A Statistical Classification of Cryptocurrencies

Daniel Traian Pele[a], Niels Wesselhöfft[b], Wolfgang K. Härdle[c], Michalis Kolossiatis[d], Yannis Yatracos[e]

[a]*Corresponding author, Department of Statistics and Econometrics, Faculty of Cybernetics, Statistics and Economic Informatics, The Bucharest University of Economic Studies, 010371, Bucharest, Romania, E-mail: danpele@ase.ro*
[b]*Humboldt-Universität zu Berlin, IRTG 1792, Dorotheenstr. 1, 10117 Berlin, Germany*
[c]*Humboldt-Universität zu Berlin, IRTG 1792, Dorotheenstr. 1, 10117 Berlin, Germany*
School of Business, Singapore Management University, 50 Stamford Road, Singapore 178899
Faculty of Mathematics and Physics, Charles University, Ke Karlovu 3, 121 16 Prague, Czech Republic
[d]*Yau Mathematical Sciences Center, Tsinghua University, Beijing, China*
[e]*Department of Mathematics and Statistics, University of Cyprus, Nicosia, 1678, Cyprus*
Central Bank of Cyprus, Nicosia 1076, Cyprus[1]

Abstract

The aim of this paper is to derive the main factors that separates cryptocurrencies from the classical assets, by using various classification techniques applied to the daily time series of log-returns. In this sense, a daily time series of asset returns (either cryptocurrencies or classical assets) can be characterized by a multidimensional vector with statistical components like variance, skewness, kurtosis, tail probability, quantiles, conditional tail expectation or GARCH parameters. By using dimension reduction techniques (Factor Analysis) and classification models (Binary Logistic Regression, Discriminant Analysis, Support Vector Machines, K-means clustering, Variance Components Split methods) for a representative sample of cryptocurrencies, stocks, exchange rates and commodities, we are able to classify cryptocurrencies as a new asset class with unique features in the tails of the log-returns distribution; more, the cryptocurrencies are classified into disjoint clusters, based on their statistical properties. The main result of our paper is the complete separation of the cryptocurrencies from the other type of assets, by using the Maximum Variance Components Split method. In addition, we observe a synchronicity in the evolution of of the cryptocurrencies, compared to the classical assets, mainly due to the tails behaviour of the log-return distribution. The codes are available via www.quantlet.de.

Keywords: cryptocurrency, classification, multivariate analysis, factor models, synchronicity
JEL Classification: C1, G1

1. Introduction

Cryptocurrencies, served as a new digital asset, have attracted much attention from investors and academics. Along with this growing popularity, the market capitalization of cryptocurrencies is increasing substantially. According to a recent report (Trans-

[1]The information and views set out in this article are those of the authors and do not necessarily reflect the official opinion of the Central Bank of Cyprus.

Preprint submitted to Elsevier

parency Market Research, 2018), the total capitalization for cryptocurrencies market was around US$ 574.3 mn in the year 2017 and is expected to reach US$ 6702.1 mn by the end of 2025. Most articles focus on Bitcoin (BTC), as it is considered the first decentralized cryptocurrency, which has the largest capitalization from its beginning till now. An extensive review of the literature regarding the Bitcoin can be found in Corbet et al. (2019). Appendix 5 lists a synthesis of the empirical findings regarding the statistical properties of cryptocurrencies, compared to classical assets.
There is no standard, widely accepted definition of cryptocurrencies allowing to identify them within the existing economic theory (Núñez et al., 2019). In general, cryptocurrencies are defined as "digital representations of value, made possible by advances in cryptography and distributed ledger technology (DLT)" (I.M.F., 2019). Followig the IMF report, cryptocurrencies can be separated into two branches: Bitcoin-like crypto assets (BLCAs) and digital tokens. BLCAs are digital assets based on distributed ledger technology and designed to work as a medium of exchange (for example: Bitcoin, Ether, Ripple (XRP), Bitcoin Cash, EOS, Stellar and Litecoin) (I.M.F., 2019). On the other hand, digital tokens can be split in four categories, depending on their economic function: payment tokens, utility tokens, asset tokens and hybrid tokens. However, there is no classification of cryptocurrencies based on their risk profile, which may be extremely beneficial for investors. In this paper we are providing a classification of assets universe, showing that cryptocurrencies poses unique statistical features, allowing them to differentiate from classical assets. Our approach is different from the existing literature, as most of the reviewed paper are using a low-dimensional approach while trying to differentiate cryptocurrencies from classical assets. Given preceding results from the literature, our contribution to the studies dealing with cryptocurrencies market is mostly empirical, proving the complete separation of cryptocurrencies from the other assets and their synchronic evolution.

Through means of dimensionality reduction techniques (Factor Analysis), we prove that most of the variation among cryptocurrencies, stocks, exchanges rates and commodities can be explained by three factors: the tail factor, the moment factor and the memory factor. Additionally, cryptocurrencies are classified into disjoint clusters, based on their statistical properties. Our results add to the findings from literature by showing that the most important factor which differentiates cryptocurrencies from classical assets is the tail behaviour of the log-returns distribution, as proven by classification techniques: Binary Logistic Regression, Discriminant Analysis, Support Vector Machines, K-means clustering and Variance Components Split. This finding is confirmed by classical Factor Analysis, performed on a static basis and also by using an expanding window approach, where the assets universe can be observed in its evolutionary dynamic. The main result of our paper is the complete separation of cryptocurrencies from classical asset types in finance, by using the Maximum Variance Components Split method and other benchmark methods, (Binary Logistic Regression, Discriminant Analysis, Support Vector Machines, K-means clustering) which provide an almost complete separation. Another important result is the discovery of synchronic evolution of cryptocurrencies, compared to classical assets types. Synchronicity refers to the fact that individual cryptocurrencies tend to develop certain similar characteristics over time that make them fully distinguishable from classical assets, i.e. they tend to behave like a homogeneous group, with certain characteristics that individualize them in the assets ecosystem. By using an expanding window approach, we are able to show that cryptocurrencies have a convergent dynamic in relationship to classical assets and this convergence is driven mainly by the tail behaviour of the log-returns distribution. Moreover, cryptocurrencies as a species exhibit a divergent evolution in relation to classical assets. A related analysis can be found in ElBahrawy et al. (2017), where the cryptocurrency market is seen as an evolutive system with several characteristics which are preserved over time. According to ElBahrawy et al. (2017), the evolution of the cryptocurrency market has been ruled by "neutral" forces, i.e. no cryptocurrency has shown any strong selective advantage over the other.

The paper is subsequently organized as follows: the second section describes the methodology used, including Factor Analysis, Binary Logistic Regression, Support Vector Machines (SVM), K-means clustering, Variance Components Split (VCS) methods and the Evolutive divergence; the third section describes the datasets and interprets the results of the classification; the fourth section describes the synchronic evolution of cryptocurrencies, while the last section concludes. The codes used to obtain the results in this paper are available via www.quantlet.de.

2. Methodology

The methodology used in this paper has four layers: first, we study the statistical properties of the daily log-returns of the selected assets and we estimate the components of a multidimensional vector describing the behaviour of the time series of assets' daily log-returns. Second, we apply data dimension reduction and orthogonalization methods (Factor Analysis) in order to retain the orthogonal factors which maximize the explained variance and could discriminate between cryptocurrencies and classical assets. Third, we employ classification techniques (Binary Logistic Regression, Discriminant analysis, Support Vector Machine, K-means clustering, Variance Components Split methods) to obtain the most influential factors discriminating between cryptocurrencies and classical assets. Fourth, we prove the validity of the synchronic evolution, showing that cryptocurrencies pose specific characteristics, allowing them to differentiate over time from classical assets.

2.1. Taxonomy variables

In order to properly classify the assets universe, we need an initial dataset of variables that have the statistical power to differentiate between cryptocurrencies and classical assets (stocks, exchange rates and commodities). Before introducing the multidimensional dataset used for taxonomy, we set the following notation:

- n: the number of assets in the dataset;
- t: the time index, $t \in \{1, \dots, T\}$, where T is the time of the last record in the dataset;
- $P_{i,t}$: the closing price for asset i in day t, with $i = 1 \dots n$, $t = 1 \dots T$;

- $R_{i,t} = \log P_{i,t} - \log P_{i,t-1}$: the daily log-return for asset i in day t, with $i = 1 \dots n$, $t = 1 \dots T$;

- $R = (R_{i,t})_{i=1,\dots,n}^{t=1,\dots,T} \in M(n,T)$: the initial matrix of the assets' daily log-returns;

- p: the number of variables used for taxonomy.

The multidimensional dataset used for taxonomy is the matrix $X_t = (x_{it,k})_{\substack{i=1\dots n \\ k=1\dots p}} \in M(n,p)$, whose components are detailed below, estimated for the time interval $[1,t]$, with $t = t_0,..,T$, where $t_0 = [T/3]$ (the integer part of $T/3$). First, we took into account the central moments of the log-returns distribution, through the following parameters:

- Variance: $\sigma_{i,t}^2 = \mathsf{E}\left\{(R_i - \mu_{i,t})^2\right\}$;

- Skewness: $\text{Skewness}_{i,t} = \mathsf{E}\left\{(R_i - \mu_{i,t})^3\right\}/\sigma_{it}^3$;

- Kurtosis: $\text{Kurtosis}_{i,t} = \mathsf{E}\left\{(R_i - \mu_{i,t})^4\right\}/\sigma_{it}^4$.

Second, we estimate the following parameters of the α-stable distribution, fitted to daily log-returns, in order to capture tail dependent behavior:

- Tail exponent: *Stable* $\alpha_{i,t} \in (0,2]$, lower values indicating heavier tails;

- Scale parameter: *Stable* $\gamma_{i,t} \geq 0$.

The α-stable distributions are a well-known class of distributions used in financial modelling (Rachev and Mittnik, 2000), capturing the fat tails and the asymmetries of the real-world log-returns distributions. The α-stable parameters were estimated using the empirical characteristic function method, following Koutrouvelis (1980, 1981), through the Matlab library *stbl* (Veillete, 2012). Third, we estimated the quantiles and the conditional tail expectations for the distribution of log-returns, in order to capture the tail behaviour:

- quantiles: $Q_{\alpha;it}$, with $\alpha \in \{0.005, 0.01, 0.025, 0.05, 0.95, 0.975, 0.99, 0.995\}$;

- conditional left tail expectation: $CTE_{\alpha,it}(R_{it}) = \mathsf{E}[R_{it}|R_{it} < Q_{\alpha;it}]$, for $\alpha \in \{0.005, 0.01, 0.025, 0.05\}$;

- conditional right tail expectation: $CTE_{\alpha,it}(R_{it}) = \mathsf{E}[R_{it}|R_{it} > Q_{\alpha;it}]$, for $\alpha \in \{0.95, 0.975, 0.99, 0.995\}$.

From a market risk perspective, the left tail quantiles can be assimilated to Value-at-Risk, the conditional left tail expectation can be regarded as Expected Shortfall, while the conditional right tail expectation can be seen as the Expected Upside.
Fourth, we estimated an GARCH(1,1) model in order to capture the ARCH/GARCH volatility model parameters. Thus, from the following variance equation of the GARCH(1,1) model:

$$\sigma_t^2 = \kappa + \theta_1 \sigma_{t-1}^2 + \omega_1 \varepsilon_{t-1}^2, \tag{1}$$

we estimate the GARCH parameter θ_{1it} and the ARCH parameter ω_{1it}. Our multidimensional dataset can be seen as a tensor $\mathscr{X} \in \mathbb{R}^{n \times p \times T'}$, where n is the number of assets, $p = 23$ is the number of variables and $T' = T - t_0$ is the number of time points.

2.2. Factor Analysis

The most popular methods used to synthesize and extract relevant information from large datasets are Principal Components Analysis (PCA) and Factor Analysis (FA) (Bartholomew, 2011). Factor Analysis has been extensively used in financial modeling for classification purposes: Stevens (1973) who applied this technique on mergers and acquisitions or Yoshino and Taghizadeh-Hesary (2015), who analyzed credit risks for financing small and medium-sized enterprises in Asia. In this paper, we use Factor Analysis to extract the main factors explaining the variation in the initial dataset. PCA itself is a linear combination of variables, while FA is a measurement model of a latent variable. The aim of Factor Analysis is to explain the outcome of the p variables in the data matrix X using fewer variables, the so-called factors (Härdle and Simar, 2012). The orthogonal factor model is given by:

$$X = QF + U + \mu, \tag{2}$$

with the following notations: X is the initial matrix of p variables, F are the common k factors ($k << p$), Q is a matrix of the non-random loadings of the common factors F, U is a matrix of the random specific factors and μ is the vector of the means of initial p variables. It also holds that the random vectors F and U are unobservable and uncorrelated.

In our paper, the initial factor pattern is extracted using the principal component method, followed by a VARIMAX rotation to insure orthogonality of the factors. The Factor Analysis is applied on the entire dataset X_T, the p initial variables being estimated for the entire time period $[1, T]$. The p-dimensional dataset X_T is then projected on the k-dimensional space defined by the k orthogonal factors, in order to observe a separation of the components of the assets.

2.3. Assets Classification

In order to perform the assets classification, we are using several classification techniques: Binary Logistic Regression, Discriminant Analysis, Support Vector Machines, K-means clustering and Variance Components Split. Most of these techniques have been successfully applied in relation to cryptocurrencies and classical assets. Thus, (Fischer et al., 2019) used Binary Logistic Regression, Support Vector Machines and Random Forests classifier to forecast the evolution of cryptocurrencies returns using minute-binned data. Mirtaheri et al. (2009) used Binary Logistic Regression to identify and characterize cryptocurrency frauds that are carried out in social media.

2.3.1. Binary Logistic Regression

The Binary Logistic Regression model quantifies the performance of each of the orthogonal factors extracted through the Factor Analysis to discriminate between cryptocurrencies and classical assets. Thus, we are estimating the following family of models:

$$P(Y_i = 1) = \frac{\exp(\beta_{0j} + \beta_{1j} F_{ji})}{1 + \exp(\beta_{0k} + \beta_{1j} F_{ji})}, \tag{3}$$

where $Y_i = 1$ for cryptocurrencies, $Y_i = 0$ for classical assets, and $F_j, j \in \{1, \dots, k\}$ are the k orthogonal factors retrieved through the Factor Analysis. Based on the explanatory power and the significance of model 3, we can derive the most important factors contributing to the specific difference of cryptocurrencies. As a performance measure for Model 3, we are using $\tilde{R}^2$ (Nagelkerke, 1991), where:

$$\tilde{R}^2 = \frac{1 - \left\{\frac{L(\mathbf{0})}{L(\widehat{\boldsymbol{\beta}})}\right\}^{\frac{2}{n}}}{1 - \{L(\mathbf{0})\}^{\frac{2}{n}}}. \tag{4}$$

In Equation 4, $L(0)$ is the likelihood of the intercept-only model, $L(\widehat{\boldsymbol{\beta}})$ is the likelihood of the full model, and $\widehat{\boldsymbol{\beta}}$ is the vector of Maximum Likelihood estimated parameters.

2.3.2. Discriminant Analysis

The aim of discriminant analysis is to classify one or more observations into *a priori* known groups, minimizing the error of misclassification (Härdle and Simar, 2012). Formally, Linear Discriminant Analysis (LDA) assumes that the input dataset is multivariate Normal: $X_i \sim \mathrm{N}(\mu_i, \Sigma)$, where X_i belong to class ω_i. The goal is to project samples X onto a line $Z = w^\top X$, where we select the projection that maximizes the standardized separability of the means over all directions. Specifically, we maximize the normalized, squared distance in the means of the classes:

$$w^* = \arg\max_w \frac{|w^\top(\mu_i - \mu_j)|^2}{s_i^2 + s_j^2}, \tag{5}$$

$$s_i^2 = \sum_{x_i \in \omega_i} (w^\top x_i - w^\top \mu_i)^2 = w^\top S_i w, \tag{6}$$

giving the Linear Discriminant of Fisher (1936):

$$w^* = S_W^{-1}(\mu_i - \mu_j), \;\; S_W = S_i + S_j. \tag{7}$$

Quadratic Discriminant Analysis (QDA) follows the same procedure, but for $X_i \sim \mathrm{N}(\mu_i, \Sigma_i)$ belong to the class ω_i, one can relax the condition of equality of covariance matrices by $\Sigma_i \neq \Sigma_j$, $i \neq j$, allowing for a non-linear classifier.

2.3.3. Support Vector Machines

Support Vector Machines (SVM) is a data classification technique, its goal being to produce a model which predicts target values based on a set of attributes (Cristianini and Shawe-Taylor, 2000). The goal is to find a projection that maximizes margin in a hyperplane of the original data, without any parametric assumptions on the underlying stochastic process. The support vectors are determined via a quadratic optimization problem i.e. given a training data set D with n samples and 2 dimensions $D = (X_1, Y_1), \dots (X_n, Y_n)$, $X_i \in \mathbb{R}^2$, $\quad Y_i \in [0, 1]$, the aim is to find a hyperplane that maximizes the margin:

$$\min_{w,b} \frac{1}{2}\|w\|^2, \text{ s.t. } \quad Y_i \left(w^\top X_i + b \right) \geq 1, i = 1, \dots, n. \tag{8}$$

2.3.4. K-means Clustering Algorithm

This clustering method was first popularized by (MacQueen, 1967), who acknowledge a couple of other researchers that independently used that method around the same time. The aim is to allocate each observation of a data set in one of $k \in \mathbb{N}$ clusters, where k is predefined, so as to minimize the within-cluster sums of squares. In brief, the algorithm proceeds as follows:

i. Take k data points and set them as the cluster centres.
ii. Iteratively, for each data point, assign it to the cluster which centre is closer to the data point (the Euclidean distance is usually used, but other distance metrics have been proposed). Update the cluster centre for the selected cluster.
iii. Repeat until convergence (*i.e.* the allocations do not change).

2.3.5. Variance Components Split methods: MVCS, GMVCS

These methods aim to separate, respectively, the components of a structure like the types of assets herein or the types of Iris flowers, and clusters defined as the components of a mixture distribution. They are based on an unusual variance decomposition in between-group variations (Yatracos, 1998, 2013). To describe the sample version of the decomposition, let $X_1, \dots, X_n$ be i.i.d. random variables. $X_{(j)}$ is the j-th order statistic, $1 \leq j \leq n$.

Consider the groups $X_{(1)}, \dots, X_{(i)}$ and $X_{(i+1)}, \dots, X_{(n)}$ with averages, respectively, $\bar{X}_{[1,i]}$ and $\bar{X}_{[i+1,n]}$, $i = 1, ..., n-1$, then

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n-1} \frac{i(n-i)}{n^2}(\bar{X}_{[i+1,n]} - \bar{X}_{[1,i]})(X_{(i+1)} - X_{(i)}). \tag{9}$$

The summands on the right side of Equation 9 measure between-groups variations. The standardized sample variance components

$$W_i = W_i(X_1, \ldots, X_n) \tag{10}$$
$$= \frac{i(n-i)}{n} \frac{(\bar{X}_{[i+1,n]} - \bar{X}_{[1,i]})(X_{(i+1)} - X_{(i)})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}, \ i = 1, \ldots, n-1, \tag{11}$$

indicate the relative contribution of the groups $X_{(1)}, ..., X_{(i)}$ and $X_{(i+1)}, ..., X_{(n)}$ in the sample variability. The index

$$\mathcal{I}_n = \max\{W_i, i = 1, \ldots, n-1\} \tag{12}$$

determines two potential clusters or parts of a structure and is based on averages and inter-point distances. When $\mathcal{I}_n = W_j$, these clusters are $\tilde{\mathcal{C}}_1 = \{X_{(1)}, \ldots, X_{(j)}\}, \tilde{\mathcal{C}}_2 = \{X_{(j+1)}, \ldots, X_{(n)}\}$. The observed $\mathcal{I}_n$-value is significant at α-level for the normal model when it exceeds the critical value $[-\ln(-\ln(1-\alpha)) + \ln n]/n$ (Yatracos, 2009); $\alpha = 0.05$ is used herein.

When $\mathcal{X}$ is the n by r data matrix of r-dimensional observations, $\mathbf{X}_j$ is the j-th row of $\mathcal{X}, j = 1, \ldots, n$. The coefficients of the orthogonal projection of $\mathcal{X}$ along the unit norm r-row vector $\mathbf{a}$ are $\mathcal{X}\mathbf{a} = (\mathbf{X}_1\mathbf{a}, \ldots, \mathbf{X}_n\mathbf{a})$.

The split in the sorted values of $\mathcal{X}\mathbf{a}$, where

$$\mathcal{I}_{\mathcal{X}}(\mathbf{a}) = \max\{W_i(\mathbf{X}_1\mathbf{a}, \ldots, \mathbf{X}_n\mathbf{a}); i = 1, ..., n-1\} \tag{13}$$

is attained, determines *along* $\mathbf{a}$ the groups $\tilde{\mathcal{C}}_{\mathcal{X},1}(\mathbf{a})$ and $\tilde{\mathcal{C}}_{\mathcal{X},2}(\mathbf{a})$ in the $\mathcal{X}$-rows which are potential clusters and parts of a structure. For example, if for the data herein $\tilde{\mathcal{C}}_{\mathcal{X},1}(\mathbf{a})$ consists of rows 1-14, cryptocurrencies (a component) among the assets (the structure) are completely separated along $\mathbf{a}$.

The Maximum Variance Component Split (MVCS) method compares known components of a structure, *e.g.* cryptocurrencies herein, with data splits for a set of unit projection directions $\mathcal{D}_M$ usually determined by M positive equidistant angles of $[0, \pi]$; *e.g.* when $r = 2$ and $M = 3$ the angles used are $\pi/3, 2\pi/3, \pi$. When one of the data split along projection direction $\mathbf{a}$ coincides with a component of the structure we have complete separation of this component along $\mathbf{a}$.

A set of projection directions $\mathcal{D}_M$ can be

$$(\Pi_{l=1}^r cos\theta_l,\ sin\theta_1\Pi_{l=2}^r cos\theta_l, ...,\ sin\theta_{r-1}cos\theta_r,\ sin\theta_r), \tag{14}$$

where θ_l takes values in $\{\frac{m\pi}{M}, m = 1, ..., M\}, l = 1, ..., r$.

The method is computationally intensive for large r and M values, thus it may be used on subsets of the $\mathcal{X}$-columns. The importance of a subset S of $\mathcal{X}$-columns in the separation of a structure's component is measured by the number N_S of projection directions 14 completely separating the component. Indications for the importance of a specific column c in S in the separation of the same component are obtained by comparing N_S with the number of projection directions N_{S-c} separating the component when c is left out and also by comparing all $N_{S-c}, c \in S$. Similar indications of importance can be used for subgroups of S-columns.

The Global Maximum Variance Component Split (GMVCS) along all projection vectors $\mathcal{D}$, to be obtained from $\max\{\mathcal{I}_{\mathcal{X}}(\mathbf{a}), \mathbf{a} \in \mathcal{D}\}$ that is called the index determines two clusters. In practice, its approximation is obtained using $\mathcal{D}_M$. The splitting of these clusters may continue (Yatracos, 2013).

2.3.6. Expanding window modelling

For observing the assets dynamic, we are using an expanding window approach, allowing to distinguish the evolution of the clusters. In fact, for $t \in \{t_0, .., T\}$, where $t_0 = [T/3]$, the p-dimensional dataset X_t is projected on the k-dimensional space defined by the main factors extracted through the Factor Analysis applied on the dataset X_T. By using this projection instead of a time-varying factor model, we are avoiding situations like changes in factors loadings, causing inconsistencies over time.

3. Data and Results

Our dataset is a combination of cryptocurrencies and classical assets (commodities, exchange rates and stocks), covering the time period 01/02/2014 - 08/30/2019 (1426 trading days), for $n = 679$ assets (see Table 1). The reason for choosing this time span for the analysis is that before 2015 the liquidity in the cryptocurrency market had been relatively low, their total market capitalization being less than US$16 billion (Feng et al., 2018).

Table 1: Dataset

Type of Asset	**Number of Assets**	**Source**
Cryptocurrencies	150	Coinmarketcap
Stocks	496	Bloomberg
Exchange rates	13	Bloomberg
Commodities	20	Bloomberg

For robustness purposes, only the assets with at least 500 observations were kept in the analysis.[2] The first component of the dataset contains a representative sample of 150 cryptocurrencies selected from the top 500 cryptocurrencies sourced from `https://coinmarketcap.com/`, accounting for 98% of total market capitalization. The second component contains a sample of the most traded commodities indexes, the third component contains a sample of the most liquid exchange rates, while the fourth component contains the constituents of the S&P500 Index, recorded at August 30th 2019. As cryptocurrencies daily data are available at all times, while the stocks data obtained from Bloomberg observe market closure days (weekends and public holidays), the cryptocurrency data were pre-processed and the returns were computed in the same way as for classical assets (for example, the Monday return compares the Monday closing price to the Friday closing price).

3.1. Factor Analysis

Factor Analysis is a classical method used to find latent variables or factors among observed variables, by grouping variables with similar characteristics. Three steps are involved:

i. Estimation of the correlation matrix for all the variables, shown in Figure 1.

[2]The complete list of the assets included in the analysis can be found here.

ii. Extraction of the factors from the correlation matrix, based on the correlation coefficients of the variables.

iii. Factor rotation, in order to maximize the relationship between the variables and relevant factors.

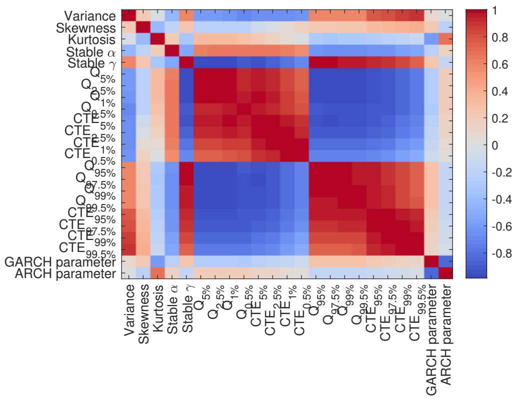

Figure 1: Correlation matrix

SFA_Cryptos

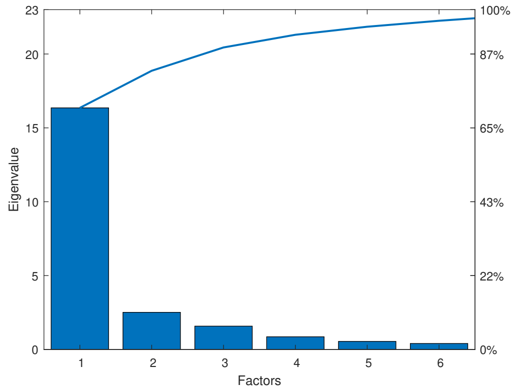

Figure 2: Scree plot

SFA_Cryptos

Based on the eigenvalues criteria, we can select those factors for which the eigenvalue is higher than 1 (see Figure 2). Accordingly, three factors were selected, accounting for 89% of the total variance. In order to test the sampling adequacy of the Factor Analysis, we are using the Kaiser-Meyer-Olkin (KMO) test, which should be greater than 0.5 for a satisfactory Factor Analysis (Tabachnick and Fidell, 2013). The overall KMO test is computed as:

$$KMO = \frac{\sum_i \sum_{i \neq j} r_{ij}^2}{\sum_i \sum_{i \neq j} r_{ij}^2 + \sum_i \sum_{i \neq j} u_{ij}^2}. \tag{15}$$

where $R = (r_{ij})_{\substack{i=1\dots n \\ j=1\dots n}}$ is the correlation matrix and $U = (u_{ij})_{\substack{i=1\dots n \\ j=1\dots n}}$ is the partial covariance matrix (Cerny and Kaiser, 1977, Kaiser, 1974).

In fact, the KMO measure represents the proportion of the variance in the input variables that might be caused by underlying factors (Kaiser, 1981). In our sample, the overall KMO value is 0.92, pointing out that the Factor Analysis is suitable for structure detection. For the factor rotation, we used the VARIMAX method, which outputs orthogonal factors, also minimizing the number of variables that have high loadings on each factor. Based on the rotated factors pattern, the following conclusions can be drawn (see Figure 4):

i. First factor: **the tail factor**, accounting for 71% of the total variance, is highly correlated with the following parameters: the tail parameter alpha and the scale parameter gamma of the stable distribution, the lower and upper quantiles of the distribution of log-returns, the conditional tail expectations and the variance of log-returns.

ii. Second factor: **the moment factor**, accounting for 11% of the total variance, is highly correlated with the variance and the skewness coefficient of the distribution of log-returns.

iii. Third factor: **the memory factor**, accounting for 7% of the total variance, is highly correlated with the GARCH and ARCH parameters of the GARCH(1,1) model estimated for log-returns and also the kurtosis of the distribution of log-returns.

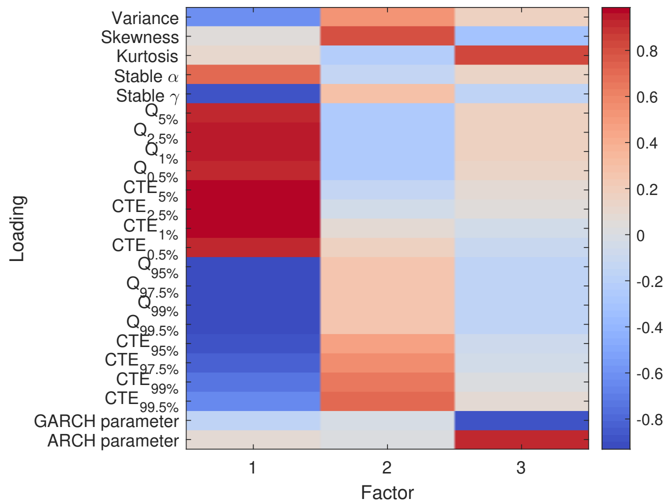

Figure 3: Loadings of the three factors

SFA_Cryptos

Based on the data revealed in Table 2, one can synthesize few characteristics of cryptocurrencies that differentiate them from the other assets. First, cryptocurrencies have higher variance of the log-return's distribution, compared to classical assets. Second, cryptocurrencies have longer tails, as indicated by the high values of quantiles and conditional tail expectations, i.e. cryptocurrencies have higher propensity for extreme values, in both tails of the log-returns distribution.

Table 2: Assets profile based on the average values of the initial variables

Variable	Commodities	Cryptocurrencies	Exchange rates	Stocks
$\sigma^2 \cdot 10^3$	3.603	43.274	0.027	1.260
$Skewness$	0.214	3.876	−1.231	−7.797
$Stable_\alpha$	1.713	1.398	1.703	1.692
$Stable_\gamma \cdot 10^3$	9.266	47.080	2.868	8.738
$Q_{0.5\%}$	−0.026	−0.159	−0.008	−0.025
$Q_{1\%}$	−0.034	−0.211	−0.010	−0.033
$Q_{2.5\%}$	−0.043	−0.300	−0.012	−0.045
$Q_{5\%}$	−0.054	−0.388	−0.014	−0.056
$CTE_{0.5\%}$	−0.042	−0.274	−0.011	−0.047
$CTE_{1\%}$	−0.056	−0.367	−0.013	−0.065
$CTE_{2.5\%}$	−0.082	−0.546	−0.017	−0.108
$CTE_{5\%}$	−0.122	−0.744	−0.020	−0.167
$CTE_{95\%}$	0.044	0.368	0.011	0.038
$CTE_{97.5\%}$	0.058	0.533	0.013	0.049
$CTE_{99\%}$	0.087	0.877	0.015	0.072
$CTE_{99.5\%}$	0.128	1.299	0.018	0.099
$Q_{95\%}$	0.026	0.171	0.007	0.024
$Q_{97.5\%}$	0.034	0.246	0.010	0.030
$Q_{99\%}$	0.046	0.377	0.012	0.040
$Q_{99.5\%}$	0.057	0.518	0.014	0.050
$ARCH$	0.111	0.494	0.079	0.698
$GARCH$	0.665	0.478	0.720	0.206
$Kurtosis$	58.608	218.732	38.167	561.702

Based on the factors estimated through the Factor Analysis, one can map cryptocurrencies and classical assets, in order to derive some clustering effect.

Figures 4 and 5 map cryptocurrencies and classical assets; the colour code is the following: green: cryptocurrencies, black: stocks, red: commodities, blue: exchange

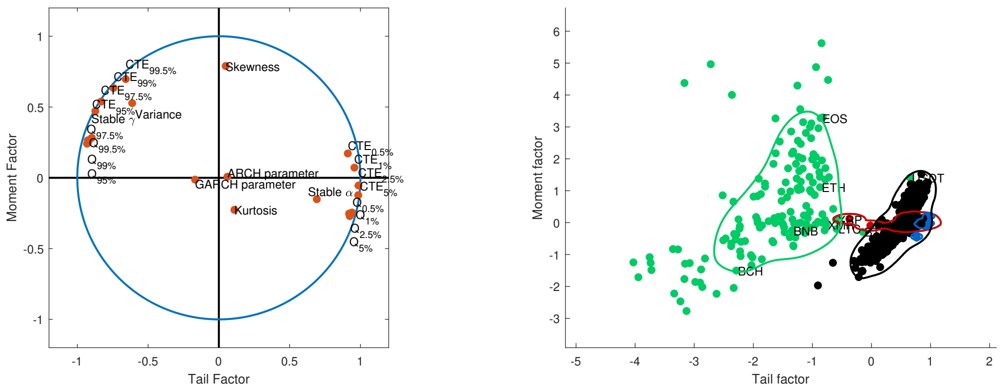

Figure 4: Loadings (left) and scores (right) based on tail and moment factor

SFA_cryptos

rates. Also, a 95% confidence region is estimated, based on the Bivariate Kernel Density; in all figures only the top 10 cryptocurrencies (according to their market capitalization) are labeled.

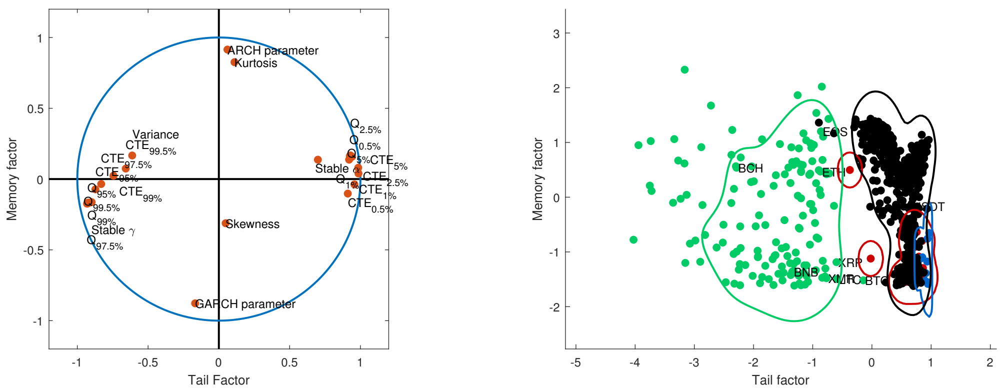

Figure 5: Loadings (left) and scores (right) based on tail and memory factor

SFA_Cryptos

As shown in Figures 4 and 5, there is a clear separation between cryptocurrencies and classical assets, mainly due to the first factor, the tail factor, while the memory and moment factor are of subliminal importance. The projection on the 3D space defined by the Factor Analysis reveals two cryptocurrencies with atypical behaviour: Bitcoin and Tether. Thus, Bitcoin (BTC), the oldest and the most traded cryptocurrency, is closer to classical stocks and commodities, in terms of the tail factor, i.e. Bitcoin can be considered at the border between the classical assets and cryptocurrencies. On the other hand, Tether (USDT), a token that attempts to be tied to the US dollar, it's indistinguishable from the classical assets.

3.2. Assets classification

In this section, we list the results of the models presented in Section 2.3, in order to assess the ability of the factors produced through the Factor Analysis to discriminate between cryptocurrencies and classical assets. First, for each of the three factors we estimated the Binary Logistic model

$$P(Y_i = 1) = \frac{\exp(\beta_{0j} + \beta_{1j} F_{ji})}{1 + \exp(\beta_{0k} + \beta_{1j} F_{ji})}, \tag{16}$$

where $Y_i = 1$ for cryptocurrencies, $Y_i = 0$ for classical assets, and $F_j, j \in \{1, 2, 3\}$ are the orthogonal factors retrieved through the Factor Analysis. Table 3 lists the estimated β_{1j} of the Binary Logistic Regression model 16, with the performance measure defined by Equation 4.

Table 3: Estimates of Binary logistic regression model

Exogenous factor	**Factor 1**	**Factor 2**	**Factor 3**
Estimated β_1	-7.879**	0.728**	-0.389**
	(1.077)	(0.102)	(0.093)
$\tilde{R}^2$	0.967	0.134	0.034

Note: Standard errors in parentheses; ** denotes significance at 95% confidence level.

As seen in Table3, the most important factor regarding the separation between cryptocurrencies and classical assets is the tail factor, while the other two factors have little influence. Second, we employed Discriminant Analysis and Support Vector Machines on the space defined by the two first factors (tail and moment). Figure 6 illustrates the classification results using Discriminant Analysis. Quadratic classifiers have a good classification power, the only cryptocurrencies which are misclassified being Bitcoin and Tether.

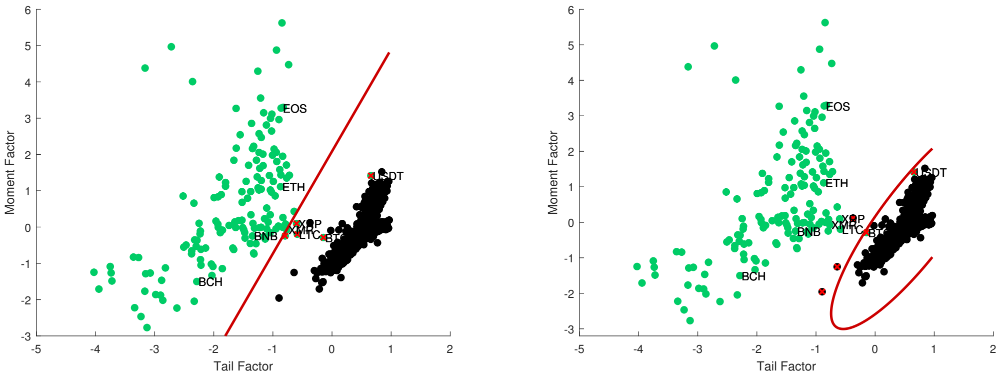

Figure 6: Discriminant Analysis: linear (left) and quadratic (right). Green dots denote cryptocurrencies, while the black dots denote the other assets; the dots highlighted in red are cases of misclassification

SFA_Cryptos

The same conclusion can be drawn by looking at the results of the Support Vector Machines non-linear classifier, according to which all cryptocurrencies are correctly classified using the tail factor and the moment factor (see Figure 7).

The k-means clustering algorithm (MacQueen, 1967) was also used for $k = 2, 3, 4$, the results showing that this method does not provide perfect classification[3]. The optimal number of clusters, as determined by the Elbow method, is $k = 10$; however, five clusters contain only cryptocurrencies, while the the other five clusters contain stocks, commodities and exchange rates, plus cryptocurrencies Bitcoin and Tether. In order to control for the influence of classical assets, these five clusters containing stocks, commodities, exchange rates, Bitcoin and Tether were merged into a single cluster. The final six clusters are shown in Figure 8, projected onto three dimensional space defined by the three factors extracted trough the Factor Analysis; the black dots denote classical assets

[3]Perfect classification is the case when a specific component is completely separated by the rest. In other words, all the members of that component, and only those, are in one cluster.

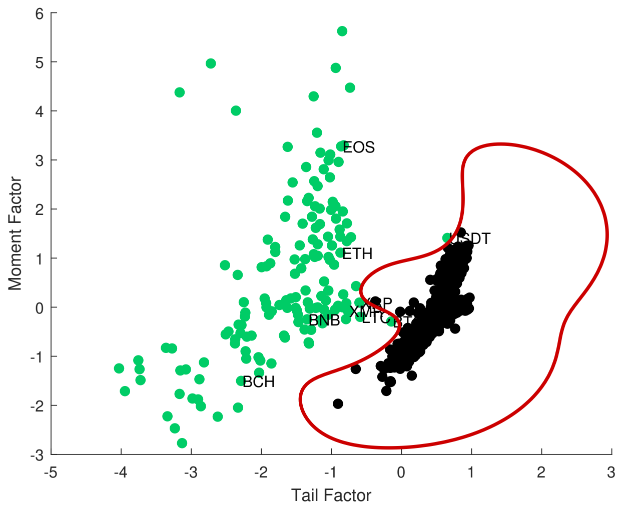

Figure 7: Support Vector Machines classification with green dots denoting cryptocurrencies, while black dots denote the other asset classes

SFA_Cryptos

(including Bitcoin and Tether), while the colored dotes denote cryptocurrencies. Each cryptocurrencies cluster was labeled with its leader in terms of market capitalization.

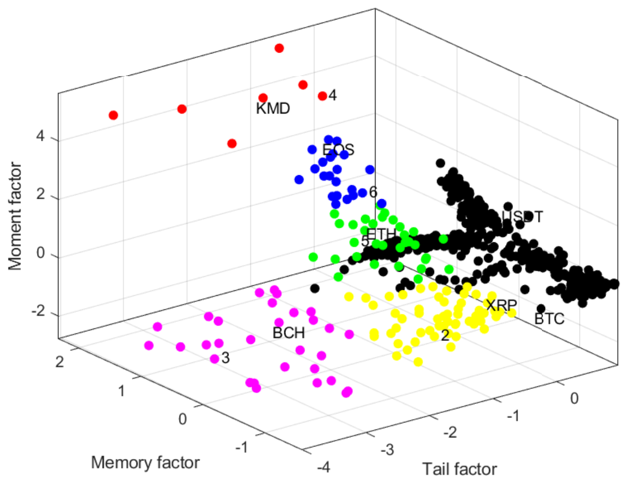

Figure 8: Projection of the clusters on the 3D space extracted trough Factor Analysis

Cluster_Cryptos

Table 4: Crytptocurrencies clusters profile

Variable	BTC&USDT	XRP	BCH	KMD	ETH	EOS
σ^2	0.01	0.02	0.05	0.23	0.03	0.07
$Skewness$	16.13	1.61	−13.34	21.37	10.01	17.60
$Stable_\alpha$	1.05	1.39	1.36	1.41	1.46	1.45
$Stable_\gamma$	0.01	0.05	0.05	0.05	0.05	0.05
$Q_{0.5\%}$	−0.04	−0.16	−0.18	−0.16	−0.16	−0.16
$Q_{1\%}$	−0.06	−0.21	−0.24	−0.20	−0.20	−0.20
$Q_{2.5\%}$	−0.08	−0.30	−0.36	−0.33	−0.28	−0.27
$Q_{5\%}$	−0.10	−0.39	−0.46	−0.56	−0.34	−0.33
$CTE_{0.5\%}$	−0.08	−0.26	−0.39	−0.31	−0.23	−0.23
$CTE_{1\%}$	−0.09	−0.33	−0.57	−0.44	−0.29	−0.29
$CTE_{2.5\%}$	−0.13	−0.46	−1.00	−0.70	−0.38	−0.39
$CTE_{5\%}$	−0.18	−0.58	−1.55	−1.02	−0.45	−0.46
$CTE_{95\%}$	0.13	0.32	0.36	0.63	0.37	0.45
$CTE_{97.5\%}$	0.17	0.43	0.48	1.06	0.54	0.72
$CTE_{99\%}$	0.32	0.64	0.67	2.18	0.92	1.36
$CTE_{99.5\%}$	0.55	0.84	0.83	4.04	1.38	2.22
$Q_{95\%}$	0.04	0.17	0.20	0.16	0.16	0.16
$Q_{97.5\%}$	0.06	0.25	0.29	0.23	0.24	0.23
$Q_{99\%}$	0.08	0.38	0.43	0.36	0.36	0.37
$Q_{99.5\%}$	0.10	0.52	0.59	0.53	0.51	0.47
$ARCH$	0.23	0.17	0.61	0.70	0.66	0.89
$GARCH$	0.76	0.79	0.39	0.30	0.29	0.11
Kurtosis	551.59	43.33	348.62	514.30	210.30	384.02
Nr. Assets.	2	57	30	7	32	22

Note: column names denote the clusters leaders, according to their market capitalization.

Thus, cryptocurrencies clusters are the following:

- Bitcoin and Tether cluster, having a similar behavior with classical assets,
- Ripple (XRP) cluster,
- Bitcoin Cash (BCH) cluster,
- Komodo (KMD) cluster,
- Ethereum (ETH) cluster,
- EOS cluster.

As shown in Table 4, there is a large variability in clusters profiles: the KMD cluster is the most volatile, with the highest variance and longest tails. The XRP cluster and BTC&USDT have the highest memory parameter, indicating a stronger persistence of volatility shocks. The validity of this classification is proven by the Multiclass Error-Correcting Output Codes (ECOC) model using Support Vector Machines (SVM) binary learners (Allwein et al., 2000), obtaining a classification accuracy of 99.41%, based on the three orthogonal factors extracted trough Factor Analysis (see Figure 9).

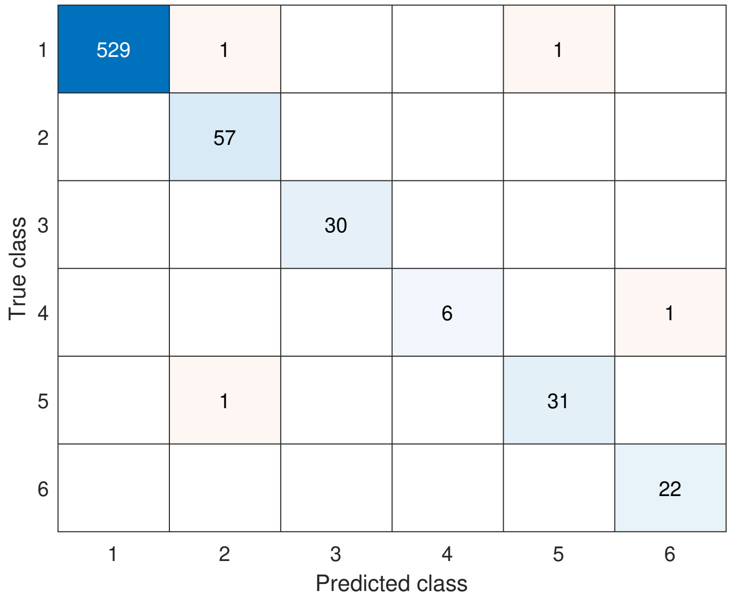

Figure 9: Confusion matrix of the ECOC classifier: 1 - classical assets, BTC and USDT, 2 - XRP cluster, 3 - BCH cluster, 4 - KMD cluster, 5 - ETH cluster, 6 - EOS cluster

Cluster_Cryptos

The results when applying the MVCS method, where the goal is to achieve separation of the components of the assets, are in accordance with those of Binary Logistic Regression, Discriminant Analysis and Support Vector Machines. In order to apply MVCS method, we are considering the following term structure: the assets-data are regarded as a matrix $X_T = (x_{iT,k})_{\substack{i=1...n \\ k=1...p}} \in M(n,p)$, with $p = 23$ columns, representing the variables used for taxonomy, and $n = 679$, representing the assets. To be concise, the 23 columns are considered to be ordered, using the following order: $Variance$,

$Skewness$, $Kurtosis$, $Stable_\alpha$, $Stable_\gamma$, $Q_{0.5\%}$, $Q_{1\%}$, $Q_{2.5\%}$, $Q_{5\%}$, $Q_{95.5\%}$, $Q_{97.5\%}$, $Q_{99\%}$, $Q_{99.5\%}$, $CTE_{0.5\%}$, $CTE_{1\%}$, $CTE_{2.5\%}$, $CTE_{5\%}$, $CTE_{95\%}$, $CTE_{97.5\%}$, $CTE_{99\%}$, $CTE_{99.5\%}$, $GARCH$, $ARCH$. The following notations are used for the MVCS method: M are the positive equidistant angles of $[0;\ \pi]$, S is a specific subset of the columns, N_S is the number of projection directions giving perfect classification when S is used, P_S is the corresponding percentage of these directions, while $minI, maxI$ are the minimum and the maximum index I value for perfect classification, respectively. The critical value for significance of the index for $\alpha = 5\%$ and $n = 679$ is 0.014.

In the following, we present the results of the MVCS method for perfect classification of cryptocurrencies from the other assets, as it was found that for all three other structures (stocks, exchange rates and commodities), none of the combinations of M and S presented below provided perfect classification.

First, due to processing power constraints, we split the data in two parts: the first part consists of columns 1-12 and the second part includes columns 13-23. For the same reason, projection directions (14) are used only for $M = 3, 6$. The number of projection directions used is M^{d-1}, with d, respectively, 12 and 11. Using all the data, perfect classification was not obtained for either data set. By omitting Tether (USDT), however, perfect classification was obtained for the fist set. Table 5 shows the results of the MVCS method, for all data, except USDT.

Table 5: Results of the MVCS method

M	S	N_S	P_S	$minI$	$maxI$
3	1-12	33	0.019%	0.041	0.135
6	1-12	27701	0.007%	0.045	0.146
3	13-23	0	0	n/a	n/a
6	13-23	0	0	n/a	n/a

The above results indicate that columns 1-12 are more important than columns 13-23 (the largest the value of P_S for a specific value of M, the more projection directions give perfect classification, and therefore the columns used are more suitable for separating the cryptocurrencies from the other assets). Table 6 shows the results of the MVCS method, when splitting these columns into disjoint subsets 1-6 and 7-12 (with USDT included).

Table 6: Results of the MVCS method, columns 1-6 and 7-12

M	S	N_S	P_S	$minI$	$maxI$
3	1-6	0	0	n/a	n/a
6	1-6	0	0	n/a	n/a
9	1-6	0	0	n/a	n/a
12	1-6	0	0	n/a	n/a
15	1-6	2	0.0003%	0.038	0.048
18	1-6	0	0	n/a	n/a
3	7-12	0	0	n/a	n/a
6	7-12	0	0	n/a	n/a
9	7-12	0	0	n/a	n/a
12	7-12	0	0	n/a	n/a
15	7-12	0	0	n/a	n/a
18	7-12	0	0	n/a	n/a

From the above, we can conclude that the most important columns for complete separation are columns 1-12, and in particular columns 1-6 (again, from P_S, for the same value of M). The projected values for all the assets, using columns 1-6 and $M = 15$, on the projection direction that provided the largest index value among those that gave perfect classification of thre cryptocurrencies, are shown in Figure 10.

The projection direction that gave the largest index value for columns 1-6 and $M = 15$ is: $(0.005, -0.001, 0, 0.052, 0.497, 0.866)$. The index value in this direction is 0.048. In total 2 (out of 759375 tried) projection directions gave perfect classification and all provided statistically significant index values for the normal model.

Next, the first six columns are further used, as they are deemed the most important, according to the above. Then, the MVCS method is applied to all six quintets (these quintets are derived by omitting one of the six columns). Again, higher values of M are

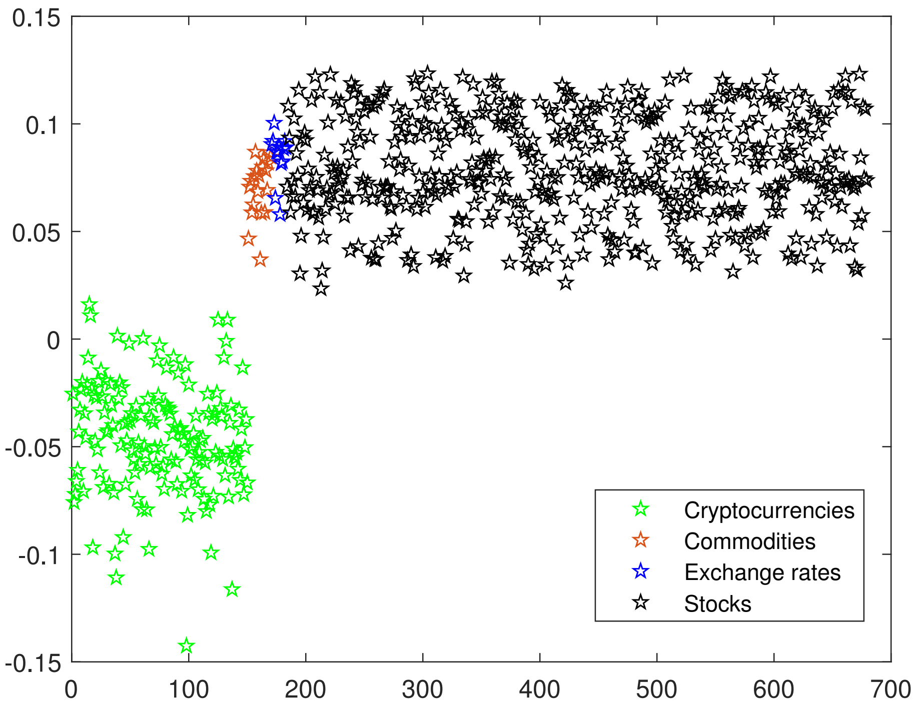

Figure 10: Projections of a subset of the data (the first 6 columns) for $M = 15$ on the projection direction that gave the largest index value among those that gave perfect classification of the cryptocurrencies.

VCS_Cryptos

used, and the results being reported in Table 7.

From the above, one can see that the least important column is the third (Kurtosis), since its omission still provided perfect classification for the cryptocurrencies. As mentioned before, perfect classification is obtained only for cryptocurrencies, while the other assets have indistinct behaviour.This result is in line with the conclusions obtained trough the other classification techniques used above (Binary Logistic Regression, Discriminant Analysis and Support Vector Machines), MVCS method showing that cryptocurrencies behaves like a totally different species in the assets universe.

Finally, it is worth noting that if USDT is excluded, by following the above procedure, we find that columns 7-12 are more important than columns 1-6. Among those columns

Table 7: Results for cryptocurrencies, all leave-one-out quintets of columns 1-6

M	S	N_S	P_S	$minI$	$maxI$
18	1,2,3,4,5	0	0	n/a	n/a
18	1,2,3,4,6	0	0	n/a	n/a
18	1,2,3,5,6	0	0	n/a	n/a
18	1,2,4,5,6	0	0	n/a	n/a
18	1,3,4,5,6	0	0	n/a	n/a
18	2,3,4,5,6	0	0	n/a	n/a
24	1,2,3,4,5	0	0	n/a	n/a
24	1,2,3,4,6	0	0	n/a	n/a
24	1,2,3,5,6	0	0	n/a	n/a
24	1,2,4,5,6	0	0	n/a	n/a
24	1,3,4,5,6	0	0	n/a	n/a
24	2,3,4,5,6	0	0	n/a	n/a
32	1,2,3,4,5	0	0	n/a	n/a
32	1,2,3,4,6	0	0	n/a	n/a
32	1,2,3,5,6	0	0	n/a	n/a
32	1,2,4,5,6	61	0.006%	0.034	0.077
32	1,3,4,5,6	0	0	n/a	n/a
32	2,3,4,5,6	0	0	n/a	n/a

(Table 8), the least important seems to be the ninth ($Q_{0.005}$) and the most important seem to be the eighth ($Q_{0.01}$) and the eleventh ($CTE_{0.025}$). We also get much more projection directions that provided perfect classification of the cryptocurrencies. This is in accordance to the observation that USDT is closely linked to the US dollar.

We can conclude that cryptocurrencies are financial instruments whose specific difference is the tail behaviour of the distribution of daily log-returns. In other words, based on the tail factor profile, we can conclude that a random asset is likely to be a cryptocurrency if it has the following properties: very long tails of the log-returns distribution (in terms of the left and right quantile and the conditional tail expectation), high variance, high value of the α-stable scale parameter and value of the α-stable tail

Table 8: Results for cryptocurrencies, all leave-one-out quintets of columns 7-12, USDT excluded

M	S	N_S	P_S	$minI$	$maxI$
18	7,8,9,10,11	171	0.16%	0.050	0.134
18	7,8,9,10,12	55	0.05%	0.061	0.109
18	7,8,9,11,12	106	0.10%	0.061	0.108
18	7,8,10,11,12	378	0.36%	0.041	0.134
18	7,9,10,11,12	33	0.03%	0.047	0.096
18	8,9,10,11,12	130	0.12%	0.048	0.119
24	7,8,9,10,11	577	0.17%	0.038	0.132
24	7,8,9,10,12	303	0.09%	0.050	0.126
24	7,8,9,11,12	403	0.12%	0.041	0.114
24	7,8,10,11,12	824	0.25%	0.041	0.132
24	7,9,10,11,12	122	0.04%	0.048	0.112
24	8,9,10,11,12	450	0.14%	0.047	0.117
32	7,8,9,10,11	1725	0.16%	0.041	0.133
32	7,8,9,10,12	886	0.08%	0.042	0.122
32	7,8,9,11,12	1198	0.11%	0.0041	0.118
32	7,8,10,11,12	2840	0.27%	0.0040	0.133
32	7,9,10,11,12	428	0.04%	0.0047	0.108
32	8,9,10,11,12	1623	0.15%	0.046	0.121

index close to 1.

4. Synchronic evolution of cryptocurrencies

For observing the assets dynamic, we are using an expanding window approach, allowing to distinguish the evolution of the clusters. In fact, for $t = t_0, \ldots, T$, the p-dimensional dataset is projected on the k-dimensional space defined by the main factors extracted through the Factor Analysis applied on the dataset X_T. By using this projection instead of a time-varying factor model, we are avoiding situations like changes in factors loadings, causing inconsistencies over time.

In order to derive the dynamics of the assets' universe, we used an expanding window approach, described below:

- The 23-dimensional dataset is estimated for the time interval $[1, t_0] = [01/02/2014, 10/31/2016]$.
- Time window is extended on a daily basis, up to T=08/30/2019 and for each step in time, the dataset is projected on the 2-dimensional space defined by the tail factor and the moment factor, estimated for the entire time period.

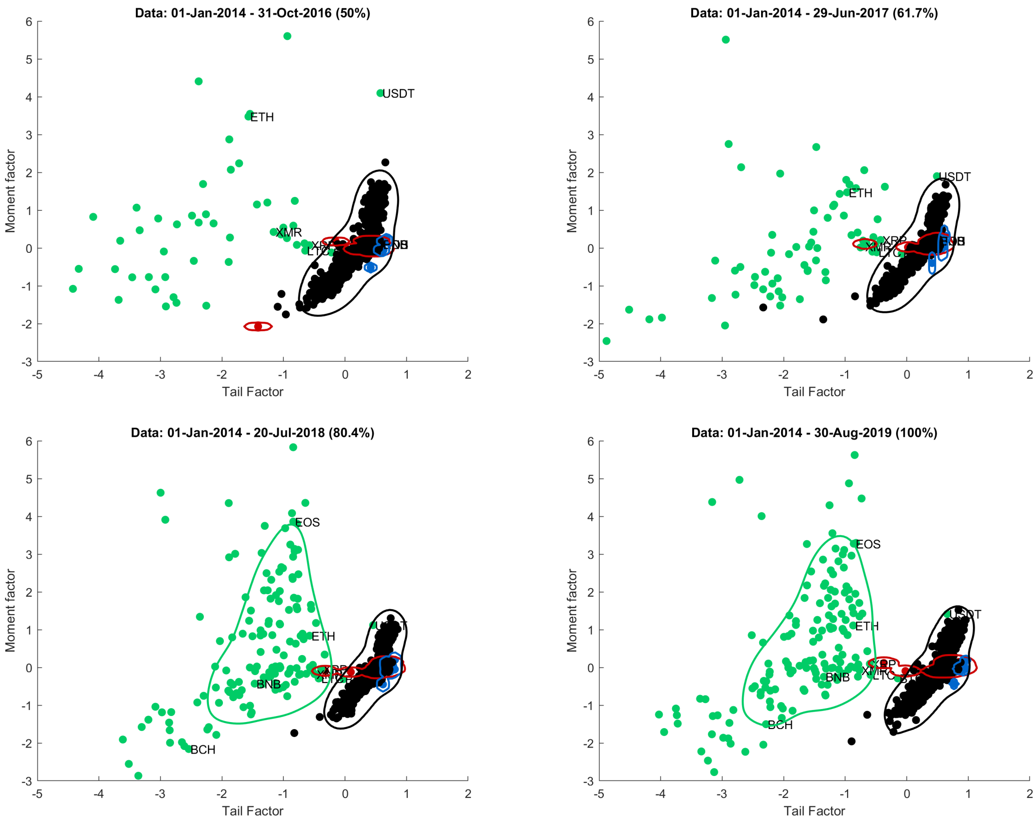

Figure 11: The evolution of the assets universe using the expanding window approach. The colour code is the following: green: cryptocurrencies, black: stocks, red: commodities, blue: exchange rates.

DFA_Cryptos

Figure 11 presents a snapshot of the evolution of the assets universe using the expanding window approach. [4] Looking at the evolution of the assets universe, it appears that individual cryptocurrencies tend to develop over time similar characteristics (synchronic evolution) that make them fully distinguishable from classical assets.

In order to test this behaviour, we are using the Likelihood Ratio associated to model 3, estimated using the expanding window approach previously described. The Likelihood Ratio for this model can be defined as:

$$LR(\widehat{\boldsymbol{\beta}}) = -2(\log L(\widehat{\boldsymbol{\beta}}) - \log L(\widehat{\boldsymbol{\beta}_{\boldsymbol{s}}})), \tag{17}$$

where $L(\widehat{\boldsymbol{\beta}_{\boldsymbol{s}}})$ is the likelihood of a saturated model that fits perfectly the sample, while $L(\widehat{\boldsymbol{\beta}})$ is the likelihood of the estimated model. In the language of Binary Logistic Regression, the Likelihood Ratio from the Equation 17 is called deviance (Hosmer and Lemeshow, 2010) and is a measure of model goodness-of-fit, with large values indicating models with poor classification power. The deviance is always positive, being zero only for the perfect fit. In order to derive the statistical significance of the classification, we compare the Likelihood Ratios of the estimated model and of the intercept-only model. Thus, we compute the difference of the likelihood ratios

$$D = LR(\widehat{\boldsymbol{\beta}}) - LR(0), \tag{18}$$

where asymptotically $D \sim \chi^2(1)$, $LR(0)$ being the likelihood ratio of the intercept-only model. In fact, we are estimating m models, where $m = T - t_0 - 1 = 740$. For each model we report the Likelihood Ratio (Figure 12) and the p-value associated to Equation 18 (see Figure 13). Large p-values indicates that the model might not differ statistically from an intercept-only model.

By examining the evolution of the Likelihood Ratios, we can observe a trend change for the tail-factor-based model, starting January 2018, when the cryptocurrency market

[4] The daily evolution of the assets universe, for the period 10/31/2016-08/30/2019, is depicted in the video *Crypto_movie*, attached to this paper as supplementary material.

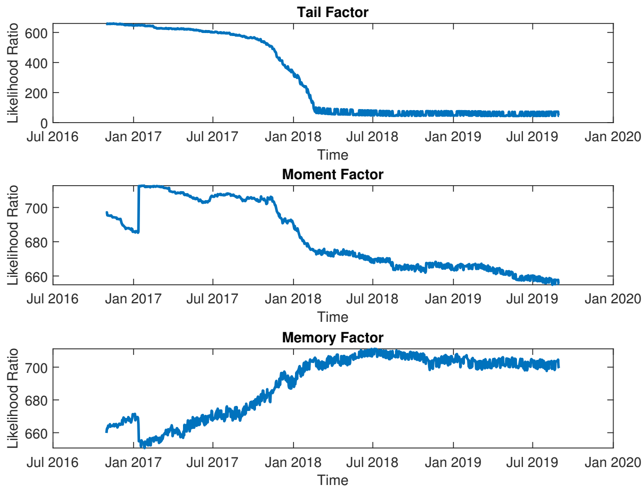

Figure 12: Likelihood Ratios for model (3), estimated on the time period 10/31/2016-08/30/2019, using an expanding window approach

CONV_Cryptos

collapsed after the historical maximum of Bitcoin from December 2017. Thus, the Likelihood Ratio converges to zero, pointing out the ability of the tail factor to discriminate between cryptocurrencies and classical assets.

The most important implication of this finding is the validity of synchronicity phenomenon among cryptocurrencies: in their evolution, the individual cryptocurrencies have developed similar characteristics (longer tails, higher volatility, higher propensity to extreme negative returns), that differentiate them from classical assets and position them as a new, different species in the ecosystem of financial instruments.

5. Conclusions

In this paper we applied various classification techniques in order to discriminate between cryptocurrencies and classical assets, like stocks, exchange rates and commodities. Through the means of dimensionality reduction techniques and classification techniques, we proved that most of the variation among cryptocurrencies, stocks, exchanges rates and commodities can be explained by three factors: the tail factor, the moment fac-

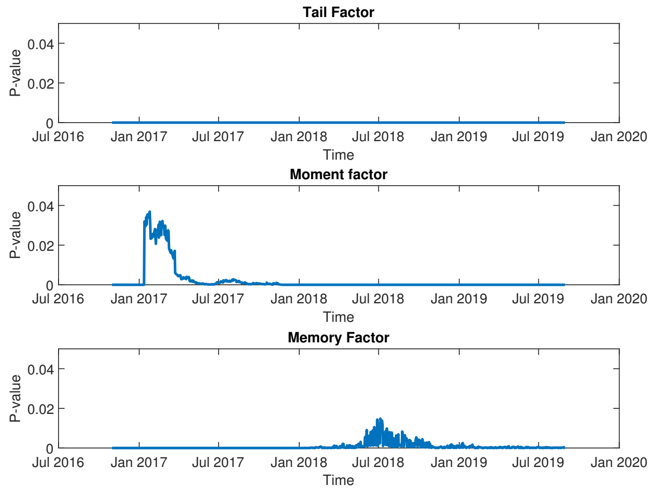

Figure 13: p-values for Equation 18, estimated on the time period 10/31/2016-08/30/2019, using an expanding window approach

CONV_Cryptos

tor and the memory factor. Our analysis revealed that the main difference between cryptocurrencies and classical assets, in terms of properties of the distribution of daily log-returns, is the tail behaviour, both in the left and in the right tail of the distribution. The moments of the distribution and the GARCH/ARCH parameters are of subliminal importance for discriminating between cryptocurrencies and classical assets.

Based on the tail factor profile, we can conclude that a random asset is likely to be a cryptocurrency if it has the following properties: very long tails of the log-returns distribution (in terms of the left and right quantile and the conditional tail expectation), high variance, high value of the α-stable scale parameter and value of the α-stable tail index closer to 1. Moreover, cryptocurrencies are completely separated by the other types of assets, as proved by Maximum Variance Components Split method. From the point of view of the risk analysts and regulators, the non-linear classification techniques applied on the factors extracted provide proficient results in order to discriminate between cryptocurrencies and the other assets.

Through the means of an expanding window approach, we are able to depict the evolutionary dynamics of cryptocurrencies universe and show how the clusters formed by projecting the multidimensional dataset on the main factors converge over time. By looking at the assets universe as a complex ecosystem, we are able to conclude that cryptocurrencies exhibit both a synchronic evolution (individual cryptocurrencies develop similar characteristics over time) and a divergent evolution, as different species, compared to classical assets.

Acknowledgements

Financial support from the Deutsche Forschungsgemeinschaft, Germany via IRTG 1792 “High Dimensional Non Stationary Time Series”, Humboldt-Universität zu Berlin, Germany, is gratefully acknowledged.

This research has received funding from the European Union’s Horizon 2020 research and innovation program FIN-TECH: A Financial supervision and Technology compliance training programme under the grant agreement No. 825215 (Topic: ICT35-2018, Type of action: CSA).

Appendix A: Cryptocurrency literature

Authors	Assets	Sample	Findings
Dyhrberg (2016a)	BTC, USD/EUR, USD/GBP, FTSE index.	2010-2015, daily data.	BTC can act as a hedge between UK equities and the USD.
Dyhrberg (2016b)	BTC, Federal funds rate, USD/EUR, USD/GBP, FTSE index, Gold futures, Gold cash.	2010-2015, daily data.	BTC is somewhere in between a currency (USD) and a commodity (Gold).
Bariviera et al. (2017)	BTC, USD/EUR, USD/GBP.	2011-2017, daily data. 2013-2016, hf data.	BTC presents large volatility and long-range memory (Hurst exponent higher than 0.5). BTC standard deviation is ten times greater than other currencies.
Baur et al. (2018)	BTC, Federal funds rate, USD/EUR, USD/GBP, FTSE index, Gold futures, Gold cash.	2010-2015, daily data.	BTC returns are not a hybrid of Gold and USD returns.
Caporale et al. (2018)	BTC, LTC, Ripple, Dash.	2013-2017, daily data.	The four cryptocurrencies exhibit persistence (Hurst exponent higher than 0.5), yet the persistency degree changes over time.
Härdle et al. (2018)	BTC, XRP, LTC, ETH, gold and S&P500.	2016-2018, daily data.	BTC, XRP, LTC, ETH exhibit higher volatility, skewness and kurtosis compared to Gold and S&P500 daily returns.
Henriques and Sadorsky (2018)	BTC and five exchange traded funds (ETFs): US equities (SPY), US bonds (TLT), US real estate (VNQ), Europe and Far East equities (EFA), and Gold (GLD).	2011-2017, daily data.	BTC can be a substitute for Gold in an investment portfolio, achieving a higher risk adjusted return.
Jiang et al. (2018)	BTC.	2010-2017, daily data.	Long-term memory and high degree of inefficiency ratio exists in the BTC market.
Klein et al. (2018)	BTC, CRIX index, Gold, Silver, crude oil, West Texas Intermediate (WTI), the S&P500, MSCI World and MSCI Emerging Markets 50.	2011-2017, daily data.	BTC returns have the highest mean and standard deviation.
Selmi et al. (2018)	BTC, Gold, Brent crude oil.	2011-2017, daily data.	Both BTC and Gold would serve the roles of a hedge, a safe haven and a diversifier for oil price movements.
Stosic et al. (2018)	Top 119 cryptocurrencies.	2016-2017, daily data.	Collective behaviour of the cryptocurrency market.
Takaishi (2018)	BTC, GBP/USD.	2014-2016, hf data.	The 1-min return distribution of BTC is fat-tailed, with high kurtosis; BTC time series exhibits multifractality.
Urquhart (2016)	BTC.	2010-2016, daily data.	Hurst statistic indicates strong anti-persistence (values lower than 0.5).
Wei (2018)	456 different cryptocurrencies.	2017, daily data.	Lower volatility for liquid cryptocurrencies. Illiquid cryptocurrencies exhibit strong return anti-persistence in the form of a low Hurst exponent.
Zhang et al. (2018)	70 % of cryptocurrencies market.	2013-2018, daily data.	Heavy tails, quickly decaying returns autocorrelations, slowly decaying autocorrelations for absolute returns, volatility clustering, leverage effects, long-range dependence, power-law correlation between price and volume.
Borri (2019)	BTC, ETH, LTC, XRP, Gold Bullion, the CBOE volatility index (VIX), the S&P400 and S&P500.	2017-2018, daily data.	Cryptocurrencies exhibit large and volatile return swings, and are riskier than most of the other assets.

References

E. L. Allwein, R. E. Schapire, and Y. Singer. Reducing Multiclass to Binary: A Unifying Approach for Margin Classifiers. *Journal of Machine Learning Research*, 1:113–141, 2000.

A. F. Bariviera, M. J. Basgall, W. Hasperué, and M. Naiouf. Some stylized facts of the Bitcoin market. *Physica A: Statistical Mechanics and its Applications*, 484:82–90, 2017. ISSN 0378-4371. doi: 10.1016/j.physa.2017.04.159. URL http://www.sciencedirect.com/science/article/pii/S0378437117304697.

D. J. Bartholomew. *Analysis of multivariate social science data*. Chapman & Hall/CRC Statistics in the Social and Behavioral Sciences. CRC Press, Boca Raton, Florida, 2nd ed. edition, 2011. ISBN 1584889616.

D. G. Baur, T. Dimpfl, and K. Kuck. Bitcoin, gold and the US dollar – A replication and extension. *Finance Research Letters*, 25:103–110, 2018. ISSN 1544-6123. doi: 10.1016/j.frl.2017.10.012. URL http://www.sciencedirect.com/science/article/pii/S1544612317305093.

N. Borri. Conditional tail-risk in cryptocurrency markets. *Journal of Empirical Finance*, 50:1–19, 2019. ISSN 09275398. doi: 10.1016/j.jempfin.2018.11.002.

G. M. Caporale, L. Gil-Alana, and A. Plastun. Persistence in the cryptocurrency market. *Research in International Business and Finance*, 46:141–148, 2018. ISSN 0275-5319. doi: 10.1016/j.ribaf.2018.01.002. URL http://www.sciencedirect.com/science/article/pii/S0275531917309200.

B. A. Cerny and H. F. Kaiser. A Study Of A Measure Of Sampling Adequacy For Factor-Analytic Correlation Matrices. *Multivariate behavioral research*, 12(1):43–47, 1977. ISSN 0027-3171. doi: 10.1207/s15327906mbr1201_3.

S. Corbet, B. Lucey, A. Urquhart, and L. Yarovaya. Cryptocurrencies as a financial asset: A systematic analysis. *International Review of Financial Analysis*, 62:182–199, 2019.

ISSN 10575219. doi: 10.1016/j.irfa.2018.09.003. URL http://www.sciencedirect.com/science/article/pii/S1057521918305271.

N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, Cambridge, 2000. ISBN 9780511801389. doi: 10.1017/CBO9780511801389.

A. H. Dyhrberg. Bitcoin, gold and the dollar - A GARCH volatility analysis. *Finance Research Letters*, 16:85–92, 2016a. ISSN 1544-6123. doi: 10.1016/j.frl.2015.10.008. URL http://www.sciencedirect.com/science/article/pii/S1544612315001038.

A. H. Dyhrberg. Hedging capabilities of bitcoin. Is it the virtual gold? *Finance Research Letters*, 16:139–144, 2016b. ISSN 1544-6123. doi: 10.1016/j.frl.2015.10.025.

A. ElBahrawy, L. Alessandretti, A. Kandler, R. Pastor-Satorras, and A. Baronchelli. Evolutionary dynamics of the cryptocurrency market. *Royal Society Open Science*, 4 (11):170623, 2017. ISSN 2054-5703. doi: 10.1098/rsos.170623.

W. Feng, Y. Wang, and Z. Zhang. Can cryptocurrencies be a safe haven: a tail risk perspective analysis. *Applied Economics*, 50(44):4745–4762, 2018. ISSN 0003-6846. doi: 10.1080/00036846.2018.1466993.

T. Fischer, C. Krauss, and A. Deinert. Statistical Arbitrage in Cryptocurrency Markets. *Journal of Risk and Financial Management*, 12(1):31, 2019. doi: 10.3390/jrfm12010031.

R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188, 1936. ISSN 20501420. doi: 10.1111/j.1469-1809.1936.tb02137.x.

W. Härdle and L. Simar. *Applied multivariate statistical analysis: [R & Matlab codes]*. Springer, Berlin and Heidelberg, 3. ed. edition, 2012. ISBN 3642172296.

W. K. Härdle, C. Harvey, and R. Reule. Understanding Cryptocurrencies. *IRTG 1792 Discussion Paper*, 2018-044, 2018. URL https://www.wiwi.hu-berlin.de/

de/forschung/irtg/results/discussion-papers/discussion-papers-2017-1/irtg1792dp2018-044.pdf.

I. Henriques and P. Sadorsky. Can Bitcoin Replace Gold in an Investment Portfolio? *Journal of Risk and Financial Management*, 11(3):48, 2018. doi: 10.3390/jrfm11030048.

D. W. Hosmer and S. Lemeshow. *Applied logistic regression*. A Wiley-Interscience publication. John Wiley, New York, 2nd ed., [repr.] edition, 2010. ISBN 0471356328.

I.M.F. Treatment of Crypto Assets in Macroeconomic Statistics. *International Monetary Fund, Statistics Department*, 2019. URL https://www.imf.org/external/pubs/ft/bop/2019/pdf/Clarification0422.pdf.

Y. Jiang, H. Nie, and W. Ruan. Time-varying long-term memory in Bitcoin market. *Finance Research Letters*, 25:280–284, 2018. ISSN 1544-6123. doi: 10.1016/j.frl.2017.12.009.

H. F. Kaiser. An index of factorial simplicity. *Psychometrika*, 39(1):31–36, 1974. ISSN 0033-3123. doi: 10.1007/BF02291575.

H. F. Kaiser. A Revised Measure of Sampling Adequacy for Factor-Analytic Data Matrices. *Educational and Psychological Measurement*, 41(2):379–381, 1981. ISSN 0013-1644. doi: 10.1177/001316448104100216.

T. Klein, H. Pham Thu, and T. Walther. Bitcoin is not the New Gold – A comparison of volatility, correlation, and portfolio performance. *International Review of Financial Analysis*, 59:105–116, 2018. ISSN 10575219. doi: 10.1016/j.irfa.2018.07.010.

I. A. Koutrouvelis. Regression-Type Estimation of the Parameters of Stable Laws. *Journal of the American Statistical Association*, 75(372):918, 1980. ISSN 01621459. doi: 10.2307/2287182.

I. A. Koutrouvelis. An iterative procedure for the estimation of the parameters of stable laws. *Communications in Statistics - Simulation and Computation*, 10(1):17–28, 1981. ISSN 0361-0918. doi: 10.1080/03610918108812189.

J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pages 281–297, Berkeley, Calif., 1967. University of California Press. URL https://projecteuclid.org/euclid.bsmsp/1200512992.

M. Mirtaheri, S. Abu-El-Haija, F. Morstatter, G. V. Steeg, and A. Galstyan. Identifying and Analyzing Cryptocurrency Manipulations in Social Media. 2009. URL http://arxiv.org/pdf/1902.03110v1.

N. J. Nagelkerke. A note on a general definition of the coefficient of determination. *Biometrika*, 78(3):691–692, 1991. ISSN 0006-3444. doi: 10.1093/biomet/78.3.691.

J. A. Núñez, M. I. Contreras-Valdez, and C. A. Franco-Ruiz. Statistical analysis of Bitcoin during explosive behavior periods. *PloS one*, 14(3):e0213919, 2019. doi: 10.1371/journal.pone.0213919.

S. T. Rachev and S. Mittnik. *Stable Paretian models in finance*. Series in financial economics and quantitative analysis. J. Wiley, Chichester, 2000. ISBN 9780471953142.

R. Selmi, W. Mensi, S. Hammoudeh, and J. Bouoiyour. Is Bitcoin a hedge, a safe haven or a diversifier for oil price movements? A comparison with gold. *Energy Economics*, 74:787–801, 2018. ISSN 01409883. doi: 10.1016/j.eneco.2018.07.007.

D. L. Stevens. Financial Characteristics of Merged Firms: A Multivariate Analysis. *The Journal of Financial and Quantitative Analysis*, 8(2):149, 1973. ISSN 00221090. doi: 10.2307/2330007.

D. Stosic, D. Stosic, T. B. Ludermir, and T. Stosic. Collective behavior of cryptocurrency price changes. *Physica A: Statistical Mechanics and its Applications*, 507:499–509, 2018. ISSN 0378-4371. doi: 10.1016/j.physa.2018.05.050.

B. G. Tabachnick and L. S. Fidell. *Using multivariate statistics*. Pearson Education, Boston, 6th ed. edition, 2013. ISBN 0205890814.

T. Takaishi. Statistical properties and multifractality of Bitcoin. *Physica A: Statistical Mechanics and its Applications*, 506:507–519, 2018. ISSN 0378-4371. doi: 10.1016/j.physa.2018.04.046.

Transparency Market Research. Cryptocurrency Market to reach US$ 6,702.1 mn by 2025. 2018. URL https://www.transparencymarketresearch.com/cryptocurrency-market.html.

A. Urquhart. The inefficiency of Bitcoin. *Economics Letters*, 148:80–82, 2016. ISSN 01651765. doi: 10.1016/j.econlet.2016.09.019.

M. Veillete. STBL: Alpha stable distributions for MATLAB. *Matlab Central File Exchange*, 2012. URL https://www.mathworks.com/matlabcentral/fileexchange/37514-stbl-alpha-stable-distributions-for-matlab.

W. C. Wei. Liquidity and market efficiency in cryptocurrencies. *Economics Letters*, 168:21–24, 2018. ISSN 01651765. doi: 10.1016/j.econlet.2018.04.003.

Y. G. Yatracos. Variance and Clustering. *Proceedings of the American Mathematical Society*, 126(04):1177–1179, 1998. ISSN 00029939. doi: 10.1090/S0002-9939-98-04524-9.

Y. G. Yatracos. The asymptotic distribution of a cluster-index for i.i.d. normal random variables. *The Annals of Applied Probability*, 19(2):585–595, 2009. ISSN 1050-5164. doi: 10.1214/08-AAP553.

Y. G. Yatracos. Detecting Clusters in the Data from Variance Decompositions of Its Projections. *Journal of Classification*, 30(1):30–55, 2013. ISSN 0176-4268. doi: 10.1007/s00357-013-9124-9.

N. Yoshino and F. Taghizadeh-Hesary. Analytical framework on credit risks for financing small and medium-sized enterprises in Asia. *Asia-Pacific Development Journal*, 21(2):1–21, 2015. doi: 10.18356/18bf1a40-en.

W. Zhang, P. Wang, X. Li, and D. Shen. Some stylized facts of the cryptocurrency market. *Applied Economics*, 50(55):5950–5965, 2018. ISSN 0003-6846. doi: 10.1080/00036846.2018.1488076.