Multivariate Hidden Markov model: An application to study correlations among cryptocurrency log-returns

Fulvia Pennoni* Francesco Bartolucci† Gianfranco Forte‡
Ferdinando Ametrano§

VERY PRELIMINARY DRAFT. PLEASE DO NOT QUOTE

Abstract

We propose a multivariate hidden Markov model to explain the price evolution of the major cryptocurrencies. We model jointly the daily log-returns of Bitcoin, Ethereum, Ripple, Litecoin, and Bitcoin Cash. The observed log-returns are assumed to be correlated according to a variance-covariance matrix conditionally on a latent Markov process having a finite number of states. For the purpose of comparing states according to their volatility we estimate the specific variance-covariance matrix of each state. Maximum likelihood estimation of the model parameters is carried out by the Expectation-Maximization algorithm. The hidden states represent different phases of the market that are identified according to estimated expected values of the log-returns and to the estimated volatility. We reach interesting results in detecting these phases of the market and the implied transition dynamics. We also find evidence of structural medium term trend in the correlation structure of Bitcoin with the other cryptocurrencies.

KEYWORDS: BITCOIN, BITCOIN CASH, DECODING, ETHEREUM, EXPECTATION-MAXIMIZATION ALGORITHM, LITECOIN, RIPPLE, TIME-SERIES

JEL classification codes: C32, C51, C53.

*Department of Statistics and Quantitative Methods, University of Milano-Bicocca (IT), email: fulvia.pennoni@unimib.it

†Department of Economics, Università di Perugia (IT), email: francesco.bartolucci@unipg.it

‡Department of Business and Law and Crypto Asset Lab, University of Milano-Bicocca (IT), email: gianfranco.forte@unimib.it

§Crypto Asset Lab, Department of Business and Law, University of Milano-Bicocca (IT), email: ferdinando.ametrano@unimib.it

1 Introduction

Following the seminal paper of Satoshi Nakamoto (Satoshi, 2008) and the creation of the Bitcoin network in 2009, an increasing number of crypto-assets have appeared. Almost all are of little interest being just clones of the first without any real functional innovation and/or trading liquidity. A few exceptions exist that have become relevant enough to be considered as investable assets. Therefore, crypto-assets time-series nowadays consist of multidimensional and complex data and these assets represent the most volatile and challenging financial market (Borri, 2019).

We aim to monitor financial asset price series for the main cryptocurrencies by using a popular statistical and unsupervised machine learning method that is based on a multivariate Hidden Markov Model (HMM); see Cappé et al. (1989), Mamon and Elliott (2007), and Zucchini et al. (2017) for details on the model in the context of time-series data and Bartolucci et al. (2013) in the context of longitudinal data. The HMM may be cast into the literature of finite mixture models (McLachlan and Peel, 2000), as it may be seen as a mixture model with a particular dependence structure across variables referred to different time points. The use of this approach is motivated by the fact that the HMM provides a flexible framework for many financial applications and it allows us to incorporate stochastic volatility in a rather simple form. A comparison with stochastic volatility models has been proposed by Genon-Catalot et al. (2000). From the pioneering work of Akaike (1998) showing that the ARMA process can be represented by a Markovian structure, many works have been proposed in the literature. Hamilton (1989), for example, proposed a model where the latent regime follows a Markov process, and several articles appeared more recently in this field; see, among others, Rossi and Gallo (2006), Mamon and Elliott (2007), Langrock et al. (2012), De Angelis and Paas (2013), Giudici and Abu Hashish (2020), and Lin et al. (2020).

In order to select the data for our application, we avoid going into the debate on the representativeness of the different cryptocurrencies. More specifically we focus on the market data referred to five cryptocurrencies: Bitcoin, Ethereum, Ripple, Litecoin, and Bitcoin Cash. The market is ruled by Bitcoin but it is in continuous and very fast

evolution. For example Trimborn and Härdle (2018) proposed a dynamic mechanism of composition for the construction of an overall index; in other cases the choice of the sample is dictated by specific objectives such as the description of the technological evolution as in Wang and Vergne (2017). In our applicative example we then follow the classical principles based on volumes and capitalization of selected crypto-assets considering those currently accounting for more than 90% of market capitalization and transaction volumes.

Unlike the prevailing literature, in which applications of switching models are focused exclusively on the estimation and prediction of volatility and consider the expected log-returns as unpredictable parameters (Ang and Bekaert, 2002), we proceed in line with De Angelis and Paas (2013) sharing the idea of also modeling the conditional means of the time-series. Furthermore, we model the log-returns of crypto-assets taking into account their correlation structure. An accurate evaluation of the conditional means might improve time-series classification. Stable periods, crises, and financial bubbles differ significantly for mean returns and structural levels of covariance.

We employ a multivariate HMM to account for the daily log-returns of the five mentioned cryptocurrencies. The reason for considering multiple time-series jointly instead of a single series (Huang et al., 2019) is that there are sideway movements in the long-term trends and we aim to identify actual trend change signals in the market. We assume that the daily log-return of each crypto is generated by a specific probabilistic distribution associated to the hidden state. The Expectation-Maximization (EM) algorithm (Baum et al., 1970; Dempster et al., 1977) is employed for the maximum likelihood estimation of the model parameters. The conditional distributions of the observed log-returns for various states of the hidden variables are taken from the Gaussian family with different means, variances, and covariances. Using the market prices collected over a three-year time period from August 2, 2017, to February 27, 2020, we identify suitable states representing relevant phases of the market. We predict the *a posteriori* most likely sequence of hidden states obtained through the so-called *decoding* based on the estimated maximum posterior probabilities to visit every state.

The remainder of the paper is organized as follows. In Section 2 we define the notation

and the quantities of interest for the proposed HMM and we illustrate the maximum likelihood estimation via the Expectation-Maximization algorithm. In Section 3 we describe the data and the reference markets. In Section 4 we show the results of our application. Finally, Section 5 provides main conclusions.

2 Proposed model

We consider a time-series $\boldsymbol{y}_t$, $t = 1, 2, \ldots$, where each element y_{tj}, with $j = 1, \ldots, r$, corresponds to the log-return of asset j among those considered. We will use $\boldsymbol{y}_t$ to denote the random vector at time t of one its realizations, in a way that will be clear from the context; the same convention will be applied to scalar random variables. In the following, we first review the HMM assumptions for our specific formulation and then we outline the steps of the EM algorithm for its estimation.

2.1 HMM assumptions

The main assumption of the HMM is that the random vectors $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots$ are conditionally independent given a hidden process $u_1, u_2, \ldots$ that follows a Markov chain with k hidden states, labelled from 1 to k. The model includes two different sub-models, named as measurement model and structural model, which are described in more detail in the following.

The measurement model corresponds to the conditional distribution of every vector $\boldsymbol{y}_t$ given the underlying variable u_t, $t = 1, 2, \ldots$. In this regard, we assume a multivariate Gaussian distribution for the overall log-returns of each cryptocurrency, that is,

$$\boldsymbol{y}_t | u_t = u \sim N_r(\boldsymbol{\mu}_u, \boldsymbol{\Sigma}_u),$$

where $\boldsymbol{\mu}_u$ and $\boldsymbol{\Sigma}_u$ are, for hidden state u, the specific mean vector and variance-covariance matrix, respectively. Obviously, the conditional means in $\boldsymbol{\mu}_u$ define the expected log-return when the underlying chain is in state u, while the elements of $\boldsymbol{\Sigma}_u$ provide measures of volatility of each asset and correlation between pairs of asset. As will be clear in the following, different constraints may be conceived on these matrices, among which the

main is that of homoschedasticity: $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_u$, $u = 1, \ldots, k$.

The above assumptions imply that the conditional distribution of the time-series $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots$ given the sequence of hidden states may be expressed as

$$f(\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots | u_1, u_2, \ldots) = \prod_t \phi(\boldsymbol{y}_t; \boldsymbol{\mu}_{u_t}, \boldsymbol{\Sigma}_{u_t}),$$

where, in general, $\phi(\cdot; \cdot, \cdot)$ denotes the density of the multivariate Gaussian distribution, in our case of dimension r, with a certain mean vector and variance-covariance matrix.

The structural model for the distribution of the latent Markov process is based on initial and transition probabilities. These parameters are defined as

$$\lambda_u = p(u_1 = u), \quad u = 1, \ldots k,$$

and

$$\pi_{v|u} = p(u_t = v | u_{t-1} = u), \quad t = 2, \ldots, \ u, v = 1, \ldots, k,$$

and are collected in the initial probability vector $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_k)'$ and the transition matrix

$$\boldsymbol{\Pi} = \begin{pmatrix} \pi_{1|1} & \cdots & \pi_{1|k} \\ \vdots & \ddots & \vdots \\ \pi_{k|1} & \cdots & \pi_{k|k} \end{pmatrix}.$$

Consequently, we can easily obtain the probability of a sequence of hidden states $u_1, u_2, \ldots$ as

$$p(u_1, u_2, \ldots) = \lambda_{u_1} \prod_{t \geq 2} \pi_{u_t | u_{t-1}}.$$

Joint together, the measurement and the structural model implies that the manifest distribution of time-series has the following density function:

$$\begin{aligned} f(\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots) &= \sum_{u_1, u_2, \ldots} p(u_1, u_2, \ldots) f(\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots | u_1, u_2, \ldots) \\ &= \sum_{u_1} \pi_{u_1} \phi(\boldsymbol{y}_1; \boldsymbol{\mu}_{u_1}, \boldsymbol{\Sigma}_{u_1}) \sum_{u_2} \pi_{u_2|u_1} \phi(\boldsymbol{y}_2; \boldsymbol{\mu}_{u_2}, \boldsymbol{\Sigma}_{u_2}) \cdots, \end{aligned}$$

which, in practice, is computed by a forward recursion (Baum et al., 1970; Welch, 2003).

This recursion requires a number of operations linearly increasing with the number of observation times because it exploits the previous factorization.

2.2 Maximum likelihood estimation

With reference to an observed time-series of log-returns $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots$, the HMM log-likelihood function is defined as

$$\ell(\boldsymbol{\theta}) = \log f(\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots),$$

where $\boldsymbol{\theta}$ is the vector of all model parameters, that is, $\boldsymbol{\mu}_u$, $\boldsymbol{\Sigma}_u$, $u = 1, \ldots, k$, $\boldsymbol{\lambda}$, and $\boldsymbol{\Pi}$. By maximizing $\ell(\boldsymbol{\theta})$ we obtain estimates of these parameters and, for this aim, we employ the EM algorithm.

The EM algorithm alternates two steps until convergence in $\ell(\boldsymbol{\theta})$ and is based on the so-called complete-data log-likelihood, corresponding to the log-likelihood that could be computed also knowing the hidden states at every time occasion. This function, denoted by $\ell^*(\boldsymbol{\theta})$, may be decomposed as the sum of three components that may be maximized separately, which are defined as

$$\begin{aligned}
\ell_1^*(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_k) &= \sum_t \sum_u w_{tu} \log \phi(\boldsymbol{y}_t | \boldsymbol{\mu}_u, \boldsymbol{\Sigma}_u) \\
&= -\frac{1}{2} \sum_t \sum_u w_{tu} [\log(|2\pi \boldsymbol{\Sigma}_u|) + (\boldsymbol{y}_t - \boldsymbol{\mu}_u)' \boldsymbol{\Sigma}_u^{-1} (\boldsymbol{y}_t - \boldsymbol{\mu}_u)], \quad (1) \\
\ell_2^*(\boldsymbol{\lambda}) &= \sum_u w_{1u} \log \pi_u, \quad (2) \\
\ell_3^*(\boldsymbol{\Pi}) &= \sum_{t \geq 2} \sum_u \sum_v z_{tuv} \log \pi_{v|u}, \quad (3)
\end{aligned}$$

where $w_{tu} = I(u_t = u)$ is a dummy variable equal to 1 if the hidden process is in state u at time t and to 0 otherwise and $z_{tuv} = I(u_{t-1} = u, u_t = v) = z_{t-1,u} z_{tv}$ is the indicator variable for the transition from state u to state v at time occasion t.

The two steps of the EM algorithm are:

- **E-step**: it computes the posterior expected value of each indicator variable w_{tu}, $t = 1, 2, \ldots$, $u = 1, \ldots, k$, and z_{tuv}, $t = 2, \ldots$, $u, v = 1, \ldots, k$, given the observed

data. These expected values correspond to

$$\hat{w}_{tu} = p(u_t|\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots), \quad (4)$$

$$\hat{z}_{tuv} = p(u_{t-1} = u, u_t = v|\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots), \quad (5)$$

and their computation is performed by suitable back-forward recursions (Baum et al., 1970; Welch, 2003).

- **M-step**: the expected complete data log-likelihood is maximized with respect to the model parameters; this function is given by the sum of functions (1)-(3), once the indicator variables have been substituted by their expected values defined in equations (4) and (5). The parameters in the measurement model are then updated in a simple way as

$$\begin{aligned} \boldsymbol{\mu}_u &= \frac{1}{\sum_t \hat{w}_{tu}} \sum_t \hat{w}_{tu} \boldsymbol{y}_t, \\ \boldsymbol{\Sigma}_u &= \frac{1}{\sum_t \hat{w}_{tu}} \sum_t \hat{w}_{tu} (\boldsymbol{y}_t - \boldsymbol{\mu}_u)(\boldsymbol{y}_t - \boldsymbol{\mu}_u)', \end{aligned}$$

for $u = 1, \ldots, k$. Under the constraint of homoschedasticity, the latter is substituted by

$$\boldsymbol{\Sigma} = \frac{1}{T} \sum_t \sum_u \hat{w}_{tu} (\boldsymbol{y}_t - \boldsymbol{\mu}_u)(\boldsymbol{y}_t - \boldsymbol{\mu}_u)',$$

with T being the number of observation times. Regarding the parameters in the structural model, we simply have

$$\begin{aligned} \pi_u &= \hat{z}_{1u}, \quad u = 1, \ldots, k, \\ \pi_{v|u} &= \frac{1}{\sum_{t \geq 2} \hat{w}_{t-1,u}} \sum_{t \geq 2} \hat{z}_{tuv}, \quad u, v = 1, \ldots, k. \end{aligned}$$

The overall vector of estimates obtained at convergence is denoted by $\hat{\boldsymbol{\theta}}$.

Since the EM algorithm may converge to a local maximum not corresponding to the global maximum, common initialization strategies involve a multi-start rule from appropriate deterministic and random starting values. Deterministic starting values of

the parameters of the measurement model, $\boldsymbol{\mu}_u$ and $\boldsymbol{\Sigma}_u$, $u = 1, \ldots, k$, are defined on the basis of the descriptive statistics (mean vector and variance-covariance matrix) of the observed log-returns. The starting values for the initial probabilities π_u are chosen as $1/k$, for $u = 1, \ldots, k$, whereas for the transition probabilities we adopt the following rule: $\pi_{v|u} = (h+1)/(h+k)$ when $v = u$ and $\pi_{v|u} = 1/(h+k)$ when $v \neq u$, where h is a suitable positive constant.

The random starting rule is instead based on values drawn from a multivariate Gaussian distribution for $\boldsymbol{\mu}_u$, $u = 1, \ldots, k$, and on suitable normalized random numbers drawn from a uniform distribution between 0 and 1 for both initial and transition probabilities. The starting values for the variance-covariance matrices are again based on their sample counterpart.

An important aspect concerns the model selection in terms of the number of hidden states. When there are not substantial reasons to use a predefined number of states, we rely on the Bayesian Information Criterion (BIC; Schwarz, 1978), which is based on the following index

$$BIC_k = -2\hat{\ell}_k + \log(T)\#\text{par}, \tag{6}$$

where $\hat{\ell}_k$ denotes the maximum of the log-likelihood of the model with k states and #par denotes the number of free parameters equal to $k[r + r(r+1)/2] + k^2 - 1$ for the heteroschedastic model and to $kr + r(r+1)/2 + k^2 - 1$ for the homoschedastic one. Based on this criterion, we estimate a series of HMMs for increasing value k and we select the number of hidden states corresponding to the minimum value of the BIC index.

We can predict the most likely sequence of hidden states, through the so-called local decoding, but we can also use the global decoding that may be implemented through the Viterbi algorithm (Viterbi, 1967; Juang and Rabiner, 1991). Computational tools required for the estimation are available upon request from the authors. They are implemented by adapting suitable functions of the `R` package `LMest` (Bartolucci et al., 2017, 2020).

3 Application

Due to the lack of regulation and established best-practices, in the crypto-asset market it is quite common to have manipulated asset prices and trading volumes. For these reasons suitable criteria must be used to select crypto-assets for quantitative analyses. Approaches based on the so-called market capitalization are unreliable because it is not easy to define the real free-floating capital for every crypto-asset. Descriptive statistics on the markets are available, but from an economic point of view they are generally not relevant. In fact, due to the anonymity/pseudonymity of the blockchain protocol it is very hard to determine the amount of capital lost forever, making the calculations of the actual capital available for trading almost always impossible.

The adopted criteria to select the cryptocurrencies for our HMM application are same underlying the Crypto Asset Lab Index (to be published in 2021); each crypto-assets must:

- be a scarce digital bearer asset that cannot be duplicated and it is cryptographically secured;
- have a value that is not pegged to any other asset or currency;
- must be traded on at least two reliable exchanges[1];
- have no more than 80% of its combined 90-day trading volume on a single reliable exchange;
- be actively traded on reliable exchanges against traditional fiat currencies, stable-coins (i.e., crypto-assets pegged to fiat currencies), Bitcoin, or Ethereum.

[1] An exchange is reliable when it:

- has not been exposed as publishing fake or inflated trading volumes;
- is registered and has obtained the license to operate in its jurisdiction;
- provides open and reliable functioning API;
- applies trading fees.

Taking the above features into consideration we limited our analysis to the following cryptocurrencies: Bitcoin (BTC), Ethereum (ETH), Ripple (XRP), Litecoin (LTC), and Bitcoin Cash (BCH). For the seek of comparability on the liquidity side, our analysis focuses on a recent time span of three years, accounting for the more recent introduction and development of some selected cryptocurrencies, in particular XRP and LTC.

The data are provided by the Crypto Currency Lab® and are referred to 940 daily quotes over a three-year period from August 2, 2017, to February, 27, 2020[2]. In what follows we show some descriptive statistics of the five time-series log-returns

$$y_{tj} = \log(x_{t+1,j}/x_{tj}), \quad j = 1, \ldots, r, \ t = 1, \ldots, T,$$

with x_{tj} denoting the closing price on day t of asset j. Then we focus on the HMM estimation. Note that in this way we dispose of a series of $T = 939$ log-returns for $r = 5$ cryptocurrencies.

Figure 1 shows the BTC prices along with the daily log-returns for the whole period of observation. We notice the high level of volatility as well the clustering phenomena also typical of other financial assets. From the chart it is immediate to recognize two periods of sharp rise in price, at the end of 2017 and in the mid-2019, together with a central period with less volatility, but affected by a sudden collapse in November 2018. Figure 2 represents the daily log-returns of the five cryptocurrencies, highlighting the volatility characteristics common in the crypto-asset market.

Table 1 reports the observed variance-covariance matrix, while Table 2 reports the observed correlations and partial correlations. The correlations are in almost all cases above 0.5, and very high for the pair BCH-ETH. The correlation structure, however, is not so obvious to interpret in terms of partial correlation, suggesting that the BTC dominance does not necessarily results in a unique co-moving driver.

[2]The data comes from a selection of 8 exchanges out of 81 available (BitFlyer, BitStamp, Bittrex, Coinbase, Gemini, itBit, Kraken, Poloniex). The cryptocurrency market is affected by a marked phenomenon of volume manipulation aimed at attracting customers from the stock exchanges. We have followed an exclusion criteria based on manipulation, also chosen by the Crypto Asset Lab® for the design of its index, and similar to that of Bitwise submitted to the SEC (see: `https://www.sec.gov/comments/sr-nysearca-2019-01/srnysearca201901-5164833-183434.pdf`) that further excludes BitFinex and Binance.

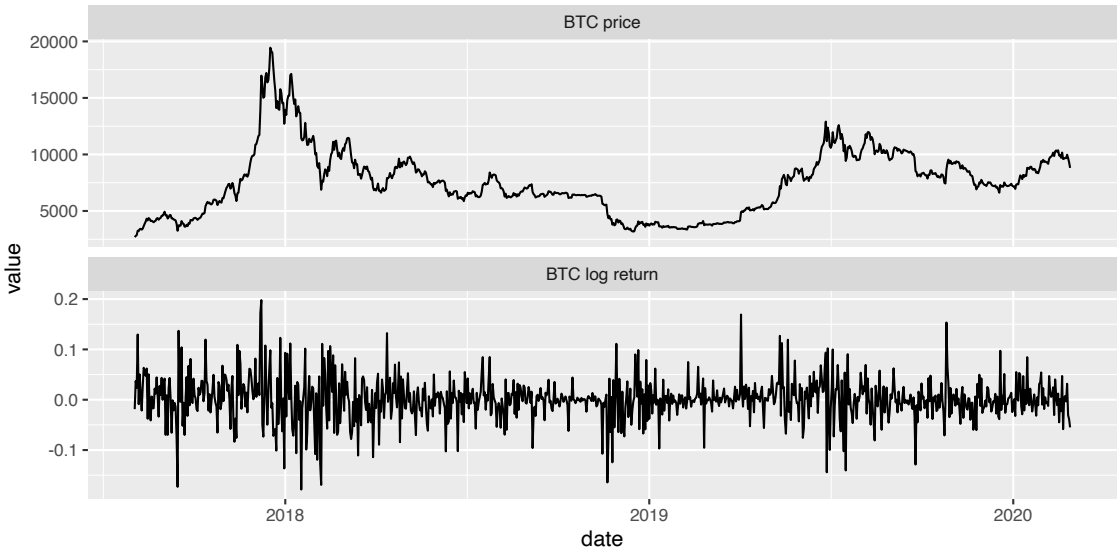


Figure 1: *Daily time-series of prices and log-returns of the BTC cryptocurrency (complete observations are referred to the period from August 2, 2017, to February 27, 2020).*

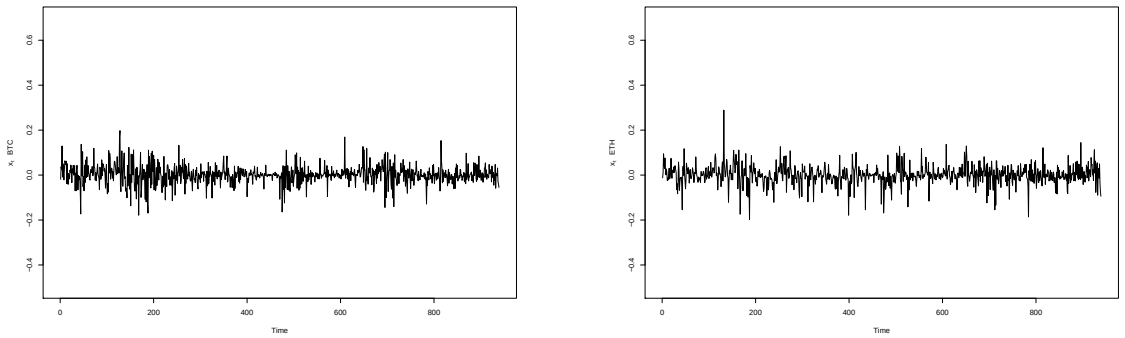
Table 1: *Observed variance-covariance matrix of the five cryptocurrencies.*

	BTC	ETH	XRP	LTC	BCH
BTC	0.15				
ETH	0.13	0.38			
XRP	0.09	0.23	0.28		
LTC	0.16	0.29	0.21	0.29	
BCH	0.19	0.45	0.27	0.35	0.61

4 Results

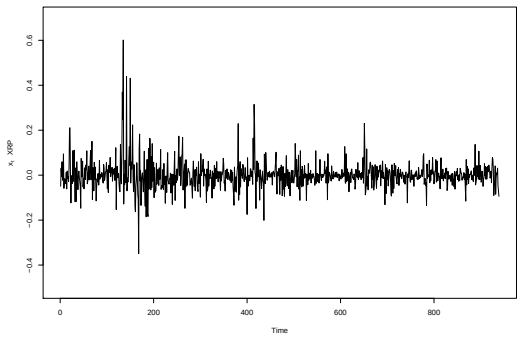
The proposed HMM model described above is estimated through the procedure presented in Section 2.2; for the sake of brevity, results are limited to the final selected model. The order (number of states, k) of the hidden distribution is selected according of the BIC (Schwarz, 1978) based on expression (6). The model is estimated for a number of hidden states ranging from 1 to 6 and the results are displayed in Table 3. The model selection strategy accounts for the multimodality of the likelihood function by using different sets of starting values for each run of the EM algorithm.

According to the results showed in Table 3, the best model corresponds to the heteroschedastic HMM with $k = 5$ hidden states with specific mean vectors and variance-

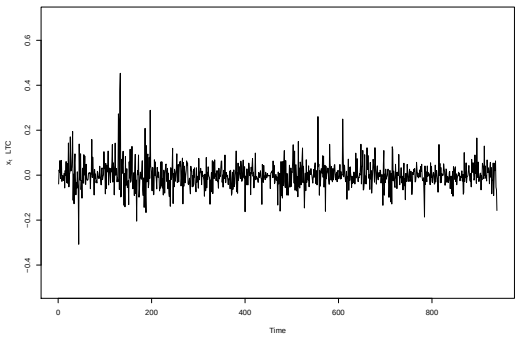


(a) BTC

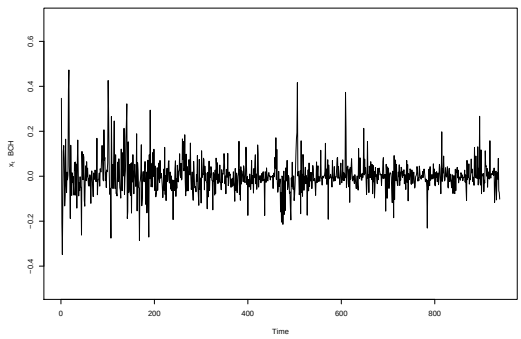
(b) ETH



(c) XRP



(d) LTC



(e) BCH

Figure 2: *Daily time-series of log-returns of the BTC, ETH, XRP, LTC, BCH cryptocurrencies based on closing prices (complete observations are referred to the period from August 2, 2017, to February 27, 2020).*

covariance matrices.

Table 2: *Observed correlation (left panel) and partial correlation (right panel) matrices of the five cryptocurrencies.*

	BTC	ETH	XRP	LTC	BCH	BTC	ETH	XRP	LTC	BCH
BTC	1.00					1.00				
ETH	0.55	1.00				-0.38	1.00			
XRP	0.44	0.71	1.00			-0.16	0.14	1.00		
LTC	0.74	0.86	0.73	1.00		0.63	0.46	0.37	1.00	
BCH	0.62	0.94	0.66	0.82	1.00	0.34	0.82	-0.04	-0.12	1.00

Table 3: *Results from the fitting of the multivariate HMMs to the daily log-returns of the BTC, ETH, XRP, LTC, BCH cryptocurrencies for increasing number of hidden states (k).*

k	log-likelihood	#par	BIC
1	7,785.46	15	-15,468.25
2	9,044.87	43	-17,795.41
3	9,334.88	68	-18,204.31
4	9,455.30	95	-18,260.35
5	9,565.06	124	-18,281.36
6	9,667.93	155	-18,274.90

4.1 HMM with five hidden states

We show the estimated expected log-returns given each state in Table 4. They represent the occurrence of a variety of situations happening on the market. According to these estimates, there are three negative regimes (1, 2, 3) and two positive regimes (4, 5). The second state, however, corresponds to cryptocurrency prices manifesting signs of stability, although the prevailing sign is negative.

From Table 5, reporting the estimated conditional variances and correlations, some interesting results emerge. First of all the correlations of BTC with the other cryptos are quite high and positive for the first three states having mainly negative or stable expected log-returns. On the other hand, the correlations for states 4 and 5 are lower and correspond to a more idiosyncratic behavior of the cryptos. It is interesting to note

Table 4: *Estimated expected log-returns for the HMM with $k = 5$ hidden states.*

	1	2	3	4	5
BTC	-0.0057	0.0054	-0.0013	0.0173	0.0159
ETH	-0.0044	-0.0016	-0.0020	0.0175	0.0126
XRP	-0.0067	-0.0051	-0.0039	0.0007	0.0629
LTC	-0.0090	0.0029	-0.0032	0.0121	0.0398
BCH	-0.0091	-0.0060	-0.0037	0.0634	-0.0016
average	-0.0070	-0.0009	-0.0028	0.0222	0.0259

that, in state 2, the correlation between BTC and XRP is high (0.68) but the partial correlation is low and negative (-0.18). Something similar happens between BTC and LTC, indicating that in this state of greater stability the dynamics of cryptocurrency are more mixed. In addition, in terms of volatility, it is clear that state 3 is the most volatile. If we therefore refer to the levels of log-returns in Table 4, states 1 and 3 are both marked by negative log-returns, but with a very different level of risk. It turns also out that state 1 is the only one characterized by significant falls of price and a marked volatility, which is typical of market crashes. We can therefore assume that also state 3, along with states 4 and 5, even if characterized by negative log-returns, represents a phase of relative stability of the prices as state 2.

Table 6 shows the estimated matrix of the transition probabilities among states. We remark that the highest persistence is estimated for states 2, 3, and 4. On the other hand, regimes 1 and especially 5 are less persistent. There is a quite high probability to transit from each state to the first meaning that in each asset take profit positions are frequent. As it stands, the first state can be considered as a "center of gravity" in terms of transition probability. Concerning the highest estimated transition probability from the less persistent state 5 to state 1 we notice that this result is not surprising, since considering state 5 as representative of main markedly positive log-returns, this transition can be read as the typical pull back following a substantial price increase.

Figure 3 illustrates the estimated posterior probabilities of being in latent state u, with $u = 1, \ldots, k$, at time t, with $t = 1, \ldots, T$, conditional on the observed time-series. Through these probabilities we are able to characterize the assets along time at different

market phases. Considering the trend line imposed on the plot and created by a smoothed local regression we notice an increasing tendency for state 3 and a decreasing tendency of states 4 and 5 over time. Moreover, apart for few exceptions there are not stable periods.

Table 5: *Estimated conditional correlations (lower triangle), variances (in bold, in diagonal) and partial correlations given all remaining variables (in italic, upper triangle) for each state of the HMM with $k = 5$ hidden states.*

State 1	BTC	ETH	XPR	LTC	BCH
BTC	**0.0019**	*-0.0404*	*0.0722*	*0.5347*	*0.1967*
ETH	0.3554	**0.0028**	*0.1060*	*0.0805*	*0.0561*
XRP	0.7705	0.3875	**0.0035**	*0.3919*	*0.0305*
LTC	0.9058	0.4016	0.8306	**0.0033**	*0.5011*
BCH	0.8501	0.3823	0.7581	0.8977	**0.0056**
State 2					
BTC	**0.0017**	*0.3531*	*-0.1846*	*-0.1072*	*0.5238*
ETH	0.7799	**0.0015**	*0.3110*	*0.2513*	*0.1188*
XRP	0.6822	0.8006	**0.0013**	*0.0845*	*0.5324*
LTC	0.6095	0.7265	0.7079	**0.0029**	*0.2916*
BCH	0.8254	0.8333	0.8579	0.7547	**0.0016**
State 3					
BTC	**0.0002**	*0.2714*	*0.2234*	*0.2655*	*0.2789*
ETH	0.6332	**0.0003**	*0.1702*	*0.0858*	*0.0227*
XRP	0.7323	0.5937	**0.0003**	*0.3167*	*0.2131*
LTC	0.7559	0.5792	0.7562	**0.0006**	*0.3488*
BCH	0.7394	0.5439	0.7179	0.7636	**0.0007**
State 4					
BTC	**0.0023**	*-0.1527*	*0.3547*	*0.1877*	*-0.3043*
ETH	0.1163	**0.0014**	*0.1897*	*0.0985*	*-0.0655*
XRP	0.6215	0.3303	**0.0021**	*0.6565*	*0.2106*
LTC	0.5977	0.3083	0.8058	**0.0028**	*-0.0709*
BCH	-0.2477	-0.0279	0.0024	-0.0802	**0.0221**
State 5					
BTC	**0.0061**	*0.1235*	*-0.0930*	*0.2351*	*0.3836*
ETH	0.2951	**0.0039**	*-0.0205*	*0.1710*	*0.0429*
XRP	0.2155	0.1047	**0.0255**	*0.0380*	*0.3890*
LTC	0.5324	0.3261	0.3044	**0.0163**	*0.3932*
BCH	0.5887	0.2729	0.4752	0.6259	**0.0136**

Figure 4 depicts the decoded states across all the days. It is based on the posterior probabilities showed in Figure 3. We estimate that the hidden state 1 is visited the

36.85%, state 2 the 16.19%, state 3 the 31.84%, state 4 the 8.41%, and state 5 the 6.71% of the $T = 939$ days. Therefore, states 4 and 5 occur in a small fraction of time occasions, especially before the second half of 2009 and also in single days.

Table 6: *Estimated transition probabilities for the HMM with $k = 5$ hidden states (in bold elements greater than 0.1).*

	1	2	3	4	5
1	**0.6879**	0.0548	**0.1722**	0.0175	0.0676
2	**0.1445**	**0.7145**	**0.1190**	0.0220	0.0000
3	**0.2035**	0.0825	**0.7140**	0.0000	0.0000
4	**0.1137**	0.0196	0.0000	**0.7757**	0.0909
5	**0.2441**	0.0791	0.0010	**0.1079**	**0.5678**

On the basis of Figures 3 and 4 some conclusions can be drawn. The recent evolution of the main cryptocurrencies is characterized by a prevalence towards phases of greater stability corresponding to states 2 and 3, and an evident reduction of episodes of marked price increase corresponding to market phases detected by states 4 and 5. These states are indeed the representation of another typical phenomenon of the crypto-assets namely, speculative bubbles. The existence of bubbles in the price dynamics of the BTC and other crypto-assets is a well-known feature of the evolution of these markets and contributed substantially to the high log-returns reported from 2009 to date. Such periods, intended as rapid price accelerations with an exponential or even explosive behavior, are one of the primary concerns for investors due to the risk posed by the subsequent bubble burst with extremes losses. Recently, Bouri et al. (2019) and Agosto and Cafferata (2020) focused on the links between crypto-assets in such periods of extreme rise and drop of prices, showing a relevant interconnection between cryptos during the price increase as well in the bubble burst. Perhaps the most relevant and widespread bubble is that of the final quarter of 2017, which quickly saw the Bitcoin reaching a value of \$10,000 and shortly thereafter peaks at more than \$20,000. The estimated posterior probabilities presented in Figure 3 are encouraging in detecting a trend of sharp decrease of these episodes that poses a serious limit for retail and institutional investors in considering cryptocurrencies as an investable asset.

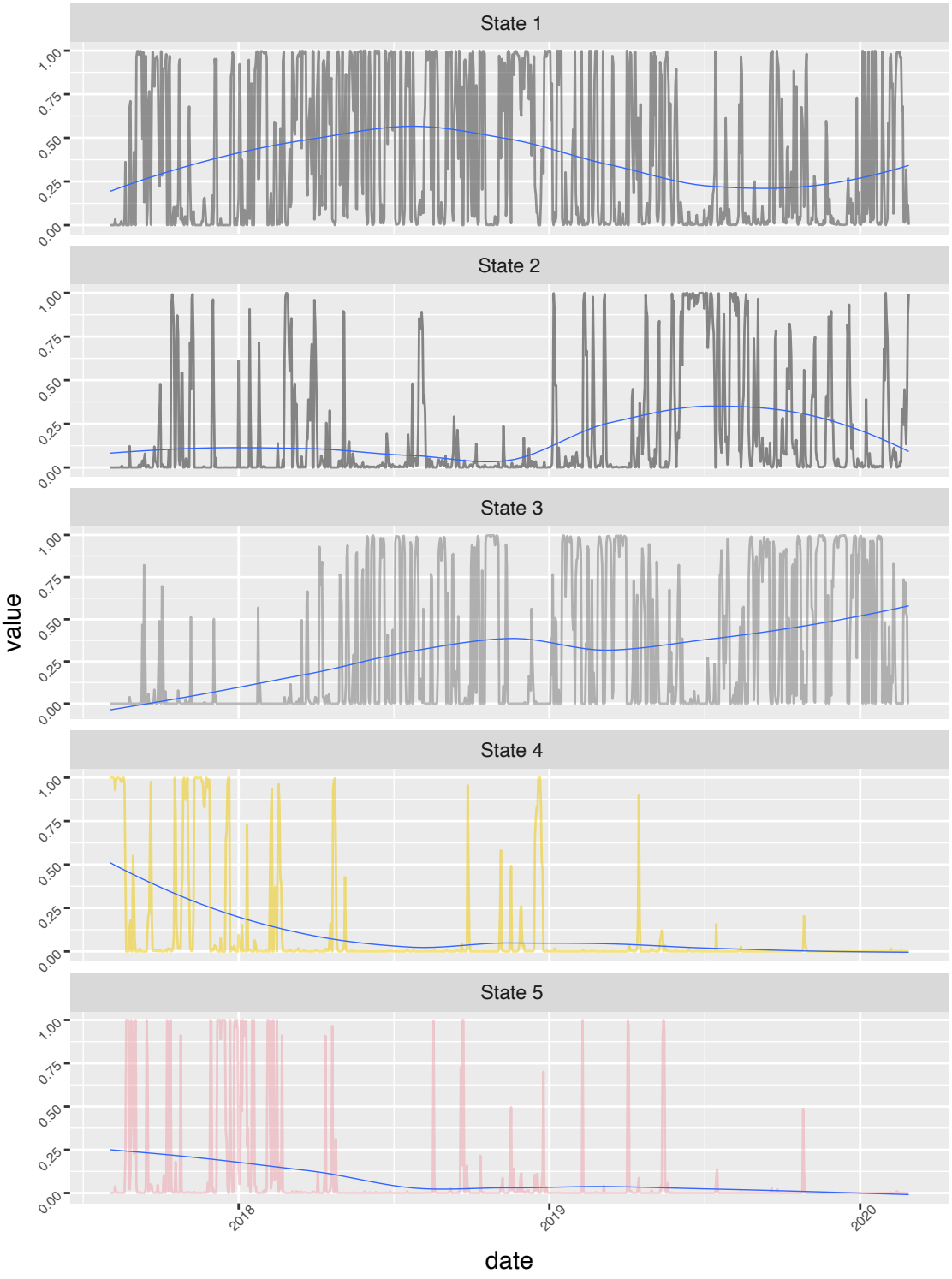


Figure 3: *Predicted posterior probabilities of the five states of the HMM with overimposed smoothed local regression lines (in blue).*

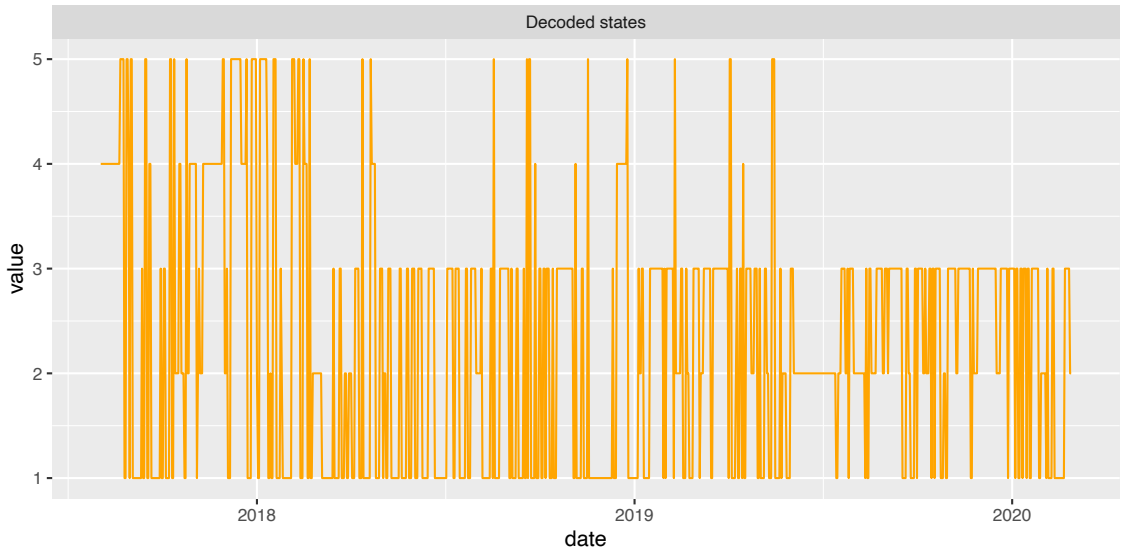


Figure 4: *Five decoded states for the estimated HMM.*

Figures 7-11 reported in the Appendix depict the observed log-returns, the predicted averages, and standard deviations over the time period for each cryptocurrency. The multivariate HMM with five states in our intention is not meant to provide unambiguous univariate predictions of log-returns or volatility, but the results are truly comforting. In particular, for Ripple, Litecoin, and Bitcoin Cash the model is able to timely detect regimes of high or low returns and volatilities.

In Figure 5 we show the realized log-returns of BTC along with the predicted values of the estimated HMM with $k = 5$ hidden states. We highlight the quite high adherence between them confirming that the model is able to account for a reverse trend in a fast way and its performance is good in both bull and bear market phases.

Finally, Figure 6 shows the estimated correlations of the BTC with all the other cryptocurrencies namely, XRP, ETH, LTC, BCH, along with the estimated trend according to a smooth local regression. A clear contribution of a multivariate model is to show if there are some structural trends in the interconnection between the assets. This aspect has received considerable attention in recent researches. Different studies investigated interconnectedness, co-movements, and volatility spillovers between cryptocurrencies applying different approaches; see Corbet et al. (2018), Chen et al. (2020), and Giudici and

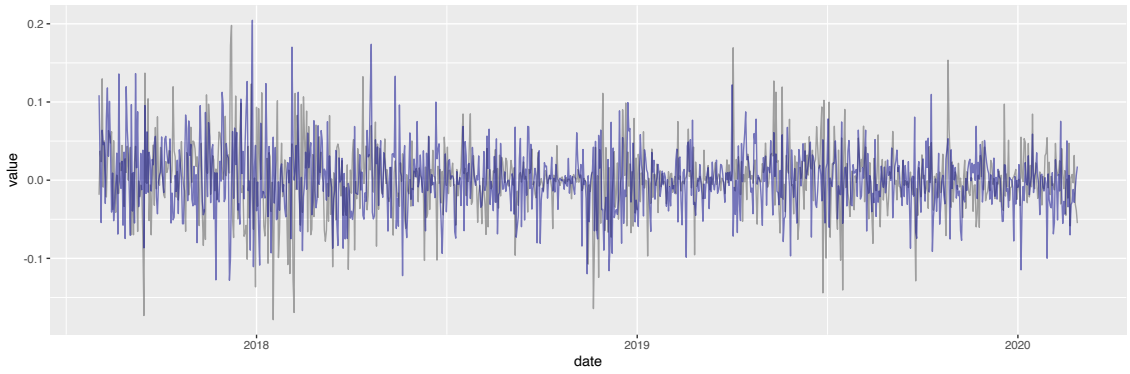


Figure 5: *Observed (gray lines) and predicted (blue lines) log-returns for the BTC of the HMM with $k = 5$ hidden states.*

Polinesi (2019) for a network analysis and Yi et al. (2018) and Giudici and Pagnottoni (2019, 2020) for a VAR analysis.

Our results confirm a medium term trend of greater correlation relative to BTC with the other cryptocurrencies. This conclusion is less evident only for Litecoin. An explanation of this evidence is the rising of a systemic risk for the whole market as it is more mature and liquid as in equity markets. Another explanation could be the rise of stablecoins pair dominance during 2018 concurring to the overall decline in the contribution of BTC pairs to total industry trade volume, with a pressure to a stronger link to USD of all cryptos and a subsequent increase in correlation.

5 Conclusions

We propose a multivariate Hidden Markov Model (HMM) to analyze log-returns of the main five cryptocurrencies: Bitcoin, Ethereum, Ripple, Litecoin, and Bitcoin cash. The narrow universe we selected fulfills the intention to concentrate on the more reliable, liquid, and less manipulated crypto-assets in the market. The choice of recent three years of data followed similar criteria of homogeneity between time-series especially with reference to the liquidity profile. The advantage of employing an HMM, as that proposed in this paper that includes state-specific expected log-returns, lies on the use of the surplus of information available in comparison to traditional regime-switching models that focus exclusively on volatility.

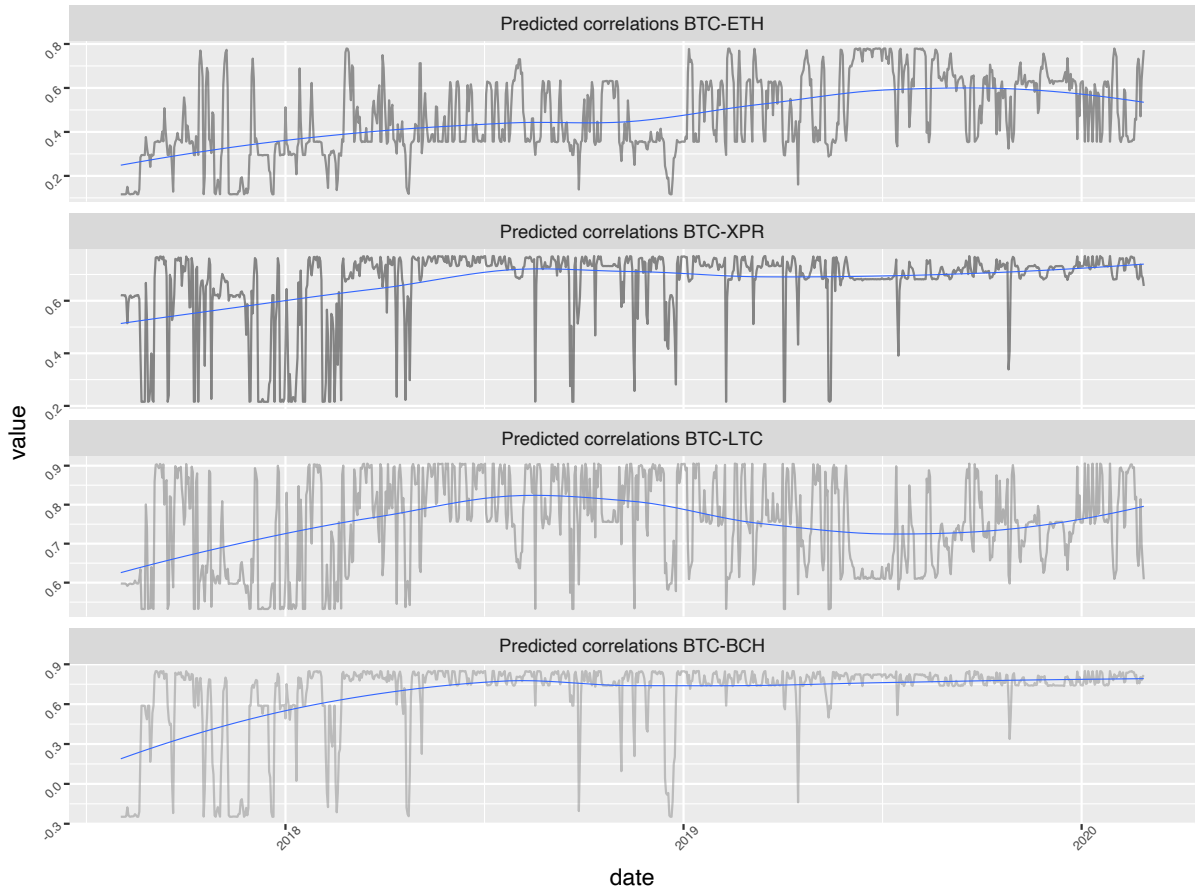


Figure 6: *Predicted correlations between BTC and the other cryptocurrencies under the HMM with $k = 5$ hidden states with overimposed smooth trend according to a local regression (blue line).*

According to the Bayesian Information Criterion, we select a model with five hidden states. Among them, states 2 and 3 describe more stable phases of the market that account for the 45% of the time, whereas state 1 represents a negative phase of the market featuring negative log-returns and high volatility, and states 4 and 5 are related to phases of a marked rise in price, and represent only the 8.41% and 6.71% of the overall time period.

From the estimated posterior probabilities and from the decoded states we can infer a trend characterized by a prevalence towards phases of greater stability detected by states 2 and 3, and an evident reduction of episodes of marked price increase detected by states 4 and 5. We show that the model is also able to provide quite remarkable

univariate predictions of log-returns and volatility for the future time occasions. Finally, we spot a trend of increase of the market correlation from the predicted correlations of the cryptocurrencies coupled to Bitcoin, coherent with the hypothesis of an increasing systematic risk observed in more mature markets, but also with the induced stronger link with the USD starting from 2018 due to the rise of stablecoins.

Appendix: additional figures

Figure 7: *Observed BTC log-returns (pink), predicted averages (green), and predicted standard deviations (blue) under the HMM with* $k = 5$ *hidden states.*

Figure 8: *Observed ETH log-returns (pink), predicted averages (green), and predicted standard deviations (blue) under the HMM with $k = 5$ hidden states.*

Figure 9: *Observed XPR log-returns (pink), predicted averages (green), and predicted standard deviations (blue) under the HMM with $k = 5$ hidden states.*

Figure 10: *Observed LTC log-returns (pink), predicted averages (green), and predicted standard deviations (blue) under the HMM with $k = 5$ hidden states.*

Figure 11: *Observed BCH log-returns (pink), predicted averages (green), and predicted standard deviations (blue) under the HMM with $k = 5$ hidden states.*

References

Agosto, A. and Cafferata, A. (2020). Financial bubbles: A study of co-explosivity in the cryptocurrency market. *Risks*, 8:34.

Akaike, H. (1998). Markovian representation of stochastic processes and its application to the analysis of autoregressive moving average processes. In *Selected Papers of Hirotugu Akaike*, pages 223–247. Springer.

Ang, A. and Bekaert, G. (2002). International asset allocation with regime shifts. *The Review of Financial Studies*, 15:1137–1187.

Bartolucci, F., Farcomeni, A., and Pennoni, F. (2013). *Latent Markov Models for Longitudinal Data*. Chapman & Hall/CRC Press, Boca Raton, FL.

Bartolucci, F., Pandolfi, S., and Pennoni, F. (2017). `LMest`: An `R` package for latent Markov models for longitudinal categorical data. *Journal of Statistical Software*, 81:1–38.

Bartolucci, F., Pandolfi, S., Pennoni, F., Farcomeni, A., and Serafini, A. (2020). *LMest: Generalized Latent Markov Models*. `R` package version 3.0.0.

Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics*, 41:164–171.

Borri, N. (2019). Conditional tail-risk in cryptocurrency markets. *Journal of Empirical Finance*, 50:1–19.

Bouri, E., Shahzad, S. J. H., and Roubaud, D. (2019). Co-explosivity in the cryptocurrency market. *Finance Research Letters*, 29:178–183.

Cappé, O., Moulines, E., and Rydén, T. (1989). *Inference in Hidden Markov Models*. Springer-Verlag, New York.

Chen, Y., Giudici, P., Hadji Misheva, B., and Trimborn, S. (2020). Lead behaviour in bitcoin markets. *Risks*, 8:4.

Corbet, S., Meegan, A., Larkin, C., Lucey, B., and Yarovaya, L. (2018). Exploring the dynamic relationships between cryptocurrencies and other financial assets. *Economics Letters*, 165:28–34.

De Angelis, L. and Paas, L. J. (2013). A dynamic analysis of stock markets using a hidden Markov model. *Journal of Applied Statistics*, 40:1682–1700.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, 39:1–38.

Genon-Catalot, V., Jeantheau, T., Larédo, C., et al. (2000). Stochastic volatility models as hidden Markov models and statistical applications. *Bernoulli*, 6:1051–1079.

Giudici, P. and Abu Hashish, I. (2020). A hidden Markov model to detect regime changes in cryptoasset markets. *Quality and Reliability Engineering International*, 36:2057–2065.

Giudici, P. and Pagnottoni, P. (2019). High frequency price change spillovers in Bitcoin markets. *Risks*, 7:111.

Giudici, P. and Pagnottoni, P. (2020). Vector error correction models to measure connectedness of Bitcoin exchange markets. *Applied Stochastic Models in Business and Industry*, 36:95–109.

Giudici, P. and Polinesi, G. (2019). Crypto price discovery through correlation networks. *Annals of Operations Research*, pages 1–15.

Hamilton, J. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57:357–384.

Huang, J.-Z., Huang, W., and Ni, J. (2019). Predicting bitcoin returns using high-dimensional technical indicators. *The Journal of Finance and Data Science*, 5:140–155.

Juang, B. H. and Rabiner, L. R. (1991). Hidden Markov models for speech recognition. *Technometrics*, 33:251–272.

Langrock, R., MacDonald, I. L., and Zucchini, W. (2012). Some nonstandard stochastic volatility models and their estimation using structured hidden Markov models. *Journal of Empirical Finance*, 19:147–161.

Lin, Y., Xiao, Y., and Li, F. (2020). Forecasting crude oil price volatility via a HM-EGARCH model. *Energy Economics*, pages 104–693.

Mamon, R. S. and Elliott, R. J. (2007). *Hidden Markov models in finance*, volume 460. Springer.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. Wiley.

Rossi, A. and Gallo, G. M. (2006). Volatility estimation via hidden Markov models. *Journal of Empirical Finance*, 13:203–230.

Satoshi, N. (2008). Bitcoin: A peer-to-peer electronic cash system. *Consulted*, 1:28.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6:461–464.

Trimborn, S. and Härdle, W. K. (2018). Crix an index for cryptocurrencies. *Journal of Empirical Finance*, 49:107–122.

Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE transactions on Information Theory*, 13:260–269.

Wang, S. and Vergne, J.-P. (2017). Buzz factor or innovation potential: What explains cryptocurrencies'returns? *PlusOne*, 12:e0169556.

Welch, L. R. (2003). Hidden Markov models and the Baum-Welch algorithm. *IEEE Information Theory Society Newsletter*, 53:1–13.

Yi, S., Xu, Z., and Wang, G.-J. (2018). Volatility connectedness in the cryptocurrency market: Is bitcoin a dominant cryptocurrency? *International Review of Financial Analysis*, 60:98–114.

Zucchini, W., MacDonald, I. L., and Langrock, R. (2017). *Hidden Markov Models for time series: an introduction using* `R`. Springer-Verlag, New York.