ahead forecasts have the same accuracy for a user specified loss function. Moveover West (1996) extended the framework to accommodate the situation where forecasts involve estimated parameters. The Diebold-Mariano-West two sided simple hypothesis is defined for a given loss differential d_{ij} as:

$$H_0\colon \bar{d}_{ij} = \mathbb{E}\left[d_{ij}\right] = 0,$$
The two forecasts have the same accuracy

$$H_A\colon \bar{d}_{ij} = \mathbb{E}\left[d_{ij}\right] \neq 0,$$
The two forecasts have different accuracy,

and using linearity of the normal distribution we end up with the test statistic defined in equation (38). Utilizing the standard normality of the test statistics (DM) the null hypothesis is rejected whenever $|DM| > 1.96$ under a 95% confidence interval.

5. Empirical results

5.1. Preliminary results and interpretation

As a preliminary evaluation of each estimator we start by giving a summary of the average MQLIKE losses for each realized measure based on Patton's data-based ranking method with a leading RCov$_{5min}$ as proxy for the true covariation process. It has been empirically evident (see Liu, Patton & Sheppard (2015) for an empirical study on the univariate counterpart) that the 5 minute RCov has a great trade-off between overall performance and mathematical simplicity, in contrast to other realized measures that tends to be more complex and in some cases only with a marginal better performance. Presented in Figure 2 we observe a graphical representation of the average losses for each estimator under the different sampling schemes. Furthermore Table 2 aids the graphical representation by depicting the precise losses of each estimator. The calculations for the daily sampling scheme was done using close-to-close returns and therefore we had two data-points, one from day $t-1$ and one from day t. The one day lead

Sampling Scheme/Estimators	RCov	BPCov	ThreshCov	MRC	BPCov*
Daily	6.9191	9.6203	7.8362	-	-
15 min	0.5336	0.6428	0.5366	1.1710	1.2653
5 min	0.4994	0.5494	0.4179	0.6750	0.7683
1 min	0.4888	0.5269	0.3894	0.5188	0.5867
5 sec	0.5918	0.7541	0.9202	0.4554	0.4883
1 sec	0.6260	1.6220	4.4431	0.4428	0.4669

Table 2: The average MQLIKE losses over the period of 250 days with RCov$_{5\text{ min}}$ as proxy for the true covariation process. The daily estimate for the realized measures are calculated using the close to close returns. Furthermore the estimators with the lowest loss for each sampling scheme are colored green, and the highest are red. MRC and BPCov* can not be calculated using too sparse sampled time frequencies such as daily sampling frequency.

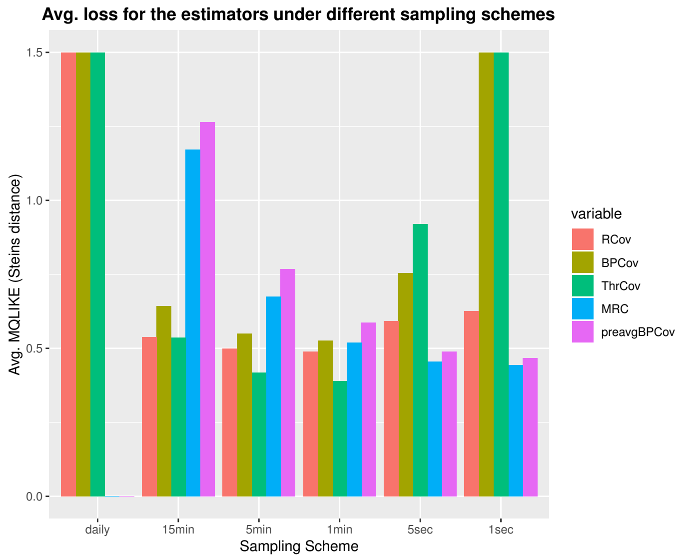

Figure 2: The figure shows the average losses over a period of 250 days for each estimator using different sampling schemes. Any extreme losses have been downscaled to 1.5 in order to produce a better overview of the differences between the estimates for each sampling scheme. In the graph, ThreshCov is "ThrCov" and BPCov* is "preavBPCov".

will contain data points from day t and $t+1$[21]. Therefore the one period ahead forecast for the daily scheme can be described as:

$$\mathbb{E}\left[\Sigma_{t-1:t+1}|\mathcal{F}_t\right] = \Sigma_{t-1:t}. \tag{40}$$

Understandably the pre-averaged versions of RCov and BPCov could not be calculated because of the sparse intraday data only yielding one data point for each asset. The daily sampling scheme are by far the worst in terms of losses with BPCov having a loss of 9.62 and RCov being the best performer with a loss of 6.92. However when we consider higher frequencies but still in the minute interval, we see that ThreshCov performs better than RCov and BPCov for 1 minute and 5 minute sampling schemes. The intuitive reason behind this, is twofold: It can be due to the truncation level of ThreshCov being more restrictive in removing jumps from the intra-day dataset under higher frequencies, than the downscaling of the jump component as done in the Bipower Realized Covariance estimator[22]. Finally, it can be the structure of BPCov, knowing that it will be biased towards zero (due to being dependable on the adjacent return observation) when introducing a high proportion of zero returns in the dataset, which will lead to a greater loss for higher frequencies. In contrast to RCov, the truncation of ThreshCov might eliminate returns that otherwise would introduce error to the estimator, thus providing with an estimate closer to the proxy. We observe that RCov for the 1- and 5-second frequencies does not explode compared to the other non noise-robust estimators. A further investigation reveals that the average QLIKE loss for each asset[23] is 0.16 and 0.15 for 1- and 5-second intervals respectively for TLT and 1.06, 0.60 for SPY. These numbers also agrees with the noise ratio depicted in Table 1 with an estimated noise ratio for SPY three times higher than TLT. Conclusively our bond proxy downscales the average MQLIKE loss for high frequencies due to the lack of a high proportion of noise as we would otherwise have in a portfolio solely consisting of equities.

Giving us more insight into the necessity of jump-robust estimators and how much they really matter under different sampling frequencies, we conduct a small study by empirically calculating the proportion of jumps observed each day. As argued in Christensen et al. (2014) jumps are often erroneously identified in lower frequencies, and in fact can be described as a burst of volatility when using higher frequency sampling schemes. A decrease of jumps in terms of the total price variation implies that RCov converges more towards the integrated covariance, since the jump component now induces a smaller effect than before (see equation (12)). Therefore theoretically, the variation between the average losses for RCov, BPCov and ThreshCov decreases when a reduction (or complete elimination) of the jump component occurs. This could be one of the reasons the variability between the average losses for the estimators is smaller for 1-minute sampling frequencies in contrast to 5- and 15-minute frequencies. However when noise is introduced, the argument fails due to the non noise-robustness of the estimators in question.

Calculating the jump proportion[24] gives us a hint on its size in terms of the different sampling schemes. The jump proportion can be expressed as the amount of jumps that contribute to the total quadratic variation during the time period. Under finite fourth moment of the noise process, $\mathbb{E}\left[\varepsilon^4\right] < \infty$, the pre-averaged Realized Variance (RV*) converges towards the quadratic variation of the price process, RV$^* \xrightarrow{\mathbb{P}} [Y]_1$ and the pre-averaged Bipower Variation (BV*) converges towards the integrated variance, BV$^* \xrightarrow{\mathbb{P}} \int_0^1 \sigma_s\, ds$. Therefore the magnitude of the jump

[21] One could argue that in order to keep the estimators within the same day, we could use open-to-close returns. However, our empirical studies yields the same extreme losses for this method, and since calculations on close-to-close returns are more widespread in the industry, we have only considered this methodology.

[22] It could also be a particular coincidence for this dataset.

[23] following from the univariate version of the MQLIKE function which is specified in Liu, Patton & Sheppard (2015) p. 297.

[24] also called jump variation.

	1 second frequency			5 second frequency			1 minute frequency		
	RV^*	BV^*	JP	RV^*	BV^*	JP	RV	BV	JP
TLT	12.1777	11.5539	10.3535	12.0281	11.6012	7.8042	12.3981	11.6756	12.1446
SPY	22.0486	21.5717	6.0719	21.8766	21.4432	6.5343	22.5028	22.0998	5.1315
Average	17.1132	16.5628	**8.2127**	16.9524	16.5222	**7.1693**	17.4505	16.8877	**8.6381**
	5 minute frequency			15 minute frequency			daily frequency		
	RV	BV	JP	RV	BV	JP	RV	BV	JP
TLT	11.9588	11.6478	8.0069	11.4771	11.2662	11.1667	11.4200	10.6945	24.3450
SPY	21.9259	21.4575	7.8219	20.9510	20.6772	10.0367	18.8640	17.3734	26.3545
Average	16.9424	16.5527	**7.9144**	16.2141	15.9717	**10.6017**	15.1420	14.0340	**25.3498**

Table 3: Reports jump variation estimates for SPY and TLT for all of the sampling schemes, together with the average. The jump variation are calculated using equation (42), RV*, BV*, RV, BV are calculated following Christensen et al. (2014), where RV*, BV* are the pre-averaged counterparts of RV and BV. Furthermore RV*, BV* and RV, BV are respectively the estimators for quadratic variation and integrated variance. The jump variation is expressed in percentages while RV*, BV* and RV, BV are expressed as percentage annualized standard deviations. Moreover the daily frequency is calculated using close-to-close returns, thus using two days for an estimation period. Calculating realized measures on few observations (2 observations for daily frequency) increases the variability of the estimators and therefore leads to higher jump variations. Thus the daily frequency should be taken with a grain of salt.

component can be consistently estimated as

$$\mathrm{RV}^* - \mathrm{BV}^* \xrightarrow{\mathbb{P}} \sum_{i=1}^{N_1^J} J_i^2, \tag{41}$$

for $n \to \infty$ and J_i being the univariate version of the one described below equation (2). Furthermore we can calculate the jump proportion as:

$$\mathrm{JP} = \frac{\mathrm{RV}^* - \mathrm{BV}^*}{\mathrm{RV}^*} \xrightarrow{\mathbb{P}} \frac{\sum_{i=1}^{N_1^J} J_i^2}{[Y]_1}. \tag{42}$$

The above equation gives us the average jump contribution to total variance.

Table 3 reports our findings for the implied jump proportion, together with the realized measures calculated using BV*, RV* and BV, RV, from Christensen et al. (2014)[25]. In Table 3 we have reported the average jump contribution to total quadratic covariation. The single most important observation to be made is that the implied jump variation for SPY inferred from the 1- and 5-second frequencies is small relative to the lower frequencies. However the same can not be said for TLT with the jump proportion being different for each sampling frequency. For 1- and 5-second on average, jumps account for a mere 10.35% and 7.80% of the total quadratic covariation. Furthermore from Christensen et al. (2014) we observe that the corresponding jump variation from SPY under 1-, 5- and 15-minute frequency are in line with the literature. However the results for the jump variation of TLT is undershot for the (comparable) 5 minute frequency with our results pointing towards a jump proportion on 8.01% and the literature revising a jump variation on average of 15%[26]. However, our results are computed on one year high frequency data on a highly diversified ETF in the period of 2018, which was a relative calm period in contrast to the general literature using a period from 1990 - 2002 of five minute intra-day data on pure longterm U.S. treasuries.

As described in Christensen et al. (2014) in *figure 3*, we observe that for equity data calculating BV*, RV* with $\theta = 1$, results in remarkably insensitive estimates to the choice of θ. Reconstructing the same plot (see Figure 3) for TLT produces the same results albeit giving us a slightly smaller jump proportion for higher thetas.

[25] Please be aware that in Christensen et al. (2014) they use a threshold filter on BV*, in order to remove the upwards bias of the finite sample estimator that is increasing in the choice of theta. This is not done in our study and thus we might be underestimating the jump proportion. However as seen in their simulation study, the bias is small.

[26] Please see Tauchen and Zhou (2011) and Andersen, Bollerslev and Diebold (2007), for "comparable" results.

Even though we indirectly can see the insensitivity of theta from Figure 1, with the multivariate pre-averaged counterparts, we further strengthen our argumentation by yielding the same results for RV* and BV*. Therefore our choice of θ is well reasoned.

Observing the relative high jump proportion of TLT in contrast to SPY under higher frequencies, together with the non-increasing nature of the jump proportion in frequencies of TLT, we conclude the following: Due to the nature of TLT the jump-robust estimators should not get excluded from higher frequencies (or lower frequencies whatsoever) based on the non-existing presence of jumps as argued for equities and currencies in Christensen et al. (2014).

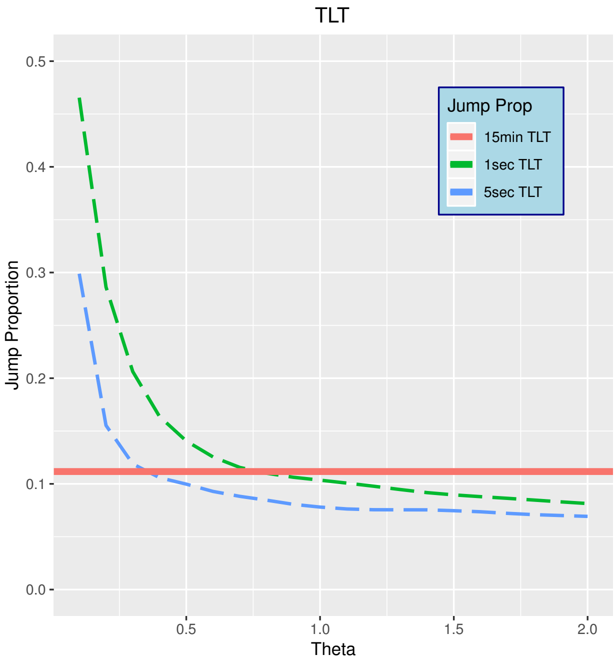

Figure 3: The figure shows the sensitivity of the jump proportion calculated using equation (42) with RV*, BV* with respect to the choice of theta for 1- and 5-second frequencies. As a comparison, the corresponding jump proportion for the 15 minute time scheme has been plotted with a solid thick line.

5.2. Main results

For the multiple comparison analysis we utilize the SPA test statistic of Hansen (2005) as theoretically described in the earlier section. The empirical use of the test in Hansen (2005) and also Hansen & Lunde (2005) focuses around a benchmark (ie ARCH(1,1) or a GARCH(1,1) in Hansen & Lunde (2005)) to test against alternative forecast models. In our situation we select the ThreshCov$_{1min}$ as the benchmark since it is the best performing estimator in terms of the lowest average MQLIKE loss function as seen in Table 2. We try to find statistical evidence about whether there exist any alternative forecasting model that are just as good or even better, by calculating the p-value in a 95% confidence interval and observing whether it is below the threshold for statistical acceptance. Hansen (2005) further argues that the test is not entirely immune to the inclusion of poor forecast models, but definitely more robust than the Reality Check of White (2000). For this reason we consider different sets of estimators (further called universes), where Ω_0 will comprise of all 28 realized measures including the poor forecasts such as the daily estimators, ThreshCov$_{1sec}$ and BPCov$_{1sec}$. The next set $\Omega_1 \subseteq \Omega_0$ comprises of all the reasonable forecasts excluding any forecasts above a 0.75 average MQLIKE loss. The last subset $\Omega_2 \subseteq \Omega_1 \subseteq \Omega_0$ contains only the best realized measure from each class. Therefore Ω_2 contains the estimators:

$$\Omega_2 = \{\text{ThreshCov}_{1min},\ \text{RCov}_{1min},\ \text{BPCov}_{1min},\ \text{MRC}_{1sec},\ \text{BPCov*}_{1sec}\}, \tag{43}$$

which is a very small size of alternative forecasting models in terms of the above mentioned articles. The results for the test are described in Table 4. We observe that we can not reject the null hypothesis of the benchmark having a worse average performance than any of the alternative forecasts in all of the universes. This further implies that the poor forecast in Ω_0 does not have an impact on the p-values since the same p-values are found in Ω_1 and Ω_2. The same result can be depicted from Table 2, but now we can further strengthen this argumentation with some statistical evidence (see also the Diebold-Mariano West test described below). Understandably, the full universe only containing 28 estimators with 249 losses each, is a tiny set which gives us one-sided results that are easy to interpret.

So we ask ourselves the question: What if we included more realized measures that performed just as well as the forecast models in each universe? To answer this question we consider a small study where we compare over 400 dummy estimators to the benchmark. The MQLIKE losses for the dummy realized measures are constructed with $t = 249$ days and utilizing the stationary bootstrap with a mean block length of 9 days across all of the losses for the alternative forecasts in each universe, Ω_i for $i = 1, 2, 3$. This yields an average expected MQLIKE loss for our dummy realized measures under the three universes of $\{\bar{L}^{\Omega_0}_{dummy}, \bar{L}^{\Omega_1}_{dummy}, \bar{L}^{\Omega_2}_{dummy}\} = \{1.3764,\ 0.4665,\ 0.3844\}$. Furthermore the full universe Ω_0 contains all of the dummy variables from the subsets, Ω_1 and Ω_0, and Ω_1 contains the dummy measures from subset Ω_2. This is done in order to observe the effects of introducing worse alternative forecast models to the test statistic. The inspiration for the amount of dummy variables comes from a mixture between Hansen (2005), Hansen & Lunde (2005) and Liu, Patton & Sheppard

Table 4: Shows the Superior Predictive Ability test of Hansen (2005) using the ThreshCov$_{1min}$ as the benchmark forecast model, with $B = 10000$ bootstrap replications. When we focus on the original sample (with 28 estimators or less) the test provide no further information on whether we hypothetically can provide any other alternative forecasts models that are just as or better than the benchmark. With a p-value of 1 in all cases, we can not reject the null hypothesis that the benchmark model is inferior to any of the alternative forecasts. Furthermore we have provided an alternative study where we consider the possibility of including more estimators by reproducing dummy losses for estimators using the stationary bootstrap approach using an average block length of 9 days. The dummies are constructed such that we resample Ω_i for $i = 1, 2, 3$ across all estimators. This further yielded an average expected MQLIKE loss for our dummy variables under the three universes of $\{\bar{L}^{\Omega_0}_{dummy}, \bar{L}^{\Omega_1}_{dummy}, \bar{L}^{\Omega_2}_{dummy}\} = \{1.3764,\ 0.4665,\ 0.3844\}$.

Benchmark: ThreshCov$_{1min}$							
Universe: Ω_0	SPA$_L$	SPA$_C$	SPA$_U$	Universe: Ω_0	SPA$_L$	SPA$_C$	SPA$_U$
Models:				*Models:*			
Original = 28 Dummies = 0 Total = 28	1	1	1	Original = 28 Dummies = 1200 Total = 1228	0.5048	0.5130	1
Universe: Ω_1	SPA$_L$	SPA$_C$	SPA$_U$	Universe: Ω_1	SPA$_L$	SPA$_C$	SPA$_U$
Models:				*Models:*			
Original = 19 Dummies = 0 Total = 19	1	1	1	Original = 19 Dummies = 800 Total = 819	0.4730	0.5050	0.6000
Universe: Ω_2	SPA$_L$	SPA$_C$	SPA$_U$	Universe: Ω_2	SPA$_L$	SPA$_C$	SPA$_U$
Models:				*Models:*			
Original = 5 Dummies = 0 Total = 5	1	1	1	Original = 5 Dummies = 400 Total = 405	0.4520	0.4942	0.5090

Table 5: Shows the pairwise comparison of the best realized measures in terms of lowest average QLIKE loss as seen in Table 2, which is equivalent to the subset Ω_2 as described above. The realized measures have been further decomposed into three different features: Not pre-averaged vs pre-averaged, Not Jump Robust vs Jump Robust and a "best of" feature between the two prior categories. In all (except one) pairwise comparisons we reject the null hypothesis of equal predictive ability for the forecast models, thus the forecast model with the lowest average loss should be considered.

Not pre-averaged vs pre-averaged		DM	$\bar{d}$	p-value
RCov_{1min} vs BPCov*_{1sec}	*Sample*	1.1722	0.0216	0.2422
ThreshCov_{1min} vs BPCov*_{1sec}	*Sample*	-4.7052	-0.0776	0.0000
BPCov_{1min} vs BPCov*_{1sec}	*Sample*	3.1402	0.0600	0.0018
RCov_{1min} vs MRC_{1sec}	*Sample*	3.2082	0.0458	0.0015
ThreshCov_{1min} vs MRC_{1sec}	*Sample*	-3.7427	-0.0533	0.0002
BPCov_{1min} vs MRC_{1sec}	*Sample*	4.0696	0.0843	0.0000
Not Jump Robust vs Jump Robust		DM	$\bar{d}$	p-value
RCov_{1min} vs ThreshCov_{1min}	*Sample*	4.4092	0.0991	0.0000
RCov_{1min} vs BPCov_{1min}	*Sample*	-3.2162	-0.0385	0.0014
Best of: Jump Robust & Pre-averaging		DM	$\bar{d}$	p-value
ThreshCov_{1min} vs BPCov_{1min}	*Sample*	-5.009	-0.1376	0.0000
BPCov*_{1sec} vs MRC_{1sec}	*Sample*	2.0804	0.0243	0.0385

(2015) where they in the latter tested more than 400 different realized measures across 31 different financial assets with more than 10 years of data and compared it to the univariate counterpart of the RCov_{5min}[27]. The most important observations to be made from this study can be summarized via three observations. Firstly, we can not find any statistical evidence that leads to a rejection of the null hypothesis. Even-though we have constructed 400 realized estimators for Ω_2 that are on average slightly above the ThreshCov_{1min} in terms of average loss, we do not have a high enough proportion of dummy measures that perform vastly better than the benchmark in order to reject the null hypothesis (the proportion of better alternative forecasts having an average loss below the average loss for the ThreshCov_{1min} is $223/405 \cdot 100 = 55.06\%$, with the lowest being an average loss of 0.3440)[28]. Moreover the small distance between SPA_L and SPA_U further implies that the worse performing alternative forecasts are only slightly worse than the benchmark, as argued in Hansen (2005) (see p. 378). The second observation to be made is that the upper bound SPA_U (also the p-value for White (2000) Reality Check) increases to some extend going from Ω_2 to Ω_0 which can be seen as "ad-hoc" evidence that the conservative upper bound is sensitive to poor forecasts. This is due to the decrease in performance for the dummy measures when going from Ω_2 to Ω_0, which indirectly can be seen on the expected average loss, $\bar{L}^{\Omega_0}_{dummy}$. The final observation stems from the small increase in the consistent p-value (SPA_C), going from Ω_2 to Ω_0, which further implies that the consistent p-value is not entirely immune to the inclusion of poor forecasts, even in our setup. Therefore as described before, it is a good idea to only include any reasonable forecast models and exclude the ones that are extremely poor. An inclusion of very poor forecasts can lead to a false acceptance of the null hypothesis and thus prevent us from concluding that an alternative forecast model might be better than the benchmark.

As hinted throughout the article, the RCov_{5min} has been a "favorite" within academic research as the "go-to" benchmark realized measure due to its simplicity while still providing many of the benefits from high-frequency data. Therefore we construct the same analysis as provided with ThreshCov_{1min}. From the analysis we reject the null hypothesis, and thus conclude that we have alternative forecasts with a better average performance. This is of course understandable since we have 6 estimators that contains a lower average loss than the RCov_{5min}. We try to remove the three best estimators in terms of average loss and see if any change happens in the statistical tests. We observe in Table 6 that we can not reject the null hypothesis in any of the universes, implying that when we have removed the three best estimators, there exists no further realized measures that significantly beats the RCov_{5min}[29]. Conclusively only via the removal of the three best estimators under the Hansen (2005) SPA test, we can find evidence that the RCov_{5min} is significantly beaten in terms of average performance measured by lower losses by the three best estimators (being ThreshCov_{1min}, ThreshCov_{5min} and MRC_{1sec}). This is also warranted in Liu, Patton & Sheppard (2015), where they argue that the univariate counterpart is beaten by a small portion of more complex realized measures including noise robust estimators (see conclusion of the above article), when using the multiple forecast comparison test called the Model Confidence Set by Hansen et al. (2011).

Table 6: Shows the Superior Predictive Ability test of Hansen (2005) using the RCov_{5min} as the benchmark forecast model, with $B = 10000$ bootstrap replications, where we have removed the three best estimators.

Benchmark: RCov_{5min}			
Universe: Ω_0	SPA_L	SPA_C	SPA_U
	0.0560	0.0820	0.1490
Universe: Ω_1	SPA_L	SPA_C	SPA_U
	0.0540	0.0750	0.1170
Universe: Ω_2	SPA_L	SPA_C	SPA_U
	0.0938	0.0938	0.1634

To better understand the characteristics of a "good" realized estimator, we also present results of pairwise comparisons of the realized measures using the test of Diebold & Mariano (1995) and West (1996). Furthermore we isolate the best cases for each realized measure in terms of their average MQLIKE loss computed in

[27] Including jump robust estimators they eventually tested 648 realized estimators.

[28] Although the test utilizes the difference of the losses for each day (see equation (21)), we can indirectly get some insight into how the benchmark would perform, by comparing the average loss of the benchmark to the average loss of the alternative forecasts.

[29] We have tried removing every combination of the six best estimators, but it was only the removal of the three best estimators that yielded an acceptance of the null hypothesis. We are well aware that this can be considered data-mining and is not a viable option when considering a large universe of alternative forecast models.

Table 2[30]. We sub-categorize each of the best realized measures into pre-averaged versions and their non noise-robust counterparts, the jump-robust and the non-jump robust measures and last the "Best of: Jump robust & pre-averaging". The first dimension is to see whether there is any statistical significant difference between the losses for the pre-averaged vs non pre-averaged measures under each sampling scheme. The next dimension has the same purpose for jump robust versions, and the last compares the two realized measures of the same category. We find that it is a reasonable act only to consider the estimators of subset Ω_2, and thus compare the best realized measures for each class. As seen in Table 5 we have computed the Diebold-Mariano West test across Ω_2 and considered every possible combination within the set. In all, with the exception of one, pairwise comparisons we reject the null hypothesis of equal predictive ability among the forecast models and solely based on the test, we should choose the forecast model with the lowest average loss (being ThreshCov_{1min})[31]. However, we can not reject the null hypothesis of equal predictive accuracy for the RCov_{1min} and the BPCov*_{1sec} realized measures. Moreover based on the Diebold-Mariano test, we can not find any evidence that the noise-robust estimators perform vastly better than the non noise-robust measures on minute frequencies, however based on Table 2 we observe that the MRC_{1sec} ranks the third best estimator after ThreshCov_{5min} and ThreshCov_{1min}, implying that there is no reason of excluding noise-robust estimators due to the "lack" in performance.

In conclusion, we can not find any statistical evidence that ThreshCov_{1min} is outperformed by any of the other alternative forecast models. Thus throughout the next sections our main focus will revolve around the use of this realized measure in contrast to any other measure when needed.

6. An application to the Risk Parity portfolio framework

In the recent years there have been a rise in the attraction of conditioning portfolio choice on volatility, so the portfolio effectively targets a constant level of volatility rather than a constant level of notional exposure. Countering the fluctuations in volatility is achieved by leveraging the portfolio at times of low volatility and downscaling it in times of high volatility. In the recent work of Moreira & Muir (2017) they find that volatility-managed portfolios increases the Sharpe ratio in the case of the equity market and a number of dynamic, mostly long-short stock strategies. This is further strengthened by the recent article of Hoyle & Sheppard (2018) studying the econometric properties of the dynamic risk parity and its impact of the sharpe ratio, and fills out the necessary and sufficient theoretical conditions for these results to hold. Although most of the research has concentrated on the equity markets, we investigate the volatility targeting approach on the two assets SPY and TLT, with TLT being composed of 20+ year treasury bonds. Moreover we investigate how well each portfolio performs in terms of achieving the constant level of volatility together with the marginal contribution of the correlation to the portfolio volatility. A portfolio closer to the constant level of volatility with a realized measure estimated on fewer days than the norm[32] is more favourable, since it performs better in mitigating the effects of a drastic sign change in the correlation from one day to another, thus proving its stability. Nevertheless we will not address how to deploy methods in order to mitigate the effects of trading costs on the performance of the volatility targeted risk parity portfolio[33].

6.1. Portfolio setup

We consider the Equally-weighted Risk contribution (ERC) portfolio, which is also known as the traditional risk parity portfolio and is constructed on the idea that no asset contributes more than its peers to the total portfolio volatility. Therefore the ERC portfolio mimic the diversification effect of equally-weighted portfolios while taking into account single and joint risk contributions of the assets.

If we consider $x = [x_{1t}, x_{2t}]$ of 2 risky assets. Then the portfolio volatility can be decomposed as,

$$\Sigma_t(x) = \sqrt{x_t^T \Sigma_t x_t} = \sqrt{\sum_{i=1}^{2} x_{it}^2 \sigma_{it}^2 + 2 \cdot x_{1t} x_{2t} \sigma_{12,t}}, \tag{44}$$

with $\sigma_{12,t} = \sigma_{1t}\sigma_{2t}\rho_t$ being the covariance of asset 1 and 2 and Σ_t being the covariance matrix. Using the above decomposition we can derive the contribution of each asset

[30] The same set as Ω_2.

[31] To alleviate confusion, we should only choose the model with lowest loss if the comparison analysis was solely based on the Diebold-Mariano test. Some consideration should be taken into account if the forecast comparison analysis consisted of a variety of pairwise and multiple forecast comparison analyses.

[32] The standard estimation window varies for each fund and depends on the frequency of the available data. However, it is quite common to use a rolling one month estimation window on daily data (see eg. Harvey et al. (2018)).

[33] That could be reported as a separate topic.

x_{it} to the portfolio volatility:

$$\frac{\partial \Sigma_t(x)}{\partial x_{it}} = \frac{\partial}{\partial x_{it}} \left(\sqrt{\sum_{i,j=1}^{2} x_{it}^2 \sigma_{it}^2 + 2 \cdot x_{it} x_{jt} \sigma_{ij,t}} \right) \tag{45}$$

$$= \frac{1}{2} \cdot \Sigma_t(x)^{-1} \cdot (2x_{it}\sigma_{it}^2 + 2x_{jt}\sigma_{ij,t})$$

$$= \frac{x_{it}\sigma_{it}^2 + x_{jt}\sigma_{ij,t}}{\Sigma_t(x)}, \quad \text{for } i \neq j.$$

Moreover, since equation (45) is the marginal contribution of each asset, then it must hold that the portfolio volatility can be further decomposed as:

$$\Sigma_t(x) = \sum_{i=1}^{2} x_{it} \cdot \frac{\partial \Sigma_t(x)}{\partial x_{it}}.$$

Thus the risk of the portfolio can be seen as the sum of the asset risk contributions[34]. For the bivariate asset dimension the ERC portfolio omits a closed form solution. If we let $x_t = (w_t, 1 - w_t)$ be a vector of weights, then the vectorized asset risk contribution can be formulated as:

$$\frac{1}{\Sigma_t(x)} \begin{pmatrix} w_t^2\sigma_{1t}^2 + w_t(1-w_t)\rho_t\sigma_{1t}\sigma_{2t} \\ (1-w_t)^2\sigma_{2t}^2 + w_t(1-w_t)\rho_t\sigma_{1t}\sigma_{2t} \end{pmatrix} \tag{46}$$

The way to find the optimal allocation scheme is to equalize the rows in the above matrix, which can be reduced to solving the equation $w_t^2\sigma_{1t}^2 = (1-w_t)^2\sigma_{2t}^2$. The unique solution is given by:

$$x_t^* = \left(\frac{\sigma_{1t}^{-1}}{\sigma_{1t}^{-1} + \sigma_{2t}^{-1}}, \frac{\sigma_{2t}^{-1}}{\sigma_{1t}^{-1} + \sigma_{2t}^{-1}} \right), \tag{47}$$

under the constraints, $0 \leq w_t \leq 1$, $\sum_{i=1}^{2} x_{it} = 1$ and $x_{it} \cdot \frac{\partial \Sigma_t(x)}{\partial x_{it}} = x_{jt} \cdot \frac{\partial \Sigma_t(x)}{\partial x_{jt}}$ for all i, j. A short-sale constraint is a plausible assumption since many investors can not take short positions due to the massive scale of the portfolio[35]. Even-though one should always consider highly liquid assets in this setup, there also exists hidden additional costs of short-selling that makes it particularly unfavourable. In order for the ERC portfolio to achieve a constant level of volatility, we rescale the portfolio weights with a leverage parameter. The leverage parameter is a scalar calculated by dividing a risk-target with the portfolio volatility. Conclusively the Risk-Targeted Equally-weighted Risk Contribution (RTERC) portfolio is stated as follows. Let $x_t = (w_1^*, w_2^*)$ be the optimal weight allocation found in equation (47), then we define the leverage parameter as:

$$\alpha_t = \frac{RT}{\sqrt{x_t^T \Sigma_t x_t}}, \tag{48}$$

with RT being the risk-target[36]. Furthermore the new weights for the RTERC portfolio can be found by rescaling the old weights with the leverage parameter

$$x_t^* = \alpha_t \cdot x_t. \tag{49}$$

Using the chain rule thus performing the same calculations as in (45), we get the portfolio volatility's sensitivity to the correlation between the two assets,

$$\frac{\partial \Sigma_t(x)}{\partial \rho_t} = \frac{x_{1t}x_{2t}\sigma_{1t}\sigma_{2t}}{\Sigma_t(x)}, \tag{50}$$

which can be used to further investigate the role of the stock bond correlation on portfolio volatility.

6.2. *RTERC portfolio in practice:*

We present a small study on the performance of the RTERC portfolio using the Threshold realized Covariance estimator with a one minute sampling scheme and comparing it to a Realized Covariance estimator calculated on daily data.

In Table 7 we present the "performance" statistics for the different lengths and methods of rebalancing. The first method is a simple rebalancing scheme, where we rebalance every x week and utilizing a lookback period of x weeks. The other method is a rolling x week/month calculation of the realized measure where the x weeks/month denotes the lookback period of the calculation and it moves in a fixed window. We give a description of each column in Table 7. From left to right: The average distance column denotes the average absolute distance the portfolio had over the entire time period to the risk-target. The Vol of Vol gives us the standard deviation of the portfolio risk, which helps us determine the stability of the portfolio over time. In an optimal situation, a portfolio with a lower vol of

[34] This is easy to verify using vector notation since the marginal risk contribution is given as, $\frac{\Sigma_t x_t}{\sqrt{x_t^T \Sigma_t x_t}}$ and thus the asset risk contribution is $x_t^T \frac{\Sigma_t x_t}{\sqrt{x_t^T \Sigma_t x_t}} = \sqrt{x_t^T \Sigma_t x_t} = \Sigma_t(x)$.

[35] A huge short position could imply a negative (over)reaction in the equity/bond market thus amounting to an artificial decrease of the prices. This has a further effect on the weight allocation, since the weights was chosen before the market fall, and thus the current allocation might lead to losses in the portfolio aswell.

[36] The volatility scaling is incorporated into the risk parity portfolio in equation (47), whereas the leverage parameter rescales the volatility once again to match the risk-target. For any other portfolio the volatility scaling is constructed on the return data, using the equation (or abbreviations of) $r_t = r_t \cdot \frac{\sigma^{\text{target}}}{\sigma_{t-k}}$, for $k \geq 1$ being arbitrary chosen (see eg. Harvey et al. (2018)).

Table 7: Shows the performance statistics of the RTERC portfolios calculated using two different methods. Moreover we present the average absolute distance to the risk target, the volatility of the portfolio risk and the mean shortfalls calculated using the historical method with $\alpha = (0.01, 0.05)$, giving us the 99%- and 95% confidence interval. The colorcoding is in terms of the standard deviation of the portfolio volatility, where a vol of vol higher than 0.70% results in yellow and 2% results in red.

Rebalanced & Rolling	Avg. Dist (%)	Vol of Vol (%)	Left Tail (%)	
			1%	5%
Rebalanced 1 Week	0.7993	3.9396	-0.8416	-0.5622
Rebalanced 3 Weeks	0.5585	0.7059	-0.5097	-0.3344
Rebalanced 5 Weeks	0.5540	0.6302	-0.4031	-0.2625
Rolling 1 week	0.8032	1.0514	-1.04848	-0.6712
Rolling 3 weeks	0.5424	0.6977	-0.6407	-0.3985
Rolling 5 weeks	0.9544	1.3721	-0.8118	-0.4653
Benchmark:				
Rebalanced 1 week	3.0030	3.9396	-8.9817	-7.9930
Rebalanced 1 month	2.9895	1.4948	-5.3418	-4.4615
Rolling 1 week	4.06986	3.8952	-21.8131	-11.8516
Rollling 1 month	2.6674	1.5092	-9.7243	-6.3657

vol will be closer to the risk-target[37]. Furthermore the Expected shortfall for a 95% and 99% confidence interval have been produced, with the only purpose of observing the reduction in extreme losses when using a realized measure with a one minute frequency. We observe that using a rolling estimation window longer than 3 weeks for ThreshCov_{1min} implies higher overall performance statistics. Utilizing a longer lookback window would result in a smaller impact of the new information being build into the newly calculated covariance estimate. Moreover this argumentation is further strengthened by the study of Harvey et al. (2018) where they show the performance of volatility-scaled equity and bond returns worsens for higher estimation windows[38] and argue that an estimation window around 20 days is reasonable. The same can not be said about the simple static rebalancing scheme which seems to improve with larger lookback windows.

The most important observations to be made is that no matter the use of weight updating scheme, the ThreshCov_{1min} outperforms the benchmarks in terms of stability (lower vol of vol) together with a better average distance to the risk-target. The portfolio volatility for the 1 week rolling weight updating schemes can be seen in Figure 4, where the lower vol of vol for the high frequency estimate implies a better risk-targeting portfolio[39]. From the correlation graph we can deduce the sensitivity of the portfolio volatilities due to changes in the correlation estimate. It is observed that a large change in the correlation estimate tend to drastically increase the portfolio volatility for the benchmark, whereas the high frequency portfolio construction does not suffer much against the variability of the correlation estimate.

Furthermore in Figure 5 we have calculated the portfolio volatility's sensitivity to the correlation estimates using equation (50), where we have fixed the average annualized asset volatilities for each asset, which became 0.06 and 0.11 for TLT and SPY respectively. It is notable that when ρ converges towards -1 the portfolio volatility increases to as much as 20% and the reverse is also true. Moreover we can conclude that the RTERC portfolio is less sensitive to correlation increases compared to drastic decreases. Therefore in periods of volatile correlation estimates it is sensible for the portfolio manager to reduce the portfolio size due to the increase in volatility. Moreover in the Appendix (A) we further compare the best performing rolling scheme for the Threshold covariance estimator against the best performing rolling scheme for the benchmark together with the 1 week rebalancing weight frequency for both portfolio constructions.

From Table 7 it is further seen that the low vol of vol further reduces the probability of very negative returns (left-tail events) which is valuable for risk averse institutional investors. Therefore it is safe to say that utilizing high frequency covariance measures as opposed to daily estimators will provide some benefits in terms of better portfolio management.

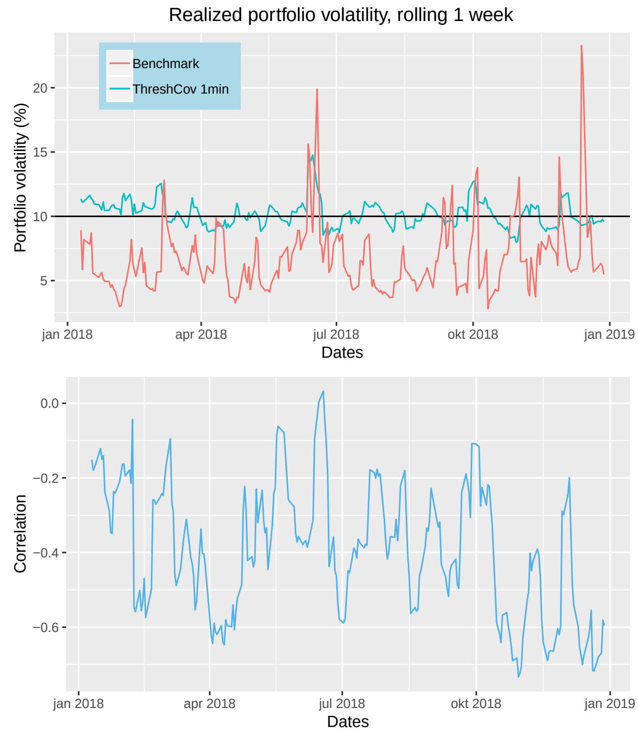

Figure 4: Shows the portfolio volatility for the rolling 1 week evalutation frequency of the RTERC using the ThreshCov_{1min} versus RCov_{daily}. Adding to the illustration we have graphed the correlation calculated from the proxy (being the "true" correlation estimate) using a one week rolling on five minute sampling scheme.

[37]If a portfolio has a low vol of vol but a high average distance, it implies the realized measure finds suboptimal portfolio weights. This was studied using BPCov^*_{15min} on a rolling 3 weeks updating scheme which yielded a portfolio risk that overshots the risk-target a high proportion of times. A graphical illustration is found in the Appendix.

[38]They use rolling exponentially decaying weights with different days of half life.

[39]Note that the figure's sole purpose is to aid Table 7 with a graphical illustration of the portfolio differences when using measures on high/low sampling frequencies. It is fairly unreasonable to use a weight updating scheme with a lookback period of 1 week when dealing with daily data, ie. calculating the covariance estimate using 5 data points.

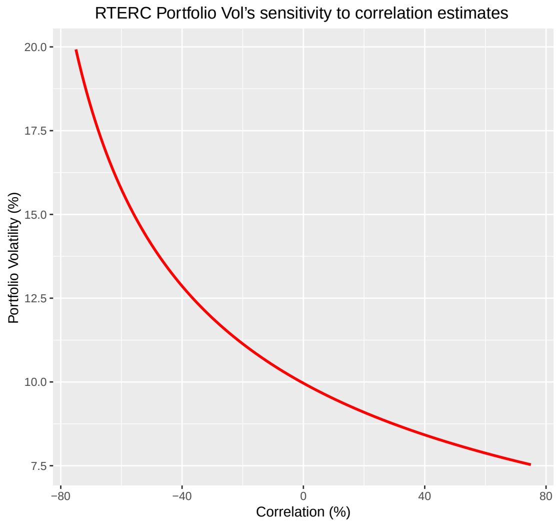

Figure 5: Shows the RTERC portfolio volatility's sensitivity to different correlation estimates. We have used fixed annualized asset volatilities of $0.06,\ 0.11$ for TLT and SPY respectively, which was found by averaging (and annualizing) the volatilities of each asset over the entire time-period. Moreover we have limited the range of the correlation, since it is higly unlikely to see a correlation of ± 1 in the market. In 2018 the range of the correlation was $\rho \in [-0.81, 0.2]$ using a rolling scheme with a lookback period of 1 month on daily data.

7. Concluding Remarks & directions of further research

Motivated by a large body of research on estimators of asset price volatility using high frequency data (our so-called "realized measures"), this article considers comparing the empirical accuracy of a collection of realized measures in a multivariate setup to further be used in a Risk Targeted Equally-weighted Risk Contribution portfolio (RTERC). In total, we consider 28 different estimators on 250 days of high frequency data with two different asset classes being equities and bonds. Furthermore we used the cleaning procedure described in the article of Barndorff-Nielsen et al. (2008) to further alleviate of outliers and invalid transaction data. Our main findings for the first part can be summarized as follows. Using the leading RCov_{5min} as a proxy for the true covariation process, we find evidence that the RCov_{5min} is significantly outperformed by ThreshCov_{1min}, ThreshCov_{5min} and MRC_{1sec}, using the testing approach of Hansen (2005). However when considering the same test for ThreshCov_{1min} we failed to reject the null hypothesis and conclusively the realized measure is considered non-inferior in contrast to the alternative forecasts. Moreover the Diebold-Mariano West test shows us that we could not accept the null hypothesis of equal predictive accuracy between ThreshCov_{1min} and the best realized estimators from Ω_2. Therefore as a rule of thumb, the estimator with the lowest average loss should be considered, which in this case is ThreshCov_{1min}.

In terms of Table 2 we also find evidence that the RCov_{1min} performs slightly better than the 5 minute counterpart, and the noise-reducing procedure (the pre-averaging) of the noise-robust estimators does not provide any statistically significant increase in performance. This could be due to the TLT containing almost no noise as compared to SPY, which in terms provide better results for minute frequencies non-noise robust estimators (especially at the 1 minute frequency). It is important to acknowledge here that since we consider a very small collection of assets, which share the same characteristic (being liquid assets on well-developed markets), our conclusion requires some adjustment before considering them for other assets. Furthermore due to the short time period being 1 year (250 days), there will be some inconsistency risk in terms of how well the best realized measures might perform next year, which could be eliminated if we considered a longer time-frame, say a 10 year period. However we can provide two general conclusions. First, sampling using a minute frequency seems to provide much of the benefits of high frequency data without exposing the measure to problems from microstructure noise. This can be seen as the RCov_{1min} and ThreshCov_{1min} provides a high empirical accuracy close to the performance of the noise-robust measures. Finally, the gains from high frequency data are greatest when microstructure noise is relatively low and when volatility is high. Thus more data will help capture the variability of the asset volatilities better, which in terms will lead to better weight allocations for different portfolio schemes (not just RTERC).

Of course the above described tests compare a set of competing realized measures with a given benchmark measure. The RCov_{5min} measure is a reasonable and widely used benchmark estimator, but one might also be interested in determining whether maintaining the estimator as the "proxy" gives it excessive preferential treatment. In order to address this question, one could use Hansen et al. (2011) Model Confidence Set (MCS), which given a set of competing realized measures identifies a subset that contains the unknown best estimator with some specific level of confidence, with the other realized measures in the MCS not being significantly different from the "true" best realized measure.

After the empirical study on the best performing realized estimator we further deducted a portfolio analysis on the RTERC portfolio where we analysed the "performance" when using the high frequency measure in contrast to the standard daily covariance estimator. We show that volatility targeting/scaling using the a 3 week (5 week) rolling weight updating scheme (simple rebalancing scheme) has one unambiguous effect: It reduces the likelihood of very negative returns and also the vol of vol in contrast to the benchmarks constructed using daily estimates. Moreover the RTERC portfolio constructed using high frequency estimates seems to be more stable to drastically changes in the correlation estimates, whereas the benchmark portfolio model suffers from spikes in the portfolio volatility. Since the improvement between the two portfolios are already visible with a one week lookback window, it further tells us that any institutional investor can use a shorter estimation window to "efficiently" estimate the covolatility, which in terms would lead to a quicker weight updating scheme that responds to drastic changes in the correlation estimates.

Bibliography:

[1] Delbaen, F & W. Schachermayer. (2006). "*The Mathematics of Arbitrage*". Springer-Verlag Berlin Heidelberg, Springer Finance (2006), DOI: 10.1007/978-3-540-31299-4.

[2] Roll, R. (1984). "*A simple model of the implicit bid–ask spread in an efficient market* ". Journal of Finance 39 1127–1139.

[3] Harris, L. (1991). "*Stock Price Clustering and Discreteness*". The Review of Financial Studies 1991 Volume 4, number 3, pp. 389-415.

[4] Black, F. (1996). "*Noise*". The Journal of Finance July (1996) Volume 41 Issue 3, pp. 528-543.

[5] Harris, L. (1990). "*Estimation of stock price variances and serial covariances from discrete observations*". Journal of Financial and Quantitative Analysis 25 291–306.

[6] Gottlieb, G. and A. Kalay. (1985). "*Implications of the discreteness of observed stock prices*". Journal of Finance 40 pp. 135–153.

[7] Hansen, P.R., A. Lunde (2006). "*Realized Variance and Market Microstructure Noise*". Journal of Business & Economic Statistics, Vol. 24, No. 2 (Apr., 2006), pp. 127-161.

[8] Andersen, T.G., T. Bollerslev & F.X. Diebold (2007). "*Roughing it up: Including Jump components in the measurement, modeling and forecasting of return volatility*". The Review of Economics and Statistics 2007 89:4, pp. 701-720.

[9] Barndorff-Nielsen, O.E., N. Shephard (2005). "*Variation, jumps, market frictions and high frequency data in financial econometrics*". Nuffield College Economics Working Paper No. 2005-W16.

[10] Barndorff-Nielsen, O.E., N. Shephard (2002). "*Econometric analysis of realised volatility and its use in estimating stochastic volatility models*". Journal of Royal Statistical society, Statistical methodology Series B, Volume64, Issue2, May (2002) pp. 253-280.

[11] Barndorff-Nielsen, O.E., N. Shephard (2004). "*Econometric Analysis of Realized Covariation: Highfrequency Based Covariance, Regression, and Correlation in Financial Economics*". Econometrica, Vol. 72, No. 3 (May, 2004), pp. 885–925.

[12] Podolskij, M., M. Vetter (2009a)."*Estimation of volatility functionals in the simultaneous presence of microstructure noise and jumps*". Bernoulli 15 (2009), no. 3, 634–658. doi:10.3150/08-BEJ167.

[13] Christensen, K., S. Kinnebrock, M. Podolskij (2010). "*Preaveraging estimators of the ex-post covariance matrix in noisy diffussion models with non-synchronous data*". Journal of Econometrics 159 (2010), pp. 116 - 133.

[14] Aït-Sahalia, Y., P.A. Mykland, L. Zhang (2005)."*How Often to Sample a Continuous-Time Process in the Presence of Market Microstructure Noise*". The Review of Financial Studies Vol. 18, No. 2. doi:10.1093/rfs/hhi016.

[15] Gatheral, J., R.C.A. Oomen (2010) "*Zero-intelligence realized variance estimation* ". Finance Stoch (2010) 14: pp. 249–283 DOI 10.1007/s00780-009-0120-1.

[16] Barndorff-Nielsen, O.E., N. Shepard (2004). "*Measuring the impact of jumps in multivariate price processes using bipower covariation*". Working paper, Nuffield College, Oxford University.

[17] Christensen, K., R.C.A. Oomen, M. Podolskij (2014). "*Fact or friction: Jumps at ultra high frequency*". Journal of Financial Economics 114 (2014) pp. 576-599.

[18] Gobbi, F., C. Mancini (2012). "*Identifying the Brownian Covariation from the Co-jumps given discrete observations*". Econometric Theory, 28, 2012, pp. 249–273. doi:10.1017/S0266466611000326.

[19] Jacod, J., V. Todorov (2009). "*Testing for common arrivals of jumps for discretely observed multidimensional processes*". The Annals of Statistics 2009, Vol. 37, No. 4, pp. 1792–1838. DOI: 10.1214/08-AOS624.

[20] Barndorff-Nielsen, O.E., N. Shephard (2006). "*Econometrics of Testing for Jumps in Financial Economics Using Bipower Variation*". Journal of Financial Econometrics, 2006, Vol. 4, No. 1, pp. 1–30.

[21] Hautsch, N., M. Podolskij (2013). "*Preaveraging-Based Estimation of Quadratic Variation in the Presence of Noise and Jumps: Theory, Implementation, and Empirical Evidence*". Journal of Business & Economic Statistics, 31:2, 165-183, DOI: 10.1080/07350015.2012.754313

[22] Barndorff-Nielsen, O.E., P.R. Hansen, A. Lunde, N. Shephard (2008). "*Multivariate realised kernels: Consistent positive semidefinite estimators of the covariation of equity prices with noise and non-synchronous trading*". Journal of Econometrics Volume 162, Issue 2, June 2011, pp. 149-169.

[23] Christensen, K., R.C.A. Oomen, M. Podolskij (2010). "*Realised quantile-based estimation of the integrated variance*". Journal of Econometrics 159 (2010) pp.74–98.

[24] Barndorff-Nielsen, O.E, P.R. Hansen, A. Lunde, N. Shephard (2008). "*Realised Kernels in Practice: Trades and Quotes*". Econometrics Journal (2008) volume 4, pp. 1 - 33.

[25] Patton, A.J. (2011a). "*Data-based ranking of realised volatility estimators*". Journal of Econometrics 161 (2011) pp. 284-303.

[26] Patton, A.J. (2011b). "*Volatility forecast comparison using imperfect volatility proxies*". Journal of Econometrics 160 (2011) pp. 246–256.

[27] James, W., C. Stein (1961). "*Estimation of Quadratic loss*". Proc. Fourth Berkeley Symp. on Math. Statist. and Prob., Vol. 1 (Univ. of Calif. Press, 1961), pp. 361-379.

[28] Jacod, J., Y. Li, P.A. Mykland, M. Podolskij, M. Vetter (2009). "*Microstructure noise in the continuous case: The pre-averaging approach*". Stochastic Processes and their Application 119 (2009) pp. 2249-2276.

[29] Christensen, K., U. Hounyo, M. Podolskij (2018). "*Is the diurnal pattern sufficient to explain intraday variation in volatility? A nonparametric approach*". Journal of Econometrics 205 (2018) pp. 336 - 362.

[30] Vetter, M., M. Podolskij (2009). "*Bipower-type estimation in a noisy diffusion setting*". Stochastic Processes and their Applications 119 (2009) pp. 2803–2831.

[31] Politis, D.N., H. White (2004). "*Automatic Block-Length Selection for the Dependent Bootstrap*". ECONOMETRIC REVIEWS Vol. 23, No. 1, pp. 53–70, 2004.

[32] Politis, D.N., J.P. Romano (1995). "*Bias-corrected nonparametric spectral estimation* ". JOURNAL OF TIME SERIES ANALYSIS Vol. 16, No. 1 pp. 67 - 101.

[33] Diebold, F.X., R.S. Mariano (1995). "*Comparing Predictive Accuracy*". Journal of Business & Economic Statistics, 13:3, pp. 253-263, DOI: 10.1080/07350015.1995.10524599.

[34] West, K.D. (1996). "*Asymptotic Inference about Predictive Ability*". Econometrica Vol. 64, No. 5 (Sep., 1996), pp. 1067-1084 (18 pages).

[35] Podolskij, M., M. Vetter (2009b). "*Understanding limit theorems*

for semimartingales: a short survey". Statistica Neerlandica (2010) Vol. 64, nr. 3, pp. 329–351. doi:10.1111/j.1467-9574.2010.00460.x.

[36] Patton, A.J., L.Y. Liu, K. Sheppard (2015)."*Does anything beat 5-minute RV? A comparison of realized measures across multiple asset classes*". Journal of Econometrics Volume 187, Issue 1, July 2015, pp. 293-311.

[37] Podolskij, M., M. Vetter. (2010). "*Understanding limit theorems for semimartingales: a short survey*". Volume64, Special Issue: Statistical Inference for Lévy Processes with Applications to Finance, August 2010, pp. 329-351.

[38] Tauchen, G., H. Zhou (2011). "*Realized jumps on financial markets and predicting credit spreads*". Journal of Econometrics Volume 160, Issue 1, January 2011, pp. 102-118.

[39] Andersen, T.G., T. Bollerslev, X. Huang (2007). "*A Reduced Form Framework for Modeling Volatility of Speculative Prices based on Realized Variation Measures*". Journal of Econometrics Volume 160, Issue 1, January 2011, pp. 176-189.

[40] Moreira, A., T. Muir (2017). "*Volatility-Managed Portfolios*". Journal of Finance, Volume 72, Issue 4, August 2017, pp. 1611-1644.

[41] Hoyle, E., N. Sheppard (2018). "*Volatility Scaling's Impact on the Sharpe Ratio*". Available at SSRN: https://ssrn.com/abstract=3279787 or http://dx.doi.org/10.2139/ssrn.3279787

[42] Politis, D.N, J.P. Romano (1994). "*The Stationary Bootstrap*". Journal of the American Statistical Association Vol. 89, No. 428 (Dec., 1994), pp. 1303-1313.

[43] Harvey, C.R., E. Hoyle, R. Korgaonkar, S. Rattray, M. Sargaison, O.V. Hemert (2018). "*The Impact of Volatility Targeting*". Journal of portfolio management (JPM) 2018, 45 (1) pp. 14 - 33.

[44] Hansen, P.R. (2005). "*A Test for Superior Predictive Ability*". Journal of Business & Economic Statistics Vol. 23, No. 4 (Oct., 2005), pp. 365-380 (16 pages)

[45] Hansen, P.R., A. Lunde, J.M. Nason, (2011). "*The Model Confidence Set*". Econometrica Vol. 79, No. 2 (March, 2011), pp. 453-497 (45 pages).

[46] Yang, J., Y. Zhou, Z. Wang, (2009). "*The stock–bond correlation and macroeconomic conditions: One and a halfcenturies of evidence*". Journal of Banking & Finance Volume 33, Issue 4, April 2009, pp. 670-680.

[47] White, H., (2000). " *A Reality Check For Data Snooping*". Econometrica, Vol. 68, No. 5 September, 2000 , pp. 1097 - 1126.

[48] Hansen, P.R., A. Lunde, (2005). "*A forecast comparison of volatility models: does anything beat a GARCH(1,1)?*". Journal of Applied Econometrics, Volume 20, Issue 7, December 2005, pp. 873-889.

[49] Carr, P., H. Geman, D. Madan, M. Yor, (2002). "*The Fine Structure of Asset Returns: An Empirical Investigation*". Journal of Business, 2002, vol. 75, no. 2 pp. 305 - 332.

[50] OECD (2018)."*OECD Economic Outlook*". Volume 2018 Issue 2, OECD Publishing, Paris. https://doi.org/10.1787/eco_outlook-v2018-2-en

[51] Campbell, J.Y, C. Pflueger, L.M Viceira, (2019). "*Macroeconomic Drivers of Bond and Equity Risks*". NBER Working Paper No. 20070 April 2014, Revised August 2018.

Appendix A Comparison of rebalancing schemes.

Figure 6 helps us to graphically illustrate the average distance and vol of vol for two weight updating schemes. It is observed that the portfolio calculated on the 3 week rolling high frequency covariance estimate is more stable than the benchmark. The same argumentation can be applied to figure (B).

Figure 6: Figure (A) gives us the rolling 3 weeks portfolio volatility calculated using the Threshold covariance estimator on 1 minute high frequency data, against the monthly rolling realized covariance estimator on daily stock data. Figure (B) shows the difference between the simple rebalancing scheme between the portfolio calculated using the high frequency covariance estimate against the benchmark.

Appendix B Portfolio volatility for the Pre-averaged Bipower realized Covariation process on a 15 minute sampling scheme

As described in a footnote, the average distance to the risk-target is quite big while the vol of vol is relatively low for the BPCov^*_{15min} realized measure. Moreover it seems that some bias is introduced into BPCov^*_{15min} when RCov_{5min} is used as a proxy for the true covariation process, which is observable in the graphical illustration below.

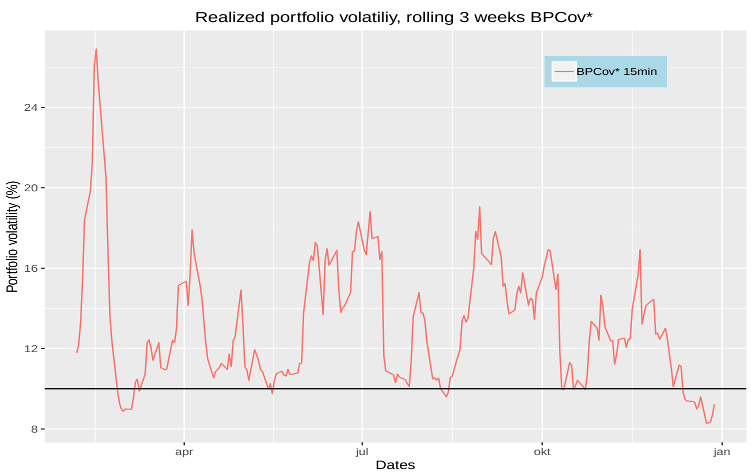

Figure 7: Shows the rolling 3 weeks portfolio volatility using the Pre-averaged Bipower Realized Covariation process on a 15 minute sampling scheme. It is reasonable to assume that the portfolio volatility would have hit the 10% vol-target more often, if we used a longer time period than 250 days.