Cointegrated portfolios and volatility modeling in the cryptocurrency market

Stefan Gabriel[a], Robert M. Kunst[b,*]

[a] *University of Vienna, Department of Finance, Vienna, Austria.*
[b] *Institute for Advanced Studies and University of Vienna, Vienna, Austria*

Abstract

We examine two major topics in the field of cryptocurrencies. On the one hand, we investigate possible long-run equilibrium relationships among ten major cryptocurrencies by applying two different cointegration tests. This analysis aims at constructing cointegrated portfolios that enable statistical arbitrage. Moreover, we find evidence for a connection between market volatility and the spread used for trading. The results of the trading strategies suggest that cointegrated portfolios based on the Johansen procedure generate the highest abnormal log-returns, both in-sample and out-of-sample. Five out of six trading strategies generate a positive overall profit and outperform a passive investment approach out-of-sample.

The second part of the econometric analysis explores Granger causality between volatility and the spread. For this analysis, we implement two types of forecasting models for Bitcoin volatility: the GARCH (generalized autoregressive conditional heteroskedasticity) family using daily price data and the HAR (Heterogeneous AutoRegressive) model family based on 5-min high-frequency data. In both categories, we also consider potential jumps in the price series, as we found that price jumps play an important role in Bitcoin volatility forecasts. The findings indicate that the realized GARCH model is the only GARCH model that can compete against the HAR-RV (Heterogeneous Autoregressive Realized Volatility) model in out-of-sample forecasting.

Keywords: cryptocurrencies, bitcoin volatility, realized variance, jump variation, cointegrated portfolios, statistical arbitrage
JEL Code: C22, C52, C53

*Robert M. Kunst, Research Group Macroeconomics and Business Cycles, Institute for Advanced Studies, Vienna, Austria, and Department of Economics, University of Vienna, Vienna, Austria; E-mail address: kunst@ihs.ac.at, robert.kunst@univie.ac.at

March 11, 2024

1. Introduction

Cryptocurrencies, especially Bitcoin (BTC), have become very popular in recent years. The market has been growing enormously, and more and more cryptocurrencies have emerged. The total market capitalization has increased from USD 5.5 billion on 1st Jan 2015 to USD 2.2 trillion on 1st Jan 2022, which is almost a 40,000% increase (CoinMarketCap, 2022). BTC with a market cap of USD 902 billion on 1st Jan 2022 was the first cryptocurrency on the market and was developed by the pseudonymous Satoshi Nakamoto.

During the recent years, the crypto market has also become increasingly attractive for research. Cryptocurrencies are known to be highly volatile financial assets that carry a high risk of total loss. In the early days after introduction of Bitcoin, research concentrated on whether it can be considered as money. Mittal (2012) argued that Bitcoin is not money and rather resembles a commodity. Kubát (2015) investigated the same issue. He found that Bitcoin has a significantly higher volatility than gold and the EUR-USD exchange rate, in line with the findings of Mittal (2012). The bottom line is that the volatility of Bitcoin is conspicuously higher than of many other typical assets and fiat currencies. Such highly volatile financial assets have also aroused interest of hedge funds and portfolio managers, which led to a strong incentive for modeling the volatility of BTC. Using Google Trends data, Urquhart (2018) found that the realized volatility of BTC together with the trading volume has a major impact on investor attention on the following day. Catania and Grassi (2017) use a robust score driven filter for modeling the volatility of 606 cryptocurrencies. They found that a model with time-varying skewness component has the best forecasting performance.

In this project, we use crypto assets to construct trading strategies that are superior to a "buy-and-hold" strategy and are also promising in an out-of-sample backtesting analysis. Profitable trading strategies that also work during a bear market and in times with high inflation rates are crucial in risk management and for hedging. Furthermore, we examine a potential connection between market volatility and the presented trading strategies. Ni et al. (2008) found a significant link between investors' trading behavior and private volatility information in the option market. Likewise, Omane-Adjepong et al. (2019) find that

three crypto markets violate the Efficient Market Hypothesis and that accurate volatility forecasting can enhance optimal portfolio hedging. The authors also emphasize the importance of taking the high volatility persistence into account. Hence, the second objective is to evaluate the forecasting performance of the volatility of BTC by estimating different types of GARCH models with different distributions and the simple Heterogeneous Autoregressive model of Realized Volatility (HAR-RV) by Corsi (2009) including several extensions. Additionally, we focus on the incorporation of price jumps as jumps can account for a substantial portion of the variation. The goal of this analysis is to answer the question whether any GARCH model can compete with the HAR-RV type models, specifically the more recent realized GARCH model. Modeling and especially accurate forecasting of the volatility of an asset is also essential for many areas in business finance. The CME Group launched options on Micro Bitcoin futures on 28th Mar 2022 (Group, 2022).

We discuss some of the related extant literature in the next section. Section 3 introduces the data by providing some descriptive statistics and presenting relevant plots. Section 4 covers the cointegration issue including the construction of cointegrated portfolios that enable statistical arbitrage. Section 5 performs the volatility forecasting analysis by estimating sevaral GARCH models and HAR-type models. Section 6 concludes and summarizes the main results.

2. Literature review

The first section of this review explores the application of cointegration in finance, emphasizing its role in developing mean-reverting trading strategies for statistical arbitrage. For instance, Yan and Wong (2022) consider pairs trading from a game-theoretical perspective and employ continuous-time vector error-correction models (VECM) to establish statistical arbitrage strategies, including the consideration of 'delayed' cointegration. In a comparison with a time-consistent dynamic pairs trading strategy based on the Markowitz mean-variance (MV) criterion proposed by Chiu and Wong (2015), Yan and Wong (2022) show that the non-Markovian strategy can yield additional abnormal returns when in-sample data suggests a high-order vector autoregression (VAR). In the context of the crypto

market, while the literature on cointegrating relationships is less extensive compared to traditional stocks, researchers and practitioners are increasingly investigating this domain due to the market's rapid growth and popularity. Sovbetov (2018) delves into the factors affecting the prices of major cryptocurrencies, utilizing an error-correction model based on the Autoregressive Distributed Lag (ARDL) method by Pesaran et al. (2001). His results suggest that cryptomarket beta, trading volume, and volatility significantly influence prices both in the long and short run. Adebola et al. (2019) study long-term interdependencies between cryptocurrencies and the gold price, finding limited evidence of bivariate cointegrating relationships between gold and cryptocurrencies. Similarly, Tan et al. (2021) explore fractional cointegration between the value at risk of altcoins and Bitcoin, revealing substantial differences in findings between pre-crash and post-crash periods. The subsequent part of this section discusses empirical work on modeling cryptocurrency volatility.

Researchers have employed various models to forecast volatility, such as GARCH models and stochastic volatility models. Cermak (2017) predicts declining volatility trends for Bitcoin using a modified GARCH(1,1) model with macroeconomic variables, however his prediction of volatility converging with fiat currencies did not materialize. Hou et al. (2019) underscore the importance of considering jumps in Bitcoin's volatility, advocating the use of stochastic volatility models. Hung et al. (2020) propose jump-robust realized measures for improved forecasting of realized GARCH models based on intraday data. Bergsli et al. (2022) compare different GARCH models and the HAR-RV models, highlighting the superiority of the latter in forecasting Bitcoin's volatility, especially because of the usage of intraday data. Yu (2019) utilizes 5-minute high-frequency data to assess five HAR models for one-step-ahead forecasts of Bitcoin's realized volatility. His results highlighted the significance of considering a leverage effect, with the HAR-RV model that accounts for leverage performing better than the HAR-RV model that solely considers a jump component. Similarly, Shen et al. (2020) expand the standard HAR-RV model by introducing a novel specification and found it to be the most accurate forecasting model. Additionally, they demonstrate that HAR models with structural breaks outperform those without structural breaks across various forecasting horizons. Additionally, researchers

have explored machine learning techniques to enhance volatility forecasts. Bouri et al. (2021) employ random forests to evaluate the impact of the US-China trade war on Bitcoin volatility forecasts, with improved performance by including relevant external factors. Aras (2021) applies a meta-learning strategy based on support vector machines, outperforming traditional GARCH-type models. D Amato et al. (2022) utilize deep learning techniques to predict cryptocurrency volatility by relying on two different neural networks and a Self-Exciting Threshold AutoRegressive (SETAR) model. Their results indicate that the recurrent Jordan Neural Network outperforms the Non-Linear Autoregressive Neural Network and the SETAR model in terms of the mean square error (MSE).

3. The data

We analyze cointegrated relationships among ten major cryptocurrencies. Daily price data have been downloaded from Yahoo Finance. The full data set contains over two and half years of price data from 31st Dec 2019 until 31st Jul 2022. The training data that is used for estimating the cointegrating vectors and in-sample evaluation is based on the sample period from 31st Dec 2019 until 29th Apr 2022 consisting of 851 days; the test period for out-of-sample evaluation uses the last three months of the data set from 30th Apr 2022 until 31st Jul 2022. Table 1 presents some descriptive statistics of all cryptocurrencies under investigation. We use adjusted closing prices in levels rather than logs, as these admit an easier interpretation. All cryptocurrencies have rather large standard deviations in comparison to their means, which is a characteristic of high volatility. This observation makes the modeling and forecasting of volatility attractive (see section 5 for Bitcoin). With a market capitalization of $ 734.59 billion, BTC is by far the dominating cryptocurrency in the market, followed by ETH with $ 339.51 billion. The total market capitalization of the ten considered cryptocurrencies amounts to a total of $ 1.233 trillion as of 29th Apr 2022, which is almost equivalent to the GDP of Spain in 2020 (Worldbank, 2022).

Figure 1 visualizes the daily logarithmic prices (log-prices) of the training period for all ten cryptocurrencies. All time series follow a similar pattern, which suggests the existence of a common stochastic trend and hence a possible long-term relationship across

Table 1
Descriptive Statistics of all ten cryptocurrencies (Training period)

Symbol	Name	Min	Max	Mean	Std	Market cap
BTC	Bitcoin	4,970.79	67,566.83	30,919.50	18,764.50	734.59B
ETH	Ethereum	110.61	4,812.09	1,740.47	1,426.85	339.51B
BNB	Binance Coin	9.39	675.68	229.22	213.02	64.18B
ADA	Cardano	0.02	2.97	0.83	0.78	27.16B
XRP	Ripple	0.14	1.84	0.59	0.38	29.41B
DOGE	Dogecoin	0.002	0.68	0.11	0.13	17.91B
LTC	Litecoin	30.93	386.45	120.50	69.17	7.04B
BCH	Bitcoin Cash	152.22	1,542.43	422.95	205.79	5.61B
XLM	Stellar	0.03	0.73	0.22	0.15	4.41B
XMR	Monero	33.01	483.58	170.15	90.31	4.00B

Notes: Market capitalization as of 29th Apr 2022 obtained from coinmarketcap.com, denoting 10^9 (1 billion) US$

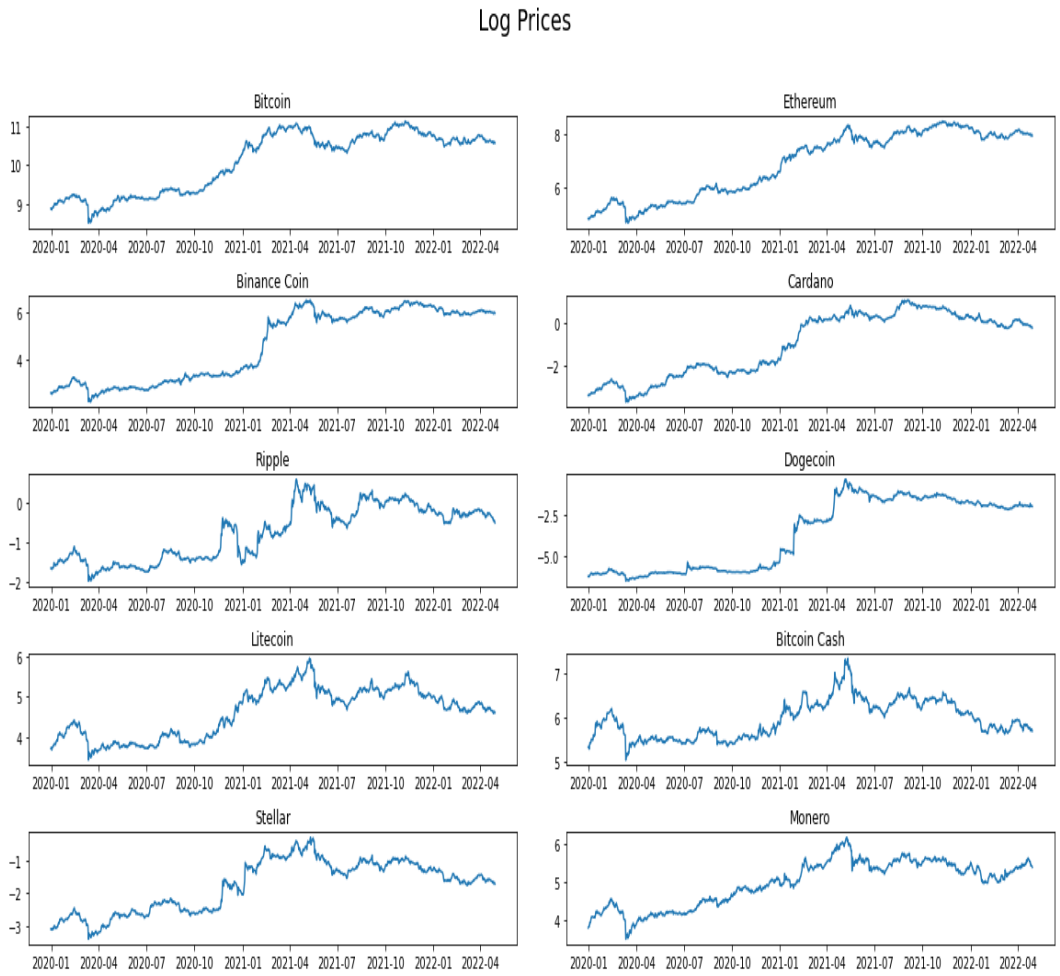

Fig. 1. Log prices of cryptocurrencies

the cryptocurrencies. To identify any potential cointegrating relationships two different cointegration tests are applied and discussed in section 4. Another noteworthy aspect is the correlation structure across returns. The level of correlation among returns of stock prices differs substantially from return correlations among cryptocurrencies. Pollet and Wilson (2010) found an average correlation of just 0.237 among the 500 largest stock pairs in the market from 1963 to 2006. In our data, however, there are five crypto pairs with a return correlation in excess of 0.8, which issue will be taken up in creating mean-reverting portfolios in section 4. Even the average correlation of the log-returns is in the upper range at 0.634. The log-returns of Dogecoin pairs have the lowest correlation.

For modeling the volatility of BTC (section 5), we use daily and intraday price data downloaded from the cryptocurrency exchange Bitstamp. The full data set contains six years of price data (1st Apr 2015 to 31st Mar 2021). For estimating the models and in-sample evaluation, we use training data for the sample period from 1st Apr 2015 until 31st Mar 2020 and an out-of-sample evaluation that covers the last year of the data set from 1st Apr 2020 until 31st Mar 2021 (COVID-19 crisis period). Figure 2 visualizes the daily log-returns of BTC for the whole sample and provides evidence for volatility clustering. There are remarkably many extreme values. The maximum daily log-return is 0.24, but the minimum is -0.49, which indicates that negative returns can be more severe. Specific statistical tools, such as normal Q-Q plots (not shown for brevity), confirm that the daily log-returns do not follow a normal distribution, as there is a lot of weight in the tails. The measured excess kurtosis may speak for the usage of a heavy-tailed distribution such as Student t when estimating GARCH models, although the marginal distribution implied by GARCH models is leptokurtic even with a normal conditional distribution (Engle (1982)). In the following, the realized variance based on 5-min high-frequency data and its stylized facts are analyzed.

Table 2 provides some descriptive statistics of the realized variance of BTC, the full sample subdivided into the training period and the test sample. The data suggest that the mean and the median of all subsamples are nearly equal. Moreover, not very surprisingly the realized variance for any subgroup is non-Gaussian, which is visible through the excess

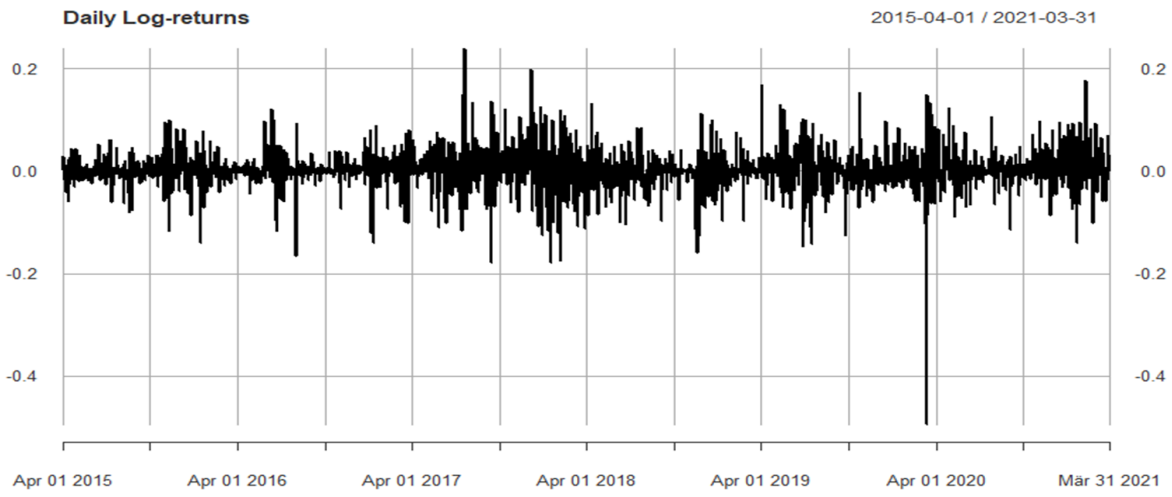

Fig. 2. Daily Log-returns of BTC

Table 2
Descriptive Statistics of the Realized Variance

	Full sample	Training period	Test period
count	2192	1827	365
mean	0.0021	0.0022	0.0017
std	0.0043	0.0046	0.0026
min	0.00003	0.00003	0.00003
25%	0.0005	0.0005	0.0005
50%	0.0009	0.0009	0.0009
75%	0.0021	0.0022	0.0018
max	0.1066	0.1066	0.0297
skewness	10.4957	10.2992	5.4818
kurtosis	189.8533	177.8116	48.0258
Jarque-Bera (p-value)	< 0.0001	< 0.0001	< 0.0001

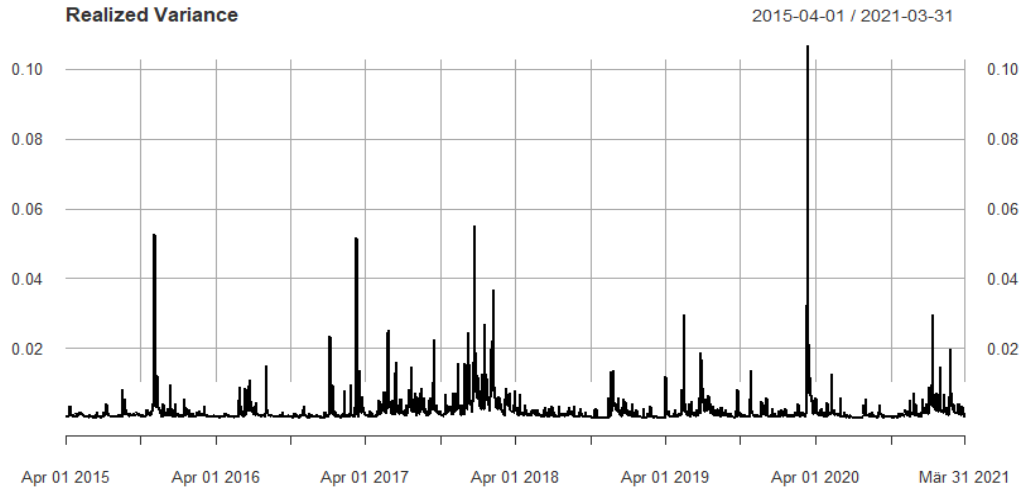

Fig. 3. Realized Variance of BTC

kurtosis and the positive skewness. The realized variance of Bitcoin follows a leptokurtic and right-skewed distribution.

Figure 3 indicates that there are huge peaks occasionally. These extreme values occur in particular at the beginning of the COVID-19 crisis, which led to considerable fear in the market and hence much volatility. An additional explanation for these extreme values could be the presence of price jumps so it makes sense to have a look at the jump variation too. The discontinuous variation of the realized variance can be estimated with any estimator that consistently estimates the integrated variance of the quadratic variation in the presence of jumps. Simply speaking, it is just defined as the difference of the realized variance and a jump-robust realized measure. Section 5.1 provides more details on these estimators.

Figure 4 shows that there are indeed many days with a relatively high jump variation. To detect significant jumps in the price process of Bitcoin the JO Jump test by Jiang and Oomen (2008) is performed. Section 5 provides further details on this test. Lastly, we want to investigate whether the realized variance of BTC follows a log-normal distribution as it is the case with many other assets. For this analysis we looked at the Q-Q-Plot and the corresponding density of the logarithmic realized variance of the full sample, of the first three years of the sample (1st Apr 2015 to 31st Mar 2018) and of the last three years (1st Apr 2018 to 31st Mar 2021).

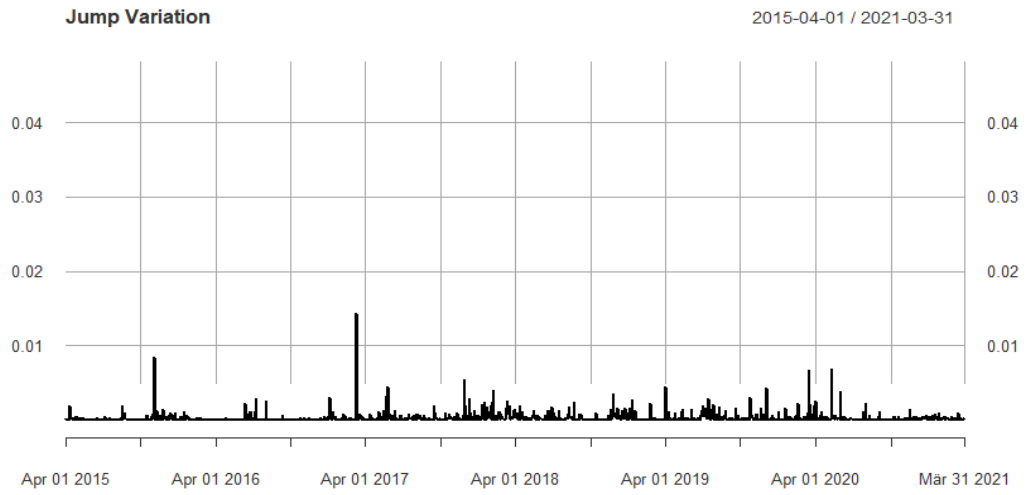

Fig. 4. Jump Variation of BTC

Our investigation insinuates that the Gaussian distribution is a rather crude approximation for the realized variance. There is too much weight in the tails. In the first three years, the evidence against the log-normality property is even stronger, whereas the log-normality property holds better for 2018–2021. The density looks more symmetric and the tails are thinner. In summary, the results suggest that there are more extreme values of the realized variance in the first three years than in the last three years. The realized variance may be time-varying. This is also in line with Figure 4, as the realized variance in the second sub-period contains less outliers except at the beginning of the COVID-19 crisis.

4. Cointegration

The portfolio constructions that we use are linear combinations of individual cryptocurrencies. The I(1) property of many financial time series is well established in the literature (see Alexander (1999)). A variable is said to be I(1) (or *first-order integrated*) if it is non-stationary but its first difference is stationary. Whereas some researchers claim that the variance of speculative prices is infinite (see Mandelbrot (1963)), and this may be particularly relevant for the highly volatile cryptocurrencies, we keep a finite variance as a technical assumption, well aware that the statistical properties of many time series procedures are no longer guaranteed if the condition of finite variance is violated.

Stationary combinations of I(1) variables are of special interest, and this is the property of *cointegration* that has been introduced by Granger (1981). To empirically establish the cointegration property and to estimate the stationary linear combinations, various procedures are available. The simplest and oldest one is the EG-2 procedure by Engle and Granger (1987). This method is now known to be inefficient and it can be regarded as outdated. Another shortcoming is the fact that EG–2 is difficult to generalize to more than two I(1) variables. The other method that we consider is the efficient maximum-likelihood procedure by Johansen (1988).

4.1. Testing for cointegration

The EG-2 procedure proceeds as follows. Consider two first-order integrated variables X and Y. Estimate the simple cointegrating regression $Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t$ by OLS to get estimates of β_0 and of β_1. Run a unit-root test on the residuals $\hat{\varepsilon}_t = Y_t - \hat{\beta}_0 - \hat{\beta}_1 X_t = Y_t - \hat{\boldsymbol{\beta}}' \mathbf{X}_t$[1]. For n regressors we can write $\hat{\varepsilon}_t = Y_t - \hat{\beta}' \mathbf{X}_t$[2]. If it rejects, the errors ε_t can be seen as I(0). $\hat{\varepsilon}_t$ can also be interpreted as the error-correction term (ECT), when estimating an error-correction model.

We use the most commonly used unit-root test, i.e. the (augmented) Dickey-Fuller test by Dickey and Fuller (1979). In this test, differenced variables are regressed on p lagged differences, on deterministic variables, and on a lagged level term. The t–value of the coefficient of the lagged term defines the test statistic. The constant p is often found via information criteria such as AIC. The null hypothesis of the Dickey-Fuller test is the existence of a unit root. With regard to deterministic parts, there exist three variants, the DF_0, the DF_μ, and the DF_τ test. For a summary overview of differences across the three types, we refer to Dickey and Fuller (1979). For DF_0 and DF_μ, the null is an I(1) process, i.e. a generalized random walk.

In a preliminary step, both X and Y are tested by DF_τ for unit roots, and ΔX and ΔY are tested by DF_μ. If the level tests do not reject and the tests in differences reject,

[1] Here, β and $\mathbf{X}_t$ are of dimension (2×1).

[2] Here, β is an $((n+1) \times 1)$ vector (including the intercept), and $\mathbf{X}_t$ is another $((n+1) \times 1)$ vector of n I(1) regressors extended by one.

both X and Y can be viewed as I(1). In this case, Y is regressed on X, and the residuals are subjected to a DF_0 test, which can be seen as the second step of EG-2. In this step, the original DF significance points are invalid and correct significance points tabulated in Phillips and Ouliaris (1990) should be used.

The Johansen procedure for testing and estimation of cointegrated systems considers a multivariate VAR(p) model. This VAR(p) model can be written as a VAR($p-1$) in differences with an additional ECT $\boldsymbol{\beta' X_{t-1}}$. In this form, it is also called a vector error-correction model (VECM).

$$\Delta \mathbf{X_t} = \boldsymbol{\delta} + \boldsymbol{\alpha\beta'} \mathbf{X_{t-1}} + \boldsymbol{\Gamma_1} \Delta \mathbf{X_{t-1}} + \ldots + \boldsymbol{\Gamma_p} \Delta \mathbf{X_{t-p+1}} + \boldsymbol{\varepsilon_t},^{3}$$

where $\boldsymbol{\beta' X_{t-1}}$ is stationary and $\boldsymbol{\alpha\beta'}$ is called the *impact matrix* $\boldsymbol{\Pi}$ which is an $n \times n$–matrix of rank r. According to this rank r, there are three cases:

1. If $r = 0$ then $\boldsymbol{\Pi}$ is a zero matrix and there is no cointegration in the system which implies that the VAR(p) model is really a VAR($p-1$) model for differences $\Delta \mathbf{X}$
2. If $\boldsymbol{\Pi}$ has full rank ($r = n$) and is non-singular then $\mathbf{X}$ is already stable
3. If the rank r of $\boldsymbol{\Pi}$ fulfills $0 < r < n$, then there are r linearly independent cointegrating vectors $\boldsymbol{\beta_j}, j = 1, \ldots, r$, such that $\boldsymbol{\beta_j'} \mathbf{X}$ is stationary.

$\boldsymbol{\beta}$ contains dynamic equilibrium conditions, and the loading matrix α describes how the components of $\Delta \mathbf{X}$ react to deviations from these conditions. To identify the rank r of the impact matrix $\boldsymbol{\Pi}$ and hence the number of cointegrating vectors, the trace test of the Johansen procedure is used. Here, we choose the 5% significance level to estimate the error-correction model given $r = r_0$. This is a so-called reduced rank regression and requires solving the canonical correlation eigenvalue problem. For details see Johansen (1988), Johansen (1991), Johansen (1995).

[3] Here, $\mathbf{X_t}$ is again a $(n \times j)$ matrix, α and β are of dimension $(n \times r)$, where r is the number of cointegrating relationships.

4.2. Trading the spread

In this section, we briefly describe the implementation of trading strategies exploiting cointegrating relationships among cryptocurrencies. The presented strategies are very similar to a market neutral pairs trading strategy where the spread of two cointegrated stocks is traded and a trader short sells the overvalued stock and takes a long position in the undervalued stock. Pairs trading and also our trading strategies only work if the spread is mean-reverting such that it is necessary to have a long run equilibrium. Gatev et al. (2006) adopted such an investment strategy for different stock pairs. Their results suggest that pairs trading yields average annualized excess returns of approximately 11%. Likewise, Tokat and Hayrullahoğlu (2022) apply pairs trading to a portfolio of 45 pairs. They find an average annual return of 15% with an average Sharpe ratio of 1.43 after considering transaction costs. The trading design of our investment strategy here works as follows. In line with Leung and Nguyen (2018) we create a cointegrated portfolio of different cryptocurrencies but in contrast to them we use ten cryptocurrencies instead of just four. The first step is to find potential cointegrating vectors that guarantee a mean-reverting spread. For this purpose, the Johansen procedure and the Engle-Granger two-step procedure are used. We first apply them to each BTC pair, then we construct a portfolio of all ten cryptocurrencies and apply them again. Although EG-2 is known to be inefficient, we still use it here as a benchmark due to its simplicity. For EG–2 with more than five explanatory variables, MacKinnon (2010) provides critical values.

On the other hand, we want to compare the performance of an investment strategy based on the EG-2 and on the Johansen procedure. Among other authors, Alexander (1999) emphasized the importance of a thorough out-of-sample performance evaluation when applying statistical arbitrage investment strategies. Therefore, we split the sample into a training period and a test period. After finding potential cointegrating vectors, two (or more) different spreads are obtained depending on the rank of the impact matrix. $spread_t^{EG}$ denotes the spread of the EG-2 and $\text{spread}_t^{J,1}$ naming the spread as a result of the Johansen procedure for the first cointegrating vector, $\text{spread}_t^{J,2}$ is based on the second vector and so on. For the optimal threshold that determines the level for buying or selling

the spread, there exist many suggestions in the literature. The optimal threshold is defined as the one that maximizes the overall profit. For a discussion we refer to Song and Zhang (2013), Ngo and Pham (2016), and Liu et al. (2020), for instance. In general, the trading strategy for in-sample evaluation works as follows:

- Long the spread if $\text{spread}_t^i \leq \mu - \tau$ for $i \in \{EG, J1, J2, \ldots\}$
- Unwind a long position at the first date when $spread_t^i \geq \mu$ for $i \in \{EG, J1, J2, \ldots\}$
- Short the spread if $\text{spread}_t^i \geq \mu + \tau$ for $i \in \{EG, J1, J2, \ldots\}$
- Unwind a short position at the first date when $spread_t^i \leq \mu$ for $i \in \{EG, J1, J2, \ldots\}$
- If there is any open position until the end, the position is closed at the last trading day.

Note that μ is the mean of the spread and τ the threshold. 'Long the spread' has the same meaning as buying one unit of the spread, which requires purchasing every cryptocurrency with a positive sign and short selling the cryptocurrencies with a negative sign as indicated by the corresponding cointegrating vector. To unwind a long position, it is necessary to sell the cryptocurrencies with a positive sign and buy back all cryptocurrencies with a negative sign. 'Short the spread' just means short selling one unit of the spread where one needs to short sell the cryptocurrencies with a positive sign and buy the cryptocurrencies with a negative sign. Unwinding a short position requires buying back the short cryptocurrency with a positive sign and selling the cryptocurrencies with a negative sign. It should be noted that short selling only happens if one long or short the spread, as defined before, but not when the position is closed since the cryptocurrencies are already in the inventory. Furthermore, in this paper we use two different thresholds for the in-sample trading strategy, i.e. $\tau \in \{\sigma, \tau^*\}$, where σ denotes the standard deviation of the spread and τ^* the optimal threshold based on a parametric approach. Generally, the total profit is simply the number of trades times the profit of each trade. The number of trades can be estimated if the distribution of $spread_t^i$ for $i \in \{EG, J1, J2, \ldots\}$ is known. For the case of $spread_t^{EG}$ this may often be approximately normal. Nevertheless, the distribution of

$\text{spread}_t^{J,1}, \text{spread}_t^{J,2}, \ldots$ is unknown. For estimating the best fitting distribution and the corresponding optimal parameters, we use the *Fitter* package in Python, which minimizes the sum of squared errors.

After this analysis, it is possible to estimate the number of trades. Let $\psi^i(.)$ denote the estimated CDF of $spread_t^i$ for $i \in \{EG, J1, J2, \ldots\}$. Then the number of trades is roughly given by

$$N * [\mathrm{P}(spread_t^i < \mu - \tau) + \mathrm{P}(\text{spread}_t^i > \text{spread}\mu + \tau)] = 851 * [\psi^i(\tau) + (1 - \psi^i(\tau))], \quad (1)$$

for $i \in \{EG, J1, J2, \ldots\}$, where N is the number of trading days. Furthermore, the profit of each trade is approximately τ. As a result, the optimal threshold solves a simple maximization problem, formally

$$\tau^* = \arg max_\tau \{851 * \tau * [\psi^i(\tau) + (1 - \psi^i(\tau))]\} \quad \text{for} \quad i \in \{EG, J1, J2, \ldots\}. \quad (2)$$

One shortcoming of this approach is that equation (1) does not exactly describe the number of profit realizations since a profit is only realized if the spread crosses the mean after it reaches the upper or lower threshold. Nevertheless, if the spread is mean-reverting due to cointegrated time series it will always converge to equilibrium after it deviates from the mean by $|\tau|$. Therefore, equation (2) holds approximately for calculating the optimal threshold τ^*. For simplicity we assume that it is only possible to long or short one unit of the spread. Finally, we compare the trading strategies with both thresholds when investing \$ 1,000 at the beginning with a passive trading strategy where \$ 1,000 are equally invested in all ten cryptocurrencies (i.e., \$ 100 per cryptocurrency) at the start then sold at the end of the period. As already mentioned, it is crucial that the trading strategies are subject to an out-of-sample backtesting analysis. For backtesting, we use a time interval of three months where we introduce two different trading strategies. For both strategies we use the same cointegrating vectors estimated from the training data. If there really exists a long-term relationship we expect that the ECT is also stationary when using the same coefficient vector but the log-prices of the test period. Since the future and hence the price data of

Table 3
Dickey-Fuller test results

Variable	levels DF_τ test statistic	differences DF_μ test statistic
BTC	-1.416	-13.5199
ETH	-1.2534	-8.6162
BNB	-1.5050	-7.4993
ADA	-0.4915	-13.1993
XRP	-2.2439	-29.9345
DOGE	-1.1167	-15.5711
LTC	-1.5778	-10.8248
BCH	-1.4857	-7.8711
XLM	-1.3145	-30.4498
XMR	-2.1101	-13.248
critical value 1%	-3.970	-3.438
critical value 5%	-3.416	-2.865

the test period are assumed to be unknown it is not possible to calculate the mean and the standard deviation of the test data. To overcome this issue, we employ a rolling window approach with two different window sizes: A long window with the same length as the test period, which reacts slowly to a sudden significant fall or increase of the spread, and a short window with a length of ten days, which is able to adjust fast to abrupt behavior of the spread. To accomplish this, we extend the out-of-sample data at the start with the last three months (ten days) of the training data for calculating the 90-day (10-day) moving average and the 90-day (10-day) rolling standard deviation. This approach allows for calculating the moving average and the rolling upper and lower threshold for each single day in the test period as time goes by. A difference to the in-sample evaluation is that the mean and both thresholds are time-varying. This implies in symbols $\tau \in \{\sigma_t(10), \sigma_t(90)\}$, because with this backtesting approach it is not possible to estimate the distribution of the test sample like it is done in-sample. Again, all out-of-sample trading strategies are compared to a passive investment strategy in the same way as described for the in-sample evaluation.

Table 4
EG-2 unit root test results (pairs and portfolio)

pairs	test statistic
BTC-ETH	-1.7321
BTC-BNB	-1.9864
BTC-ADA	-2.0167
BTC-XRP	-2.8475
BTC-DOGE	-1.9706
BTC-LTC	-2.1886
BTC-BCH	-1.5833
BTC-XLM	-2.9836
BTC-XMR	-2.834
critical value 1%	-3.96
critical value 5%	-3.37
portfolio	
spread_t^{EG}	-5.7634
critical value 1%	-6.00
critical value 5%	-5.47

Note: Critical values follow MacKinnon (2010).

4.3. Empirical results

Table 3 indicates that all variables follow non-stationary processes with a unit root. We use the DF_τ specification for all variables as we assume that all cryptocurrencies are trending. It is crucial that all variables are I(1) to be able to apply cointegration tests. Dickey-Fuller test results for first differences—here the DF_μ variant is used as growth rates do not trend—are shown in the right part of Table 3: the null hypothesis is now generally rejected at the 1% significance level. This means that all variables are integrated of order one, I(1).

The results of EG-2 in Table 4 are surprising, as no pair is cointegrated at a reasonable significance level. A portfolio containing more than two and maybe even all cryptocurrencies, however, may still have a long-run relationship. To consider all ten cryptocurrencies, we run the following regression:

$$\begin{aligned} BTC_t &= \beta_0 + \beta_1 ETH_t + \beta_2 BNB_t + \beta_3 ADA_t + \beta_4 XRP_t + \beta_5 DOGE_t \\ &\quad + \beta_6 LTC_t + \beta_7 BCH_t + \beta_8 XLM_t + \beta_9 XMR_t + \varepsilon_t \end{aligned} \tag{3}$$

Estimating this regression enables calculating the spread that can be used for statistical arbitrage. The spread is given by

$$\begin{aligned} \text{spread}_t^{EG} &= BTC_t - 6.8297 - 0.2034ETH_t - 0.1130BNB_t - 0.0684ADA_t \\ &\quad +0.1231XRP_t + 0.0667DOGE_t - 0.8213LTC_t + 0.5001BCH_t \\ &\quad -0.1459XLM_t - 0.1016XMR_t \end{aligned} \tag{4}$$

If there is cointegration in the system, the spread as defined in equation (4) should be stationary. Applying the Dickey-Fuller test yields the bottom row of Table 4. Whereas EG-2 does not find any cointegrating relationship for any Bitcoin pair, the test for multiple regression residuals suggests that the portfolio containing all ten cryptocurrencies is cointegrated at the 5% significance level. Figure 5 shows the time series of $spread_t^{EG}$. For our purposes, we omit the intercept of the cointegrating vector in the construction of the spread, as it cannot affect its stationarity.

Equation (4) informs that to long one unit of the spread one needs to buy 1, 0.1231, 0.0667 and 0.5001 units of BTC, XRP, DOGE and BCH, respectively. In addition, it is necessary to short sell 0.2034, 0.1130, 0.0684, 0.8213, 0.1459 and 0.1016 shares of ETH, BNB, ADA, LTC, XLM and XMR, respectively. Shorting one unit of the spread works the other way around. In contrast to classical stocks, it is feasible to buy fractional shares up to eight decimal places of a crypto asset.

We now consider the results of the Johansen procedure. In contrast to EG-2 there is cointegration for some cases. With a 5% (10%) significance level, the null of no-cointegration is rejected three (five) times. As convened above, we generally go for a 5% significance level, whereupon we conclude that the three cryptocurrencies pairs BTC-BNB, BTC-XRP and BTC-DOGE are cointegrated. The BTC-DOGE relationship is rather surprising, as the log-returns of Bitcoin have the lowest correlation with the log-returns of DOGE. For the optimal VAR lag order, we adopted the optimum of 5 found by AIC and FPE. The stricter criteria BIC and HQIC would have selected an optimum of only 1.

The trace test rejects $H_0 : r = 0$ at the 1% significance level and $H_0 : r \leq 1$ at the

Table 5
Johansen test results (pairs)

Variables pair	H_0	Test statistic
BTC-ETH	$r \leq 1$	1.7348
	$r = 0$	14.2598*
BTC-BNB	$r \leq 1$	5.1611**
	$r = 0$	20.9306***
BTC-ADA	$r \leq 1$	2.1820
	$r = 0$	14.0817*
BTC-XRP	$r \leq 1$	1.8931
	$r = 0$	20.009***
BTC-DOGE	$r \leq 1$	3.2619*
	$r = 0$	16.1165**
BTC-LTC	$r \leq 1$	0.969
	$r = 0$	10.1242
BTC-BCH	$r \leq 1$	0.6756
	$r = 0$	12.8733
BTC-XLM	$r \leq 1$	1.7671
	$r = 0$	11.087
BTC-XMR	$r \leq 1$	1.8913
	$r = 0$	11.255

Notes: The VAR lag order is chosen by minimizing the AIC for lags up to 10. Critical values are 2.7055, 3.8415, 6.6349 for $H_0 : r \leq 1$ and 13.4294, 15.4943, 19.9349 for $H_0 : r = 0$, in both cases at 10%, 5%, 1% significance.

Table 6
Johansen test results (portfolio)

		critical value		
H_0	Test statistic	10%	5%	1%
$r \leq 3$	113.224	120.367	125.618	135.982
$r \leq 2$	157.392	153.634	159.529	171.09
$r \leq 1$	207.712	190.871	197.377	210.037
$r = 0$	274.497	232.103	239.247	253.253

Notes: Critical values are based on MacKinnon et al. (1999)

Table 7
Cointegrating vectors (Johansen procedure)

	$\beta_{J,1}$	$\beta_{J,2}$
BTC	1.0000	0.0000
ETH	0.0000	1.0000
BNB	-0.5311***	1.8189***
ADA	-0.5807*	-0.9590
XRP	0.0266	0.3268
DOGE	0.4241	-0.8650*
LTC	-1.0504***	-5.4176***
BCH	0.2079	3.8990***
XLM	0.9163***	3.2049***
XMR	-0.2590	-2.3504***

Notes: *, **, and *** denote significance at the 10%, 5%, and 1% level, respectively. The values for BTC and ETH are restricted and are not tested.

5% significance level, but $H_0 : r \leq 2$ is only rejected at the 10% significance level, which we consider as insufficient evidence. In summary, we conclude that the rank of the impact matrix is 2, so we estimate a VECM with a cointegrating rank of $r_0 = 2$. Table 7 shows both significant cointegrating vectors. In $\beta_{J,1}$ there are only four significant coefficients, namely BNB, ADA, LTC and XLM, while ADA is only significant at the 10% level. $\beta_{J,2}$ has a more desirable outcome with only two insignificant parameters (ADA and XRP). LTC and BCH represent the largest entries (in absolute terms) of the second vector and are highly significant.

Figure 5 shows the time series of all error-correction terms, which should be stationary and represent the three spreads.

It is immediately apparent that all spreads fluctuate around zero and that $spread_t^{J,2}$ has the highest standard deviation. A stationary time series with high variance is pleasant, as a high standard deviation yields higher average returns. Figure 5 confirms that the residuals of equation (4) look indeed stationary even the standard deviation compared to $spread_t^{J,2}$ is relatively low. The time series of $spread_t^{J,1}$ in Figure 5 also looks stationary with a slightly higher variation but there was a huge deviation from the long-run equilibrium at the beginning of the year 2021. This may have been caused by the fast increase of the BTC price, which peaked at around $ 50,000 in March 2021, as $\beta_{J,1}$ puts the greatest weight

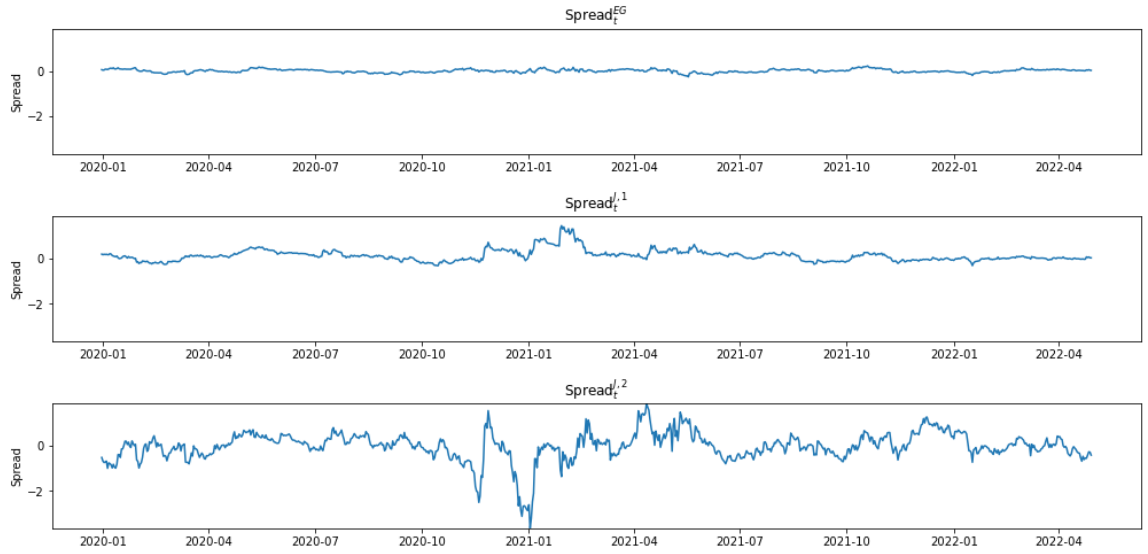

Fig. 5. Time series of all Spreads

on BTC with one unit. By contrast, the start of the COVID-19 crisis did not result in a significant disequilibrium.

ECT 2 shows a similar picture since the time series of $spread_t^{J,2}$ fluctuates around zero. Again, at the start of the year 2021 the cryptocurrencies were in disequilibrium. However, the spread converges back to its long-run equilibrium relatively fast, which indicates a long-term relationship between the variables. $\beta_{J,2}$ puts zero weight on BTC by construction. The exclusion of any BTC-ETH interaction may be also responsible for the better fit of the ECT based on the EG-2 for two reasons. First, the return correlation matrix (not shown for brevity) insinuates that Bitcoin forms a group of closely related assets with Ethereum, Litecoin and Bitcoin Cash. This is why the cointegrating vectors may depend strongly on the BTC-ETH, BTC-LTC and BTC-BCH interaction. While $\beta_{J,1}$ and $\beta_{J,2}$ exclude the BTC-ETH relationship by definition, β_{EG} allows any interaction. Second, the Johansen trace test rejects the null of no-cointegration between BTC and ETH at least at the 10% significance level indicating the importance of the BTC-ETH interaction.

Table 8 provides the estimates of the loading matrix α. The adjustment parameters in ECT_1 are all positive while in ECT_2 the parameters are all negative (except for XMR). The coefficients are rather small, however, which implies a slow speed of adjustment. Since ten variables are involved in the VECM, the interpretation of the parameters is complex.

Table 8
Loading matrix

	ECT_1	ECT_2
BTC	0.0133**	-0.0077***
ETH	0.0202***	-0.0088***
BNB	0.0401***	-0.0100***
ADA	0.0344	-0.0094**
XRP	0.0049	-0.0058
DOGE	0.0263**	-0.0044
LTC	0.0147*	-0.0051
BCH	0.0116	-0.0099***
XLM	0.0055	-0.0177***
XMR	0.0203***	0.0013

Notes: *, **, and *** denote significance at the 10%, 5%, and 1% level, respectively.

Table 9
Optimal thresholds

spread	Estimated CDF	τ^*
$spread_t^{EG}$	$\psi^{EG}(.) = N(\mu = 6.8297, \sigma = 0.0733)$	0.055
$spread_t^{J,1}$	$\psi^{J,1}(.) = LN(\mu = 0.5561, \sigma = 0.3869)$	0.209
$spread_t^{J,2}$	$\psi^{J,2}(.) = t_{3.0768}(\mu = 0.0351, \sigma = 0.3889)$	0.37

When considering the significance in ECT_1 we found that only BTC, ETH, BNB, DOGE, LTC and XMR are significant at the 10% significance level, which means that only these variables adjust to deviations from the long-run equilibrium. In ECT_2, the loading coefficients significant at 5% include BTC, ETH, BNB, ADA, BCH and XLM, which means that these variables react to the error-correction term. The variables with a true $\alpha = 0$ are weakly exogenous for the cointegrating vector as defined by Engle et al. (1983), which implies that Ripple (XRP) is the only exogenous variable.

Table 9 shows distribution estimates for $spread_t^i$ for $i \in \{EG, J1, J2\}$. Three different specifications achieve the best fit: a normal distribution for $spread_t^{EG}$; a log-normal distribution for $spread_t^{J,1}$; and a t–distribution for $spread_t^{J,2}$. Now it is possible to solve equation (2) for each CDF to obtain the theoretical optimal threshold.

For the further analysis, we use the following notation for the sake of simplicity:

- Trading strategy 1: Use $spread_t^{EG}$ with $\tau = \sigma$
- Trading strategy 2: Use $spread_t^{EG}$ with $\tau = \tau^*$

Table 10
Summary of all trading strategies (In-sample: 2019/12/31-2022/04/29)

	Trading strategy					
Strategy	1	2	3	4	5	6
Total return realizations	25	26	9	10	17	28
Total transactions	50	52	18	20	34	56
Largest log-return	15.02%	15.00%	60.22%	55.18 %	162.41%	139.54%
Lowest log-return	7.72%	4.18%	16.33%	11.50%	26.56%	-1.56%
Average log-return	11.11%	9.18%	32.81%	29.87%	87.90%	64.84%
Total log-return	277.83%	238.66%	295.31%	298.74%	1,494.37%	1,815.42%

- Trading strategy 3: Use $spread_t^{J,1}$ with $\tau = \sigma$
- Trading strategy 4: Use $spread_t^{J,1}$ with $\tau = \tau^*$
- Trading strategy 5: Use $spread_t^{J,2}$ with $\tau = \sigma$
- Trading strategy 6: Use $spread_t^{J,2}$ with $\tau = \tau^*$
- Trading strategy 7: Passive investing approach

We note that both spreads based on the Johansen procedure have a much higher standard deviation than the spread of the EG-2 (see Table 9). Since the profit for two subsequent trades equals approximately τ it can be conjectured that $spread_t^{J,1}$ and $spread_t^{J,2}$ have a higher average return. Moreover, $spread_t^{EG}$ and $spread_t^{J,2}$ seem to have a higher speed of mean reversion than $spread_t^{J,1}$ resulting in more mean-crossings and thus more trades. When looking at the optimal thresholds in Table 9 and comparing them to the standard deviations of the spreads it is discernable that τ^* of all spreads is smaller than σ leading to more profit realizations. Table 10 summarizes the most important key data.

As expected, trading strategies 3 and 4 result in fewer return realizations due to the slow mean reversion. A return realization occurs after unwinding a long/short position that is why two transactions in total (long/short the spread and unwind long/short position) are needed. Nevertheless, the overall returns of both strategies are higher than the total return of trading strategy 1 and 2 according to $spread_t^{EG}$, even though strategies 1 and 2 lead to almost three times as many transactions than strategies 3 and 4. Another outcome that stands out is the clear superiority of trading strategies 5 and 6. We conjecture that the main

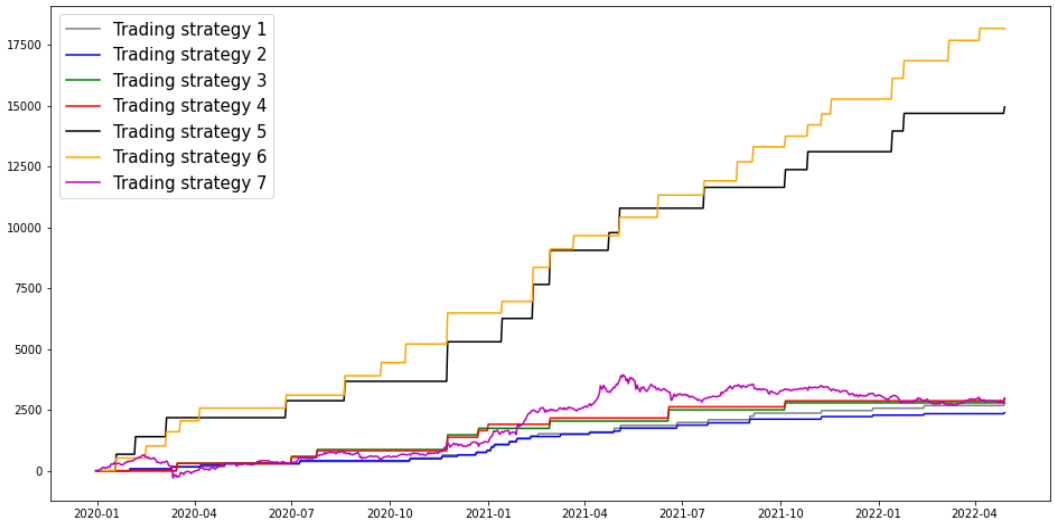

Fig. 6. Cumulative returns – All trading strategies (In-sample)

driver of this dominance is the fast mean-reverting $spread_t^{J,2}$ and its relatively high standard deviation. Even though the average return of strategy 5 is around 20 percent points larger than of strategy 6, the significantly more transactions of strategy 6 predominate yielding a higher overall return. On the one hand, both trading strategies resting upon the Johansen approach with the theoretical optimal threshold outperform the strategies where σ is used as threshold. On the other hand, the mean-reverting portfolio constructed with $spread_t^{EG}$ and $\tau = \tau^*$ has a weaker performance than strategy 1 where just the standard deviation is used. Because of this result the effect of the usage of τ^* is ambiguous but because of the superiority in two out of three cases we assume that estimating the distribution of the spread still makes sense for calculating the optimal threshold. Finally, the next plot illustrates the time series of the cumulative returns of all six trading strategies and the passive investment approach.

Figure 6 shows that trading strategies 5 and 6 are clearly superior to all other strategies. Moreover, we found that if all cryptocurrencies bought at the beginning were sold at almost any time point during 2021, the passive investment strategy would be superior to all trading strategies except both strategies generated by $spread_t^{J,2}$. This is a strong result as it suggests that the risk associated with trading has no remarkable advantage over a long-term passive investing strategy. However, due to the crypto bear market starting at the beginning of

Table 11
Summary of all trading strategies (In-sample)

Strategy	Final wealth
Trading strategy 1	\$3,778.31
Trading strategy 2	\$3,386.56
Trading strategy 3	\$3,953.14
Trading strategy 4	\$3,987.43
Trading strategy 5	\$15,943.67
Trading strategy 6	**\$19,154.22**
Trading strategy 7	\$3,789.56

2022 strategy 7 becomes less attractive ex post and approaches a similar level as trading strategies 3 and 4. Figure 6 indicates that trading strategies 3 and 4 outperform the passive investment strategy, while both strategies constructed by $spread_t^{EG}$ generate less profit than strategy 7. Table 11 summarizes the final wealth of all strategies including the \$1,000 investment at the beginning. Trading strategy 6 with a final wealth of \$19,154.22 is second to none and yields a five-time higher final wealth than the passive investing strategy. Whether the superiority of the strategies generated by $spread_t^{J,2}$ is also present out-of-sample, appear in the next section.

As before, for simplicity the further analysis considers the following notation:

- Trading strategy 1: Use $spread_t^{EG}$ with $\tau = \sigma_t(90)$
- Trading strategy 2: Use $spread_t^{EG}$ with $\tau = \sigma_t(10)$
- Trading strategy 3: Use $spread_t^{J,1}$ with $\tau = \sigma_t(90)$
- Trading strategy 4: Use $spread_t^{J,1}$ with $\tau = \sigma_t(10)$
- Trading strategy 5: Use $spread_t^{J,2}$ with $\tau = \sigma_t(90)$
- Trading strategy 6: Use $spread_t^{J,2}$ with $\tau = \sigma_t(10)$
- Trading strategy 7: Passive investing approach

First of all, it is immediately apparent that the process of all three spreads has the same pattern. Until the mid of June 2022, the spread is more or less mean-reverting followed

Table 12
Summary of all trading strategies (Out-of-sample: 2022/04/30-2022/07/31)

	Trading strategy					
Strategy	1	2	3	4	5	6
Total return realizations	4	12	5	10	3	11
Total transactions	8	17	10	17	6	16
Largest log-return	7.74%	6.44%	8.76%	8.76%	40.43%	39.42%
Lowest log-return	-21.01%	-31.45%	-20.95%	-18.09%	-39.43	-119.34%
Average log-return	-0.34%	0.24%	0.97%	0.50%	7.84%	5.13%
Total log-return	-1.37%	2.84%	4.86%	5.04%	23.51%	56.44%
Final wealth in $	986.29	1028.46	1048.62	1050.44	1235.12	1564.36

Final wealth for Trading strategy 7: $ 552.77

by a huge downward movement. This suggests that during this time interval something happened causing a disequilibrium of the cryptocurrencies. As it is the case in-sample the mean-reversion of $spread_t^{J,1}$ is very slow leading to a continuing downward movement, while in July 2022 $spread_t^{EG}$ fluctuates at a relatively constant level and $spread_t^{J,2}$ even slowly converges back to its long-run equilibrium at the end of the period indicating the best mean-reversion. Another explanation for the better mean-reverting property of $spread_t^{J,2}$ are the findings of Table 8 since the more significant and negative loading parameters lead to an adjustment to return to the long-run equilibrium after the disequilibrium. Furthermore, all trading strategies with a 90-days rolling window generate a loss at the end of the test period by unwinding the last long position because of the significant deviation from the long-run equilibrium. Using a shorter window size results in significantly more transactions than the longer 90-days window.

Table 12 summarizes the key facts of all out-of-sample trading strategies. All trading strategies except strategy 1 produce a positive return at the end of the period. Strategies with a shorter window size seem to outperform trading strategies with a 90-day window length. Just as in the in-sample evaluation, trading strategy 2 based on $spread_t^{EG}$ yields the lowest positive total log-returns. Likewise, trading strategy 6 based on $spread_t^{J,2}$ yields the highest total return with 56.44% despite a loss of more than 100%. This outcome could provide evidence that a good in-sample trading performance with a cointegrated cryptocurrency portfolio is also a good indicator for a promising out-of-sample performance. This can be pinned down by the fact that regardless of the window size both trading

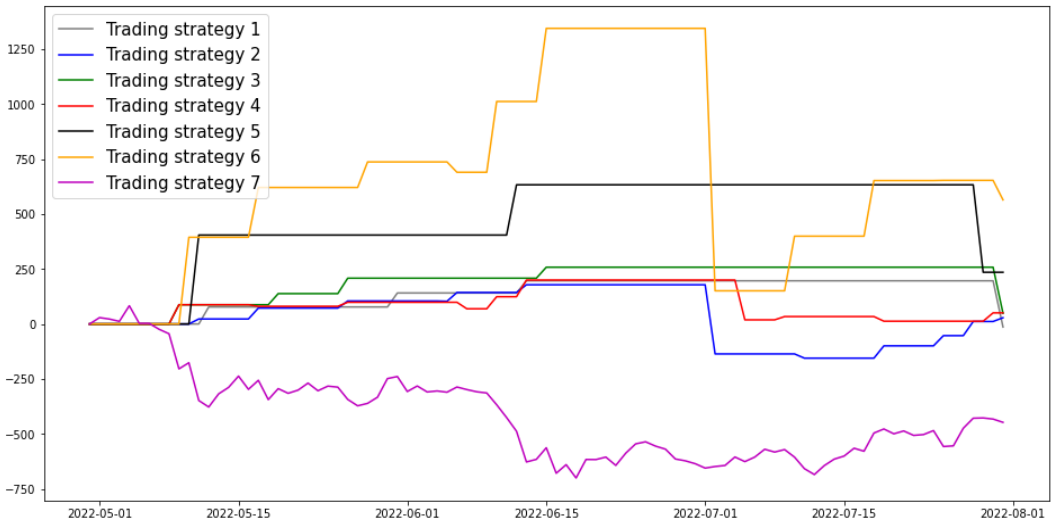

Fig. 7. Cumulative returns (Out-of-sample)

strategies based on $spread_t^{J,2}$ outperform all other trading strategies. In any case, the main takeaway of the out-of-sample backtesting analysis is that the strategies with the 10-days window length are superior to the strategies with a window size of three months by resulting in a considerable higher total return. When comparing the first two trading strategies one can see that using a short window generates a positive profit while using the 90-days window leads to a loss. Figure 7 compares the cumulative returns of all six trading strategies to the passive investment approach.

All trading strategies yield a higher total return than the passive investment approach because of the crypto bear market during 2022. These results suggest that cointegrated cryptocurrency portfolios can be useful for hedging against a downtrend market. As already implied by Table 12, strategy 6 is by far the best performing trading strategy. In addition, trading strategy 5 tends to outperform the first four trading strategies and even strategy 6 at the beginning of July but the costly unwinding of the last long position leads to a huge loss and hence a weaker performance than trading strategy 6. However, the bottom line is that the superiority of trading strategy 6 based on $spread_t^{J,2}$ is also prevailing out-of-sample. Figure 7 (see also Table 12) shows that trading strategy 1 and the passive investment strategy result in a loss while the other strategies yield positive profits. To relate this outcome to the high inflation rates in Europe and the US nowadays, these findings indicate

that statistical arbitrage strategies based on cointegrated cryptocurrency portfolios are indeed a profitable shield against high inflation rates. Considering an annual inflation of 10% for simplicity, it is necessary to generate a final wealth of at least $\$1,000 * (1.1)^{\frac{3}{12}} = \$1,024.11$ after the three months test period, which can be achieved with five out of six statistical arbitrage strategies.

4.4. Volatility and the spread

In this section we investigate the impact of market volatility on the spread that is used for statistical arbitrage. The time series of the three spreads show that there are huge spikes occasionally, which may be caused by high volatility in the market. This becomes especially apparent when looking at $spread_t^{J,1}$ and $spread_t^{J,2}$. As a proxy for the overall volatility in the cryptocurrency market we use the Crypto Volatility Index (CVI), which is the counterpart to the well-known CBOE Volatility Index (VIX). Figure 8 shows the CVI during the in-sample period. It indicates that there are three excessive spikes caused by, among other things, the start of the COVID-19 pandemic. Both episodes of a strong disequilibrium of $spread_t^{J,1}$ and $spread_t^{J,2}$ are periods of high market volatility implied by the CVI. This suggests that there may be a connection between the spread and market volatility.

Fig. 8. Crypto Volatility Index

In order to find empirical evidence for a potential relatedness between volatility and the spread we apply a simple Granger causality test (see Granger (1969)). Usually, Granger causality is tested via a restriction test in a VAR. For the following analysis we estimate three bivariate VAR models to test for Granger causality between the CVI and each individual spread. To ensure a stable VAR it is necessary that each variable is I(0). $spread_t^{EG}$, $spread_t^{J,1}$ and $spread_t^{J,2}$ are covariance stationary by definition, so the only critical variable is CVI. For CVI, the DF_μ test rejects the null of a unit root at the 5% significance level indicating that CVI is covariance stationary, resulting in stable VAR models.

Table 13
Granger causality test results

H_0	p-value	Conclusion	VAR lag order
CVI does not Granger cause $spread_t^{EG}$	0.4265	Not reject	1
CVI does not Granger cause $spread_t^{J,1}$	0.0256	Reject	5
CVI does not Granger cause $spread_t^{J,2}$	0.2531	Not reject	8

Notes: The VAR lag order is chosen by minimizing the AIC for lags up to 10

The results of Table 13 suggest that there is only one case where the CVI Granger causes the spread. Whereas this appears to be a weak result for a strong connection between volatility and the spread used for establishing the trading strategies, the CVI is only a noisy proxy for the overall cryptocurrency market volatility and not a reliable indicator for the volatility of the three cointegrated portfolios based on just ten specific cryptocurrencies. Nevertheless, we found evidence that the CVI granger cause $spread_t^{J,1}$ implying that the CVI is useful in forecasting $spread_t^{J,1}$. Thus, there is evidence that the volatility of each individual cryptocurrency in the portfolio helps to generate a more accurate forecast for the spread. Hence, the incorporation of volatility for statistical arbitrage strategies leads to a better out-of-sample performance especially due to the high volatile crypto market. On the basis of this conjecture, the next section deals with a rigorous volatility analysis of the most popular cryptocurrency BTC. In addition, we make out-of-sample one-step ahead volatility forecasts and consider potential price jumps. The same analysis can be applied to the other nine cryptocurrencies.

5. Volatility modeling

Understanding the sources and the dynamics of volatility in financial markets is crucial in risk management, portfolio allocation, derivative pricing and other related fields. In this section, we introduce two kinds of volatility models, which are used for estimating and forecasting the volatility of BTC in the period from 1st Apr 2015 until 31st Mar 2021. On the one hand, we use different types of GARCH models including the standard GARCH model, the EGARCH, the GJR-GARCH and the more recent realized GARCH model. On the other hand, six different HAR-RV models are also estimated.

5.1. Theoretical framework

This subsection discusses the methodology of the four GARCH models and of the HAR-type models.

Generalized autoregressive conditional heteroskedasticity (GARCH) models of order p and q as proposed by Bollerslev (1986) are used for modeling the conditional variance. They allow for a high persistence in the process. In the following only the simple case $p = q = 1$ is considered, which reduces the process to

$$r_t = \mu_t + \sigma_t z_t,$$

$$\sigma_t^2 = \omega + \alpha_1 r_{t-1}^2 + \beta_1 \sigma_{t-1}^2,$$

where r_t denotes the observed log-returns of Bitcoin and μ_t the conditional mean. z_t is a standardized i.i.d. error term with $\mathrm{E}[z_t] = 0$ and $\mathrm{V}[z_t] = 1$, formally $z_t \sim i.i.d.(0, 1)$. In the following analysis z_t is specified as following either a normal distribution or a Student's t-distribution. σ_t is a conditionally deterministic function depending on the history of the process. To ensure stability it must hold that $\alpha_1 + \beta_1 < 1$.

The exponential GARCH (EGARCH) model is an asymmetric GARCH model introduced by Nelson (1991) that considers leverage effects between positive and negative shocks.

The conditional variance in the EGARCH(1,1) model is given by

$$\log(\sigma_t^2) = \omega + \alpha_1 z_{t-1} + \gamma_1(|z_{t-1}| - E[|z_{t-1}|]) + \beta_1 \log(\sigma_{t-1}^2),$$

where α_1 describes the sign effect and γ_1 captures the magnitude of z_t. An advantage of the EGARCH model is that, because of the exponential functional form, the model does not need any constraints on coefficient parameters, it is always defined.

The GJR-GARCH model introduced by Glosten et al. (1993) is another modification of the standard GARCH model. It is also an asymmetric model that values positive and negative shocks of the conditional variance differently. Using the indicator function $I(.)$, the conditional variance of the GJR-GARCH(1,1) is given by

$$\sigma_t^2 = \omega + \alpha_1 r_{t-1}^2 + \gamma_1 r_{t-1}^2 I_{r_{t-1}<0} + \beta_1 \sigma_{t-1}^2,$$

where γ_1 covers the leverage effect.

The realized GARCH model by Hansen et al. (2012) exploits realized measures of volatility by including a measurement equation. The measurement equation links the observed realized measure to the latent volatility of the returns (Hansen et al. (2012)). The realized GARCH model also accounts for leverage effects. This asymmetric reaction to shocks is a useful property when modeling the conditional variance of stock returns because, as pointed out by Black (1976), positive and negative news may affect future volatility asymmetrically. The following analysis uses the realized GARCH(1,1) model with a log-linear specification suggested by Hansen et al. (2012). Formally, the model is described by the following GARCH and measurement equation:

$$\begin{aligned} \log(\sigma_t^2) &= \omega + \alpha_1 \log(x_{t-1}) + \beta_1 \log(\sigma_{t-1}^2) \\ \log(x_t) &= \xi + \delta \log(\sigma_t^2) + \tau(z_t) + u_t, \quad u_t \sim N(0, \lambda) \end{aligned} \tag{5}$$

$\tau(z_t) = \eta_1 z_t + \eta_2(z_t^2 - 1)$ in equation (5) serves as leverage function and is a simple quadratic function on the basis of Hermite polynomials.

In contrast to the standard GARCH(1,1) model the first lag of the squared return is replaced with the first lag of a realized measure x_t. The volatility process is stable as long as $\beta_1 + \delta\alpha_1 \in (-1, 1)$. For more details, see Hansen et al. (2012). In this project, we use two realized measures, i.e. $x_t \in \{\mathrm{RV_t}, \mathrm{MedRV_t}\}$. The realized variance (RV) in period t is defined as

$$\mathrm{RV}_t = \sum_{i=1}^{N} r_{t,i}^2, \tag{6}$$

which sums up N squared returns over the entire day. $r_{t,i} = log(p_{t,i}) - log(p_{t,i-1})$, where $p_{t,i}$ denotes the ith closing price in period t. Furthermore, $N = 1/\Delta$, where Δ denotes the sampling frequency. For calculating RV, we use 5-min high-frequency data as suggested by Andersen et al. (2008). The second realized measure is the median realized variance (MedRV) estimator introduced by Andersen et al. (2012) that is robust to price jumps and to the presence of zero intraday returns in finite samples. The estimator is defined by the following formula:

$$\mathrm{MedRV}_t = \frac{\pi}{6 - 4\sqrt{3} + \pi} \left(\frac{N}{N-2}\right) \sum_{i=2}^{N-1} \mathrm{med}(|r_{t,i-1}|, |r_{t,i}|, |r_{t,i+1}|)^2 \tag{7}$$

Using high-frequency data and measuring the realized variance ex-post has become a popular research field in the early 2000s. One of the most popular estimators for the ex-post variance of an asset is the realized variance estimator. Modeling the RV was challenging in the beginning, as it is difficult to account for the long-memory property of the ex-post measure of the return variance with simple ARIMA models. One solution to overcome this issue is to use so-called autoregressive fractionally integrated moving average (ARFIMA) models, which handle the long memory of time series by generalizing the integer differencing order d of ARIMA models to real-valued differences. A disadvantage of such models is that they are difficult to estimate; especially the estimation of d is a demanding task.

Corsi (2009) developed the first simple long-memory model for estimating and forecasting the realized ex-post variance. The Heterogeneous Autoregressive model of Realized Volatility (HAR-RV) is based on the Heterogeneous Market Hypothesis by Müller et al. (1997) who claim that agents have heterogeneous preferences in terms of the time horizon

of their investment decisions. On the one hand, there are dealers and market makers who trade on a daily basis; on the other hand, there are insurance companies and pension funds that trade at a lower frequency. Agents react to new information differently and as a result create volatility. In detail, Corsi (2009) assumes that the daily realized volatility depends on the last daily, weekly and monthly realized volatility caused by short-term, medium-term and long-term investors, respectively. Many extensions of the model have emerged and were intensively discussed in the literature. As a next step, the theoretical framework of the simple HAR-RV model is described.

Let p_t denote the logarithmic price of Bitcoin in period t. Then the diffusion process is given by the following stochastic differential equation

$$dp_t = \mu_t dt + \sigma_t dW_t, \tag{8}$$

where μ_t describes a drift with a finite and continuous variation process, σ_t is a cadlág stochastic volatility process independent of W_t, and W_t denotes a standard Brownian motion. The integrated variance (IV) which is equivalent to the quadratic variation (QV) of this process for one trading day is then defined by

$$QV_t = IV_t = \int_{t-1}^{t} \sigma_s^2 ds$$

It can be shown that the integrated variance can be consistently estimated using the realized variance (RV) if the number of squared intraday returns goes to infinity, which is equivalent to $\Delta \to 0$. Formally,

$$\text{plim}_{N\to\infty} RV_t = IV_t = QV_t.$$

This simple framework, however, does not incorporate any jumps in the price process, whereas Hung et al. (2020) found that Bitcoin is very prone to jumps. For this reason, as a next step we consider a jump-diffusion model that was introduced by Merton (1976). Now

the jump-diffusion process is a Brownian semimartingale given by

$$dp_t = \mu_t dt + \sigma_t dW_t + \kappa_t dq_t, \tag{9}$$

where q_t is a Poisson process counting the number of jumps in the process and κ_t measures the magnitude of the corresponding discrete jumps. The quadratic variation of this process has now two components, namely

$$QV_t = IV_t + JV_t,$$

where $\sum_{t-1<s\leq t} \kappa_s^2$ is the jump variation (JV) or discontinuous variation which is simply the sum of squared jump sizes for a given period. Even though the quadratic variation now has two components the realized variance estimator is still consistent. When the sample points within period t approach infinity

$$\text{plim}_{N\to\infty} RV_t = QV_t = IV_t + JV_t.$$

As a result, RV_t contains the continuous variation and the discontinuous variation of the jumps in the price process. One may be interested in estimating just the continuous part IV_t of the process. Barndorff-Nielsen and Shephard (2004) introduced the realized bipower variation estimator given by

$$\text{BV}_t = \mu_1^{-2} \sum_{i=2}^{N} |r_{t,i-1}||r_{t,i}|, \tag{10}$$

where

$$\mu_p = \text{E}(|Z|^p) = 2^{p/2} \frac{\Gamma((1+p)/2)}{\Gamma(1/2)} \tag{11}$$

$Z \sim N(0,1), p \geq 0$, and $\Gamma(.)$ denotes the Gamma function. In particular, it follows that $\mu_1 = \sqrt{2/\pi}$. The authors show that, in the presence of jumps, $\text{plim}_{N\to\infty} BV_t = IV_t$. Consequently, the difference between the realized variance and the realized bipower variation

consistently measures the jump variation (JV), which means in mathematical terms

$$\text{plim}_{N\to\infty}(RV_t - BV_t) = JV_t = \Sigma_{t-1<s\leq t}\kappa_s^2$$

Of course, any other consistent estimator of the integrated variance in the presence of jumps can be used. The MedRV estimator is an alternative to the realized bipower variation estimator, as it was shown by Andersen et al. (2012) that, in the presence of jumps,

$$\text{plim}_{N\to\infty}(MedRV_t) = IV_t.$$

In the following, we also use this newer jump robust estimator when estimating the discontinuous part of the realized variance. To ensure the positivity of the jump variation we apply the max-function and define the discontinuous jump variation as

$$JV_t = \max\{RV_t - MedRV_t, 0\}$$

This would imply that there is a non-negative jump variation every day, which is not plausible since there should be significant and insignificant jumps in the price process. To detect significant price jumps the JO Jump test by Jiang and Oomen (2008) is performed to test for the presence of jumps in the high-frequency price series of Bitcoin. Theodosiou and Zikes (2011) compared several tests for jumps in the price series and they found that the JO jump test has the highest power among all other tests considered at a 5-min sampling frequency. In addition, the swap variance test by Jiang and Oomen (2008) also performs well in the presence of zero intraday returns because of the usage of the integrated sixticity in the denominator of the test statistic (Theodosiou and Zikes (2011)). Jiang and Oomen (2008) found that the accumulated difference of the simple arithmetic return $R_{i,n}$ and log-return $r_{i,n}$ should equal half the realized variance in the absence of jumps. If this difference is too large (in absolute terms), this would indicate the existence of jumps. For other suggestions of jump tests, see Ait-Sahalia and Jacod (2009) and Barndorff-Nielsen and Shephard (2006).

The null and alternative hypothesis of the JO jump test is given by:

$$\begin{aligned} H_0 &: \text{There are no jumps in period} \quad t \\ H_A &: \text{There is at least one jump in period} \quad t. \end{aligned}$$

If there are N equispaced returns in period t and $N \to \infty$, the test statistic of the ratio test is given by

$$\text{JOjumpTest}_{t,N} = \frac{NBV_t}{\sqrt{\Omega_{SwV}}} * \left(1 - \frac{RV_t}{SwV_t}\right) \xrightarrow{d} N(0,1),$$

where $S\omega V_t = 2\Sigma_{i=1}^{N}(R_{t,i} - r_{t,i})$, BV is the bipower variation given by (10) and RV denotes the realized variance stated in equation (6). Furthermore

$$\Omega_{SwV} = \frac{\mu_6}{9} \frac{N^3 \mu_{6/p}^{-p}}{N-p-1} \sum_{i=0}^{N-p} \prod_{k=1}^{p} |r_{t,i+k}|^{6/p},$$

is an estimator of the integrated sixticity, $\int_{t-1}^{t} \sigma_s^6 ds$. μ_6 of Ω_{SwV} is the same as in equation (11) evaluated at $p = 6$.

The estimator of the integrated sixticity depends on a power parameter p, which is jump-robust for $p = 4$ and $p = 6$ (Jiang and Oomen (2008)). We choose $p = 4$, as using higher powers of returns can make the estimator upward-biased, which leads to a deterioration of the power of the test. In fact, Jiang and Oomen (2008) found that using either the realized quadpower or sixthpower sixticity estimator makes little difference, which is also in line with the findings of Theodosiou and Zikes (2011).

Now it is possible to distinguish the continuous and discontinuous variation of the realized variance. Formally, the continuous part is given by

$$C_t = I(\text{JOjumpTest}_{t,N} \leq \phi_\alpha) * RV_t + I(JOjumpTest_{t,N} > \phi_\alpha) * MedRV_t, \tag{12}$$

where $I(.)$ is the indicator function and ϕ_α is the α–quantile of the standard normal dis-

tribution. The corresponding discontinuous jump variation of the quadratic variation is defined as

$$J_t = I(\text{JOjumpTest}_{t,N} > \phi_\alpha) * \max\{RV_t - MedRV_t, 0\} \tag{13}$$

We use the 5% significance level for identifying statistically significant jumps.

5.2. HAR and related models

The simplest version of all Heterogeneous Autoregression of Realized Volatility (HAR-RV) models was introduced by Corsi (2009) and assumes that the price process of Bitcoin is generated by equation (8) which implies a continuous price process and the absence of any jumps. Then, the HAR-RV model has the following form:

$$RV_t^{(d)} = \beta_0 + \beta_1 RV_{t-1}^{(d)} + \beta_2 RV_{t-1}^{(w)} + \beta_3 RV_{t-1}^{(m)} + \varepsilon_t, \tag{14}$$

where ε_t is a mean-zero error term and $RV_{t-1}^{(d)}, RV_{t-1}^{(w)}$ and $RV_{t-1}^{(m)}$ are the corresponding first lag of daily, average weekly and average monthly realized variances which are defined as

$$\begin{aligned} RV_{t-1}^{(w)} &= \frac{1}{7}(RV_{t-7} + RV_{t-6} + \ldots + RV_{t-1}), \\ RV_{t-1}^{(m)} &= \frac{1}{30}(RV_{t-30} + RV_{t-29} + \ldots + RV_{t-1}) \end{aligned} \tag{15}$$

In contrast to most applications, we use seven days and 30 days when computing the average weekly and monthly realized variance, respectively. BTC can be traded seven days a week and 24 hours a day. There do not exist any trading hours and trading days as with typical stock exchanges. Apart from that, there exist ample studies that show that the typical assumption of a homoskedastic and normally distributed error term is violated, therefore we also use the logarithmic transformation of equation (14) as suggested by Corsi (2009). Using the logarithm of the realized variance has two advantages. On the one hand, it ensures the positivity of the partial variances and on the other hand it reduces the heteroskedasticity of the error term and the assumption $\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$ holds approximately due to the log-normal property of the realized variance. The logarithmic version takes the

form

$$\log(RV_t^{(d)}) = \beta_0 + \beta_1 \log(RV_{t-1}^{(d)}) + \beta_2 \log(RV_{t-1}^{(w)}) + \beta_3 \log(RV_{t-1}^{(m)}) + \varepsilon_t.$$

The HAR-RV-J model was introduced by Andersen et al. (2007). In contrast to the HAR-RV model, it does not assume a continuous price process but a jump diffusion process given by (9). This model uses an additional explanatory variable when predicting the daily realized variance. The HAR-RV-J is defined as

$$RV_t^{(d)} = \beta_0 + \beta_1 RV_{t-1}^{(d)} + \beta_2 RV_{t-1}^{(w)} + \beta_3 RV_{t-1}^{(m)} + \beta_4 J_{t-1}^{(d)} + \varepsilon_t, \tag{16}$$

where $J_{t-1}^{(d)}$ is a jump component estimated by (13). Applying a log-transformation of (16) is complicated by the existence of days with no jump and with a zero jump component. To deal with this issue, Andersen et al. (2007) suggest the logarithmic HAR-RV-J model

$$\begin{aligned} \log(RV_t^{(d)}) &= \beta_0 + \beta_1 \log(RV_{t-1}^{(d)}) + \beta_2 \log(RV_{t-1}^{(w)}) + \beta_3 \log(RV_{t-1}^{(m)}) \\ &\quad + \beta_4 \log(1 + J_{t-1}^{(d)}) + \varepsilon_t \end{aligned}$$

A further extension of the standard HAR-RV model was again introduced by Andersen et al. (2007). The main idea of the HAR-RV-CJ model is to use the property of the realized variance and decompose the RV into its continuous and discontinuous component as described in (12) and (13). This yields six explanatory variables instead of the three in HAR-RV model. Formally, the HAR-RV-CJ model is defined as

$$RV_t^{(d)} = \beta_0 + \beta_1 C_{t-1}^{(d)} + \beta_2 C_{t-1}^{(w)} + \beta_3 C_{t-1}^{(m)} + \beta_4 J_{t-1}^{(d)} + \beta_5 J_{t-1}^{(w)} + \beta_6 J_{t-1}^{(m)} + \varepsilon_t, \tag{17}$$

where

$$
\begin{aligned}
C_{t-1}^{(w)} &= \frac{1}{7} * (C_{t-7} + C_{t-6} + \ldots + C_{t-1}) \\
J_{t-1}^{(w)} &= \frac{1}{7} * (J_{t-7} + J_{t-6} + \ldots + J_{t-1}) \\
C_{t-1}^{(m)} &= \frac{1}{30} * (C_{t-30} + C_{t-29} + \ldots + C_{t-1}) \\
J_{t-1}^{(m)} &= \frac{1}{30} * (J_{t-30} + J_{t-29} + \ldots + J_{t-1})
\end{aligned}
$$

The logarithmic HAR-RV-CJ model takes the form

$$
\begin{aligned}
\log(RV_t^{(d)}) &= \beta_0 + \beta_1 \log(C_{t-1}^{(d)}) + \beta_2 \log(C_{t-1}^{(w)}) + \beta_3 \log(C_{t-1}^{(m)}) \\
&\quad + \beta_4 \log(1 + J_{t-1}^{(d)}) + \beta_5 \log(1 + J_{t-1}^{(w)}) + \beta_6 \log(1 + J_{t-1}^{(m)}) + \varepsilon_t
\end{aligned}
$$

5.3. Evaluation of model performance

For the evaluation, the data set is divided into a training set from 31st Mar 2015 to 31st Mar 2020 containing 1,827 data points and a test set of the last year of the data set (1st Apr 2020 to 31st Mar 2021) containing 365 data points. In total sixteen different models are estimated and subjected to an in-sample and out-of-sample performance evaluation. The following ten GARCH models are estimated:

- Standard GARCH with normally distributed standardized error terms (sGARCH-norm)
- Standard GARCH with t-distributed standardized error terms (sGARCH-t)
- EGARCH with normally distributed standardized error terms (EGARCH-norm)
- EGARCH with t-distributed standardized error terms (EGARCH-t)
- GJR-GARCH with normally distributed standardized error terms (GJR-GARCH-norm)
- GJR-GARCH with t-distributed standardized error terms (GJR-GARCH-t)

- Realized GARCH with normally distributed standardized error terms and realized variance as realized measure (RealGARCH-RV-norm)
- Realized GARCH with t-distributed standardized error terms and realized variance as realized measure (RealGARCH-RV-t)
- Realized GARCH with normally distributed standardized error terms and median realized variance as realized measure (RealGARCH-MedRV-norm)
- Realized GARCH with t-distributed standardized error terms and median realized variance as realized measure (RealGARCH-MedRV-t)

Furthermore, six different HAR-RV models are estimated, containing

- HAR-RV
- Log-HAR-RV
- HAR-RV-J
- Log-HAR-RV-J
- HAR-RV-CJ
- Log-HAR-RV-CJ

For comparing the in-sample fit of the models, four different information criteria including the Akaike (AIC), Bayesian (BIC), Hannan-Quinn (HQIC) and Shibata (SIC) information criteria are used as well as the maximum value of the log-likelihood function. The model with the lowest information criteria and highest log-likelihood is considered as the best model. Formally, the information criteria are defined as follows:

$$\begin{aligned} AIC &= \frac{-2LL}{N} + \frac{2m}{N}, \\ BIC &= \frac{-2LL}{N} + \frac{m \log N}{N}, \\ HQIC &= \frac{-2LL}{N} + \frac{2m \log(\log(N))}{N}, \\ SIC &= \frac{-2LL}{N} + \log \frac{N + 2m}{N} \end{aligned}$$

where LL denotes the maximum value of the log-likelihood function, N is the number of observations and m stands for the number of parameters. However, this procedure

does not make sense when comparing GARCH models with the HAR-RV models since the conditional variance of daily returns estimated from GARCH models might not, on average, correspond to the realized measures estimated from intraday data. Even though these two types of models are not directly comparable because they do not exactly predict the same volatilities, in order to perform an in-sample comparison we use the mean squared error (MSE) and Quasi-Likelihood (QLIKE) loss function. When performing the forecasting analysis, we conduct a one-step ahead forecast where we estimate the model with the training data and re-estimate the model after each shift in start and end of the estimation interval. For evaluating performance, the MSE and QLIKE loss functions are used, as these two are robust when evaluating the out-of-sample performance by divergence between the predicted values and the volatility proxy of an observable variable (Patton (2011)). The two loss functions are defined by

$$\begin{aligned} MSE &= \frac{1}{N}\sum_{t=1}^{N}\left(\sigma_t^2 - \hat{\sigma}_t^2\right)^2 \\ QLIKE &= \frac{1}{N}\sum_{t=1}^{N}\left(\frac{\sigma_t^2}{\hat{\sigma}_t^2} - \log\left(\frac{\sigma_t^2}{\hat{\sigma}_t^2}\right) - 1\right) \end{aligned}$$

Moreover, for an out-of-sample robustness check the realized variance and MedRV serve as a proxy for the true variance σ_t^2 in order to compare GARCH and HAR models. In Panel A of Table 17, HAR-RV is considered as a benchmark model by looking whether any other model outperforms the standard HAR-RV model when applying the Diebold-Mariano test at the 5% significance level. Panel B uses MedRV as a variance proxy and evaluates which model is outperformed by the RealGARCH-MedRV-std. The Diebold-Mariano test is applied at the same significance level. This test was originally introduced by Diebold and Mariano (1995), but we use the modified version due to Harvey et al. (1997). The orginal test assumes uncorrelated forecast errors, while the modified version allows for autocorrelation in the errors. The test works as follows: Let $d_t = L(\hat{x}_t^{(1)}) - L(\hat{x}_t^{(2)})$, where $L(\hat{x}_t^{(i)})$ is an out-of-sample loss function, i.e. $L(\hat{x}_t^{(i)}) \in \{MSE, QLIKE\}$ for forecast $\hat{x}_t^{(i)}$ obtained with model M_i. Assuming d_t is covariance-stationary and has finite moments, the

null hypothesis is equal predictive accuracy or formally, $H_0 : E(d_t) = 0$. Moreover, the test statistic is given by

$$DM = \frac{\frac{1}{N}\sum d_t}{\sqrt{\hat{\sigma}_d^2}} \xrightarrow{d} N(0,1),$$

where $\hat{\sigma}_d^2$ is a long-run variance estimate of $\frac{1}{N}\sum d_t$. For details, see Diebold and Mariano (1995) and Harvey et al. (1997). If DM is significantly < 0 (> 0), one can conclude that model M_1 is better (worse) than M_2. Furthermore, we use the Mincer-Zarnowitz test proposed by Mincer and Zarnowitz (1969) for out-of-sample evaluation by using the same two proxies mentioned previously. The test works as follows:

1. Estimate the simple OLS regression (Mincer-Zarnowitz regression): $\theta_t = \alpha + \beta\hat{\theta}_t + e_t$, where θ_t is the true value, i.e. $\theta_t \in \{RV_t, MedRV_t\}$ and the regressor $\hat{\theta}_t$ is the value of the forecast, i.e. $\hat{\theta}_t \in \{RV_t, \sigma_t^2\}$

2. Test the joint hypothesis $H_0 : \alpha = 0, \beta = 1$ by using an F-test and focus on the R^2.

3. If the null hypothesis gets rejected the forecast is considered as being biased and inefficient

However, as pointed out by Andersen and Bollerslev (1998), σ_t^2 is often subject to an estimation error which may cause biased estimates for β for GARCH models. To overcome this problem the authors suggest concentrating on the R^2 of the Mincer-Zarnowitz regression instead. Therefore, we also use the R^2 as decision criterion and consider the model with the highest R^2 as the best forecasting model.

5.4. Empirical results

Jump test results

Table 14 shows the numbers of rejections of the null hypothesis of no jumps for different significance levels. The results indicate that there are many rejections of the null, which suggests the usage of a jump-robust estimator like the MedRV estimator. Even with a significance level of 0.1% the null hypothesis is rejected 392 times which equates approxi-

Table 14
JO Jump test results, Number of days: 2,192

Significance level	$\alpha = 5\%$	$\alpha = 1\%$	$\alpha = 0.1\%$
Number of rejections	713	537	392

Table 15
In-sample fit (GARCH Models)

Model	Point estimates									
	μ	ω	α_1	β_1	γ_1	δ	ξ	η_1	η_2	λ
sGARCH-norm	0.0022**	0.0001***	0.206***	0.791***						
sGARCH-std	0.0017***	0.00002***	0.131***	0.868***						
EGARCH-norm	0.0016*	-0.4501***	-0.058***	0.925***	0.331***					
EGARCH-std	0.0016***	-0.06	0.053**	0.99***	0.312***					
GJR-GARCH-norm	0.0015**	0.0001***	0.159***	0.779***	0.103***					
GJR-GARCH-std	0.0017***	0.00001***	0.146***	0.879***	-0.052**					
RealGARCH-RV-norm	0.0016**	-0.988***	0.406***	0.432***		1.22***	1.237***	-0.028*	0.065***	0.615***
RealGARCH-RV-std	0.0016***	0.2538	0.508***	0.456***		0.953***	-1.152***	-0.061**	0.108***	0.624***
RealGARCH-MedRV-norm	0.0014	-1.299***	0.378***	0.405***		1.373***	2.111***	-0.024	0.055***	0.606***
RealGARCH-MedRV-std	0.0015***	0.390	0.554***	0.409***		0.948***	-1.345***	-0.046**	0.106***	0.601***

Notes: *, **, and *** denote significance at the 10%, 5%, and 1% level, respectively. Values in parentheses are HAC standard errors to account for the presence of autocorrelation and heteroskedasticity in the error terms.

mately every 5th trading day.

In-sample results

Table 15 presents estimates for all ten GARCH models, with most coefficients significant at least at the 5% level. Notably, the conditional mean across all models hovers around zero and all models are stable (asymptotically stationary). Of particular interest, RealGARCH models with t-distributed innovations ($\alpha_1 = 0.508$ and 0.554) assign greater weight to the realized measure compared to normally distributed returns ($\alpha_1 = 0.406$ and 0.378). This implies a more responsive reaction to sudden volatility changes. Consequently, in scenarios like the onset of the COVID-19 crisis, realized GARCH models with t-distributed standardized returns offer improved conditional variance estimation.

Table 16 further reveals that logarithmic transformations in HAR models result in significantly higher R^2 values. Additionally, coefficients in the continuous component of realized variance are highly significant. Notably, the jump component in HAR-RV-J only shows significance at the 10% level, while the Log-HAR-RV-J model reaches 1% significance. Surprisingly, jump variations in HAR-RV-CJ and Log-HAR-RV-CJ are all insignificant,

Table 16
In-sample fit (HAR Models)

Model	Point estimates and R^2							
	β_0	β_1	β_2	β_3	β_4	β_5	β_6	R^2
HAR-RV	0.0005***	0.3885***	0.1559***	0.2401***				0.267
	(0.0002)	(0.0668)	(0.0457)	(0.0825)				
Log-HAR-RV	-0.7769***	0.5045***	0.2737***	0.1213***				0.6227
	(0.1447)	(0.0281)	(0.0339)	(0.0398)				
HAR-RV-J	0.0005***	0.4293***	0.1645***	0.2344***	-0.5656*			0.2716
	(0.0002)	(0.0846)	(0.0472)	(0.0808)	(0.3317)			
Log-HAR-RV-J	-0.5905***	0.5425***	0.2645***	0.1156***	-100.58***			0.6253
	(0.1453)	(0.0294)	(0.0342)	(0.0388)	(27.4939)			
HAR-RV-CJ	0.0006***	0.4103***	0.1267**	0.2908***	-0.0098	0.5187	-1.0096	0.2699
	(0.0002)	(0.0702)	(0.0575)	(0.0814)	(0.3891)	(0.5490)	(0.8421)	
Log-HAR-RV-CJ	-0.6903***	0.5258***	0.2358***	0.1368***	15.219	-2.4985	-49.4889	0.628
	(0.1592)	(0.0270)	(0.0360)	(0.0437)	(38.2334)	(103.1854)	(161.6001)	

Notes: *, **, and *** denote significance at the 10%, 5%, and 1% level, respectively. Values in parentheses are HAC standard errors to account for the presence of autocorrelation and heteroskedasticity in the error terms.

Table 17
In-sample results – Loss functions

	Panel A: RV as proxy		Panel B: MedRV as proxy	
Model	MSE	QLIKE	MSE	QLIKE
sGARCH-norm	**0.1451**	0.3653	0.1446	0.3610
sGARCH-std	0.1585	0.4748	0.1604	0.4531
EGARCH-norm	0.1784	0.3783	0.1579	0.3770
EGARCH-std	0.1815	0.4396	0.1930	0.4796
GJR-GARCH-norm	0.1454	0.3773	**0.1427**	0.3712
GJR-GARCH-std	0.1639	0.4868	0.1666	0.4625
RealGARCH-RV-norm	0.1666	**0.3480**	0.1734	**0.3348**
RealGARCH-RV-std	0.3143	0.4810	0.3453	0.5280
RealGARCH-MedRV-norm	0.1661	0.3517	0.1724	0.3353
RealGARCH-MedRV-std	0.4083	0.5072	0.4426	0.5542
HAR-RV	0.1580	0.3529		
Log-HAR-RV	0.1606	0.3924		
HAR-RV-J	0.1570	0.3576		
Log-HAR-RV-J	0.1606	0.3907		
HAR-RV-CJ	0.1574	0.3571		
Log-HAR-RV-CJ	0.1586	0.3908		

Notes: The MSE is multiplied by 10^5

Table 18
Information criteria and Log-Likelihood

Model	AIC	BIC	HQIC	SIC	LL
sGARCH-norm	-3.824	-3.8115	-3.8191	-3.8236	3496.9
sGARCH-std	-4.141	-4.1254	-4.1349	-4.1405	3787.3
EGARCH-norm	-3.8257	-3.8106	-3.8201	-3.8257	3499.8
EGARCH-std	**-4.1624**	**-4.1444**	**-4.1558**	**-4.1625**	**3808.4**
GJR-GARCH-norm	-3.8287	-3.8136	-3.8231	-3.8287	3502.5
GJR-GARCH-std	-4.1418	-4.1237	-4.1352	-4.1419	3789.6
RealGARCH-RV-norm	-3.7768	-3.7496	-3.7668	-3.7768	3459.1
RealGARCH-RV-std	-4.1233	-4.0961	-4.1132	-4.1233	3775.6
RealGARCH-MedRV-norm	-3.7585	-3.7314	-3.7485	-3.7586	3442.4
RealGARCH-MedRV-std	-4.1123	-4.0852	-4.1023	-4.1124	3765.6

possibly due to the lacking overnight component because of continuous Bitcoin trading.

Table 17 enables a potential in-sample performance evaluation of two different types of volatility models with robust loss functions. Intriguingly, the GARCH family outperforms the HAR family. Panel A shows that sGARCH-norm exhibits the lowest MSE, and RealGARCH-RV-norm displays the lowest QLIKE. In Panel B, RealGARCH-RV-norm retains the lowest QLIKE, while GJR-GARCH-norm yields the smallest MSE.

Table 18 presents information criteria and the maximum Log-Likelihood value (LL) for the ten estimated GARCH models. To compare classical GARCH models with realized GARCH models, we exclusively employ the partial log-likelihood of the realized GARCH models. Notably, the EGARCH-std model is identified as having the best in-sample fit according to AIC, BIC, HQIC, SIC, and Log-Likelihood values. Interestingly, realized GARCH models exhibit inferior in-sample performance compared to GARCH models without a realized measure. The next section explores if these findings hold out-of-sample.

Out-of-sample results

The out-of-sample analysis, as presented in Table 19, unveils noteworthy insights. Firstly, we observe that realized GARCH models featuring t-distributed error terms consistently outperform all other considered GARCH models. This superiority holds true for both Panel A and Panel B.

Table 19
Out-of-sample results : Loss functions

	Panel A: RV as proxy		Panel B: MedRV as proxy	
Model	MSE	QLIKE	MSE	QLIKE
sGARCH-norm	0.0527	0.3041	0.0524	0.3318**
sGARCH-std	0.0548	0.3068	0.0541	0.3189**
EGARCH-norm	0.0536	0.3358	0.0527	0.3619**
EGARCH-std	0.0697	0.4244	0.0749**	0.4975**
GJR-GARCH-norm	0.0539	0.3295	0.0532	0.3507**
GJR-GARCH-std	0.0542	0.2879	0.0537	0.3035**
RealGARCH-RV-norm	0.0711	0.4767	0.0749**	0.5502**
RealGARCH-RV-std	0.0493	0.2490*	0.0487	0.2537**
RealGARCH-MedRV-norm	0.0993	0.5719	0.1050**	0.6558**
RealGARCH-MedRV-std	0.0455	**0.2399***	**0.0434**	**0.2203**
HAR-RV	0.0475	0.3136		
Log-HAR-RV	0.0462	0.2665*		
HAR-RV-J	0.0502	0.9138		
Log-HAR-RV-J	0.0463	0.3131		
HAR-RV-CJ	0.0476	0.3171		
Log-HAR-RV-CJ	**0.0451***	0.2594*		

Notes: The MSE is multiplied by 10^5, * in Panel A denotes that the MSE/QLIKE of the corresponding model is significantly less than the MSE/QLIKE of the HAR-RV model when applying the Diebold-Mariano test at the 5% significance level, ** in Panel B denotes that the MSE/QLIKE of the corresponding model is significantly greater than the MSE/QLIKE of the RealGARCH-MedRV-std model when applying the Diebold-Mariano test at the 5% significance level

A key revelation is the performance of the RealGARCH-MedRV-std model. It displays lower MSE and QLIKE values compared to its counterpart, the RealGARCH-RV-std. This suggests that the incorporation of a jump-robust realized measure, such as MedRV, significantly enhances forecast accuracy.

Furthermore, despite EGARCH-std showing the best in-sample fit according to information criteria, its out-of-sample performance is notably weaker, ranking third in terms of both MSE and QLIKE. This highlights the discrepancy between in-sample fitness and out-of-sample forecasting accuracy.

In Panel A, the Log-HAR-RV-CJ model excels, boasting the lowest MSE and emerging as the only model with a statistically significantly smaller MSE than the HAR-RV model. When evaluated using the QLIKE loss function, four models, including both realized GARCH models with t-distributed innovations, outperform the HAR-RV model.

Panel B underscores the supremacy of the RealGARCH-MedRV-std, with the smallest MSE and QLIKE values. While the difference in MSE is statistically significant for only three GARCH models, the RealGARCH-MedRV-std outperforms all other GARCH models at the 5% significance level when the QLIKE loss function is considered.

Further, the introduction of a more recent realized GARCH model employing t-distributed error terms proves to be a significant development. This model stands as the sole contender against the HAR models, with the RealGARCH-MedRV-std exhibiting the smallest QLIKE across all models. The asymmetry of the QLIKE loss function indicates a systematic overestimation of realized volatility by the RealGARCH-MedRV-std.

Within the realm of HAR models, the application of log specifications for forecasting outperforms using levels. Moreover, the Log-HAR-RV-CJ model emerges as the best performing model even though the three jump components are all insignificant.

Our examination of the Mincer-Zarnowitz test, as detailed in Table 20, reveals intriguing and somewhat unexpected outcomes. On one hand, the null hypothesis is rejected in Panel A and Panel B at the 5% significance level for several models, suggesting a prevailing bias in forecasts. Surprisingly, models where the null hypothesis is rejected tend to exhibit the highest R-squared values, indicating a superior ability to explain variations in the dependent

Table 20

Out-of-sample results - Mincer-Zarnowitz test

	Panel A: RV as proxy				Panel B: MedRV as proxy			
			p-value				p-value	
	α	β	$H_0 : \alpha = 0, \beta = 1$	R^2	α	β	$H_0 : \alpha = 0, \beta = 1$	R^2
Model								
sGARCH-norm	-0.0001	1.202	0.0857	0.1963	-0.0002	1.103	0.5778	0.2037
	(0.0002)	(0.1276)			(0.0002)	(0.1145)		
sGARCH-std	0.000003	1.150	0.0849	0.194	-0.0001	1.085	0.7064	0.1766
	(0.0002)	(0.1230)			(0.0002)	(0.1229)		
EGARCH-norm	-0.0006	1.56	0.0003	0.2034	-0.0007	1.453	0.0078	0.2172
	(0.0003)	(0.1621)			(0.0003)	(0.1448)		
EGARCH-std	-0.0002	0.662	≈ 0	0.2406	-0.0003	0.632	≈ 0	0.2242
	(0.0002)	(0.0618)			(0.0002)	(0.0617)		
GJR-GARCH-norm	-0.0001	1.228	0.0249	0.1828	-0.0002	1.133	0.5469	0.1916
	(0.0002)	(0.1363)			(0.0002)	(0.1222)		
GJR-GARCH-std	0.00002	1.099	0.2492	0.1984	-0.0001	1.039	0.9445	0.1812
	(0.0002)	(0.1160)			(0.0002)	(0.1159)		
RealGARCH-RV-norm	-0.0002	0.688	≈ 0	0.2654	-0.0004	0.677	≈ 0	0.3169
	(0.0002)	(0.0601)			(0.0002)	(0.0522)		
RealGARCH-RV-std	0.0003	0.725	≈ 0	0.3185	0.0001	0.729	≈ 0	0.3309
	(0.0002)	(0.0555)			(0.0002)	(0.0544)		
RealGARCH-MedRV-norm	-0.0003	0.5764	≈ 0	0.2917	-0.0006	0.5713	≈ 0	0.3353
	(0.0002)	(0.047)			(0.0002)	(0.0406)		
RealGARCH-MedRV-std	0.0004	0.710	≈ 0	**0.3864**	0.0002	0.723	≈ 0	**0.4111**
	(0.0001)	(0.0468)			(0.0001)	(0.0454)		
HAR-RV	-0.0001	0.9738	0.2401	0.2975				
	(0.0002)	(0.0785)						
Log-HAR-RV	0.0001	1.1669	0.0013	0.3335				
	(0.0002)	(0.0862)						
HAR-RV-J	0.00001	0.8601	0.0127	0.2696				
	(0.0002)	(0.0743)						
Log-HAR-RV-J	0.0002	1.0460	0.0230	0.3238				
	(0.0002)	(0.0793)						
HAR-RV-CJ	-0.0001	0.9351	0.0842	0.2995				
	(0.0002)	(0.0751)						
Log-HAR-RV-CJ	0.0002	1.0965	0.0092	0.3443				
	(0.0002)	(0.0794)						

Notes: Values in parentheses are the corresponding standard errors

variable. Of particular note is the RealGARCH-MedRV-std model, which achieves the highest R-squared value in Panel A, surpassing all GARCH and HAR models. This occurs despite the rejection of the null hypothesis at a very low significance level and a systematic tendency to overestimate forecasts. In Panel B, the RealGARCH-MedRV-std once again outperforms all other GARCH models in terms of R-squared, even though it produces biased forecasts. In conclusion, our findings from the Mincer-Zarnowitz test challenge conventional expectations, revealing an intriguing relationship between bias and explanatory power in forecasts. The RealGARCH-MedRV-std consistently emerges as the superior forecasting model, aligning with prior research in this domain.

6. Conclusion

In conclusion, our cointegration analysis reveals that the ten considered cryptocurrencies exhibit common stochastic trends, allowing for the creation of mean-reverting portfolios. Out-of-sample results underscore the profitability of trading strategies during the test period, indicating a robust long-term relationship. Notably, cointegrated portfolios based on the Johansen procedure consistently outperform those constructed with the Engle-Granger two-step procedure. Furthermore, $spread_t^{J,2}$, exhibiting superior in-sample performance, leads to the highest out-of-sample returns. In contrast to volatility modeling, where in-sample fitness does not guarantee out-of-sample success, trading strategies built upon $spread_t^{J,2}$ consistently deliver remarkable returns. This analysis reveals compelling arbitrage opportunities in the unregulated crypto market through mean-reverting portfolios, capable of yielding positive returns even in bear markets while outperforming a "buy-and-hold" strategy. It is worth noting that we have not accounted for potential short-selling constraints and transaction costs, which may impact profitability and warrant further investigation.

Furthermore, Granger causality results establish a link between trading strategies and volatility, motivating the extension of volatility modeling for Bitcoin. The RealGARCH-MedRV-std model outperforms other GARCH models, including the HAR-RV model, particularly when employing a jump-robust realized measure and t-distributed innovations. Of utmost importance is the RealGARCH-MedRV-std's superior performance when utilizing realized variance as a proxy.

Remarkably, the Mincer-Zarnowitz regression reveals that models with biased and inefficient forecasts achieve higher R^2. This study underscores the significance of considering jumps when modeling and forecasting Bitcoin's volatility, with GARCH models demonstrating superior accuracy compared to the standard HAR-RV model, especially when incorporating high-frequency data and a jump-robust realized measure with a heavy-tailed distribution.

In summary, our research emphasizes the importance of factoring in jumps when modeling the volatility of Bitcoin, suggesting that GARCH models can offer more accurate

forecasts than the standard HAR-RV model when augmented with high-frequency data. Future research could explore the robustness of results, considering HAR-RV models with a leverage effect and longer forecasting horizons (e.g., 1 week and 1 month). Additionally, incorporating a volatility component for each cryptocurrency when constructing statistical arbitrage strategies based on cointegrated portfolios may enhance out-of-sample performance. Expanding the volatility analysis to other cryptocurrencies and providing one-day ahead forecasts for each cryptocurrency can help traders avoid significant losses when closing long or short positions.

Acknowledgement

The authors wish to thank Jaroslava Hlouskova for helpful comments on an earlier version.

References

Adebola, S. S., Gil-Alana, L. A., and Madigu, G. (2019). Gold prices and the cryptocurrencies: Evidence of convergence and cointegration. *Physica A: Statistical Mechanics and Its Applications*, 523:1227–1236.

Ait-Sahalia, Y. and Jacod, J. (2009). Testing for jumps in a discretely observed process. *The Annals of Statistics*, 37:184–222.

Alexander, C. (1999). Optimal hedging using cointegration. *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 357:2039–2058.

Andersen, T. G. and Bollerslev, T. (1998). Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *International Economic Review*, 39(4):885–905.

Andersen, T. G., Bollerslev, T., and Diebold, F. X. (2007). Roughing it up: Including jump components in the measurement, modeling, and forecasting of return volatility. *The Review of Economics and Statistics*, 89(4):701–720.

Andersen, T. G., Bollerslev, T., and Huang, X. (2008). A reduced form framework for modeling volatility of speculative prices based on realized variation measures. *SSRN Scholarly Paper. Rochester, NY.*

Andersen, T. G., Dobrev, D., and Schaumburg, E. (2012). Jump-robust volatility estimation using nearest neighbor truncation. *Journal of Econometrics*, 169:75–93.

Aras, S. (2021). On improving garch volatility forecasts for bitcoin via a meta-learning approach. *Knowledge-Based Systems*, 230:107393.

Barndorff-Nielsen, O. E. and Shephard, N. (2004). Power and bipower variation with stochastic volatility and jumps. *Journal of Financial Econometrics*, 2(1):1–37.

Barndorff-Nielsen, O. E. and Shephard, N. (2006). Econometrics of testing for jumps in financial economics using bipower variation. *Journal of Financial Econometrics*, 4(1):1–30.

Bergsli, L. O., Lind, A. F., Molnar, P., and Polasik, M. (2022). Forecasting volatility of bitcoin. *Research in International Business and Finance*, 59:101540.

Black, F. (1976). Studies of stock price volatility changes. *Proceedings of the 1976 Meeting of the Business and Economic Statistics Section, American Statistical Association*, pages 177–181.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307–327.

Bouri, E., Gkillas, K., Gupta, R., and Pierdzioch, C. (2021). Forecasting realized volatility of bitcoin: The role of the trade war. *Computational Economics*, 57(1):29–53.

Catania, L. and Grassi, S. (2017). Modelling crypto-currencies financial time-series. *SSRN Scholarly Paper. Rochester, NY.*

Cermak, V. (2017). Can bitcoin become a viable alternative to fiat currencies? an empirical analysis of bitcoin s volatility based on a garch model. *SSRN Scholarly Paper. Rochester, NY.*

Chiu, M. C. and Wong, H. Y. (2015). Dynamic cointegrated pairs trading: Mean variance time-consistent strategies. *Journal of Computational and Applied Mathematics*, 290 (December):516–534.

Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 7(2):174–196.

D Amato, V., Levantesi, S., and Piscopo, G. (2022). Deep learning in predicting cryptocurrency volatility. *Physica A: Statistical Mechanics and Its Applications*, 596:127–158.

Dickey, D. A. and Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74(366):427–431.

Diebold, F. and Mariano, R. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, 50:987–1007.

Engle, R. F. and Granger, C. W. J. (1987). Co-integration and error correction: Representation, estimation, and testing. *Econometrica*, 55(2):251–276.

Engle, R. F., Hendry, D. F., and Richard, J.-F. (1983). Exogeneity. *Econometrica*, 51(2):277–304.

Gatev, E., Goetzmann, W. N., and Rouwenhorst, K. G. (2006). Pairs trading: Performance of a relative value arbitrage rule. *SSRN Scholarly Paper. Rochester, NY.*

Glosten, L. R., Jagannathan, R., and Runkle, D. E. (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. *The Journal of Finance*, 48(5).

Granger, C. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438.

Granger, C. (1981). Some properties of time series data and their use in econometric model specification. *Journal of Econometrics*, 16 (1):121–130.

Group, C. (2022). Cme group announces launch of micro-sized bitcoin and ether options.

Hansen, P. R., Huang, Z., and Shek, H. H. (2012). Realized garch: A joint model for returns and realized measures of volatility. *Journal of Applied Econometrics*, 27(6):877–906.

Harvey, D., Leybourne, S., and Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of Forecasting*, 13(2):281–291.

Hou, A. J., Wang, W., Chen, C. Y.-H., and Härdle, W. K. (2019). Pricing cryptocurrency options: The case of bitcoin and crix. *SSRN Scholarly Paper. Rochester, NY.*

Hung, J.-C., Liu, H.-C., and Yang, J. J. (2020). Improving the realized garch s volatility forecast for bitcoin with jump-robust estimators. *The North American Journal of Economics and Finance*, 52:101–165.

Jiang, G. J. and Oomen, R. C. (2008). Testing for jumps when asset prices are observed with noise - a swap variance approach. *Journal of Econometrics*, 144(2):352–370.

Johansen, S. (1988). Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control*, 12(2):231–254.

Johansen, S. (1991). Estimation and hypothesis testing of cointegration vectors in gaussian vector autoregressive models. *Econometrica*, 59(6):1551–1580.

Johansen, S. (1995). *Likelihood-Based Inference in Cointegrated Vector Autoregressive Models*. Oxford University Press.

Kubát, M. (2015). Virtual currency bitcoin in the scope of money definition and store of value. *Procedia Economics and Finance*, 30:409–416.

Leung, T. and Nguyen, H. (2018). Constructing cointegrated cryptocurrency portfolios for statistical arbitrage. *SSRN Scholarly Paper. Rochester, NY.*

Liu, R., Wu, Z., and Zhang, Q. (2020). Pairs-trading under geometric brownian motions: An optimal strategy with cutting losses. *Automatica*, 115:108912.

MacKinnon, J., Haug, A., and Michelis, L. (1999). Numerical distribution functions of likelihood ratio tests for cointegration. *Journal of Applied Econometrics*, 14(5):563–577.

MacKinnon, J. G. (2010). Critical values for cointegration tests. QED Working Paper. No. 1227.

Mandelbrot, B. (1963). The variation of certain speculative prices. *The Journal of Business*, 36:394–419.

Merton, R. C. (1976). Option pricing when underlying stock returns are discontinuous. *Journal of Financial Economics*, 3(1):125–144.

Mincer, J. and Zarnowitz, V. (1969). The evaluation of economic forecasts. *Economic Forecasts and Expectations: Analysis of Forecasting Behavior and Performance*, pages 3–46.

Mittal, S. (2012). Is bitcoin money? bitcoin and alternate theories of money. SSRN Scholarly Paper. Rochester, NY.

Müller, U. A., Dacorogna, M. M., Dav, R. D., Olsen, R. B., Pictet, O. V., and von Weizsäcker, J. E. (1997). Volatilities of different time resolutions—analyzing the dynamics of market components. *Journal of Empirical Finance*, 4(2-3):213–239.

Nelson, D. B. (1991). Conditional heteroskedasticity in asset returns: A new approach. *Econometrica*, 59(2):347–370.

Ngo, M.-M. and Pham, H. (2016). Optimal switching for the pairs trading rule: A viscosity solutions approach. *Journal of Mathematical Analysis and Applications*, 441(1):403–425.

Ni, S. X., Pan, J., and Poteshman, A. M. (2008). Volatility information trading in the option market. *The Journal of Finance*, 63(3):1059–1091.

Omane-Adjepong, M., Alagidede, P., and Akosah, N. K. (2019). Wavelet time-scale persistence analysis of cryptocurrency market returns and volatility. *Physica A: Statistical Mechanics and Its Applications*, 514:105–120.

Patton, A. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, 160(1):246–256.

Pesaran, M. H., Shin, Y., and Smith, R. J. (2001). Bounds testing approaches to the analysis of level relationships. *Journal of Applied Econometrics*, 16(3):289–326.

Phillips, P. and Ouliaris, S. (1990). Asymptotic properties of residual based tests for cointegration. *Econometrica*, 58(1):165–193.

Pollet, J. M. and Wilson, M. (2010). Average correlation and stock market returns. *Journal of Financial Economics*, 96(3):364–380.

Shen, D., Urquhart, A., and Wang, P. (2020). Forecasting the volatility of bitcoin: The importance of jumps and structural breaks. *European Financial Management*, 26(5):1294–1323.

Song, Q. and Zhang, Q. (2013). An optimal pairs-trading rule. *Automatica*, 49(10):3007–3014.

Sovbetov, Y. (2018). Factors influencing cryptocurrency prices: Evidence from bitcoin, ethereum, dash, litcoin, and monero. SSRN Scholarly Paper. Rochester, NY.

Tan, Z., Huang, Y., and Xiao, B. (2021). Value at risk and returns of cryptocurrencies before and after the crash: Long-run relations and fractional cointegration. *Research in International Business and Finance*, 56:101347.

Theodosiou, M. and Zikes, F. (2011). A comprehensive comparison of alternative tests for jumps in asset prices. Working Papers 2011 2. Central Bank of Cyprus.

Tokat, E. and Hayrullahoğlu, A. C. (2022). Pairs trading: Is it applicable to exchange-traded funds? *Borsa Istanbul Review*, 22(4):743–751.

Urquhart, A. (2018). What causes the attention of bitcoin? *Economics Letters*, 166:40–44.

Yan, T. and Wong, H. Y. (2022). Equilibrium pairs trading under delayed cointegration. *Automatica*, 144:110498.

Yu, M. (2019). Forecasting bitcoin volatility: The role of leverage effect and uncertainty. *Physica A: Statistical Mechanics and Its Applications*, 533:120707.