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

Fulvia Pennoni†

*Bartolucci F.**, *Forte G.*** *and Ametrano F.***

† *Department of Statistics and Quantitative Methods*
University of Milano-Bicocca
Email: fulvia.pennoni@unimib.it

* *University of Perugia*, ** *University of Milano-Bicocca*

Outline

- Introduction
- Multivariate hidden Markov model
- Maximum likelihood estimation
- Application to the market of five cryptocurrencies: Bitcoin (BTC), Ethereum (ETH), Ripple (XRP), Litecoin (LTC), and Bitcoin Cash (BCH)
- Conclusions

Introduction

- We propose a statistical and an unsupervised machine learning based on a multivariate Hidden Markov model (HMM) to jointly analyse financial asset price series of the major cryptocurrencies
- HMM provides a flexible framework for many financial applications and it allows us to incorporate stochastic volatility in a rather simple form
- With respect to the regime-switching models the HMM estimate state-specific expected log-returns along with state volatility
- We aim to estimate and predict volatility considering the expected log-returns as unpredictable parameters by considering the conditional means of the time-series

Introduction

- We model the log-returns of crypto-assets taking into account their correlation structure
- We assume that the daily log-return of each cryptocurrency is generated by a specific probabilistic distribution associated to the hidden state
- The evaluation of the conditional means improve the time-series classification: stable periods, crises, and financial bubbles differ significantly for mean returns and structural levels of covariance

Proposed Hidden Markov Model (HMM)

- We denote by:

  $\boldsymbol{y}_t$ the random vector at time t where each element y_{tj}, $j = 1, \ldots, r$, corresponds to the log-return of asset j

- We assume that the random vectors $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots$ are conditionally independent given a hidden process

- The hidden process is denoted as $u_1, u_2, \ldots$

- We assume that it follows a Markov chain with a finite number of hidden states labelled from 1 to k

Proposed HMM

- We model the conditional distribution of every vector $\boldsymbol{y}_t$ given the underlying latent variable u_t by a multivariate Gaussian distribution that is

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

where $\boldsymbol{\mu}_u$ and Σ_u are, for hidden state u, the specific mean vector and variance-covariance matrix (heteroschedastic model)

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

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

where, in general, $\phi(\cdot; \cdot, \cdot)$ denotes the density of the multivariate Gaussian distribution of dimension r

Proposed HMM

- The parameterization of the distribution of the structural model of the latent Markov process is based on:

- The initial probability defined as:

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

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

- The transition probability defined as:

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

collected in the transition matrix:

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

Maximum likelihood estimation

- The log-likelihood function for $\boldsymbol{\theta}$ vector of all model parameters is defined as

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

- The complete-data log-likelihood is defined as

$$\begin{aligned}
\ell_1^*(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k, \Sigma_1, \ldots, \Sigma_k) &= \sum_t \sum_u w_{tu} \log \phi(\boldsymbol{y}_t | \boldsymbol{\mu}_u, \Sigma_u) \\
&= -\frac{1}{2} \sum_t \sum_u w_{tu} [\log(|2\pi\Sigma_u|) + (\boldsymbol{y}_t - \boldsymbol{\mu}_u)' \Sigma_u^{-1} (\boldsymbol{y}_t - \boldsymbol{\mu}_u)], \\
\ell_2^*(\boldsymbol{\lambda}) &= \sum_u w_{1u} \log \pi_u, \\
\ell_3^*(\Pi) &= \sum_{t \geq 2} \sum_u \sum_v z_{tuv} \log \pi_{v|u},
\end{aligned}$$

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 0 otherwise, z_{tuv} denotes the transition in t from u to v

Maximum likelihood estimation

- Maximization of the log-likelihood is performed through the Expectation-Maximization algorithm (Baum *et al., 1970; Dempster* et al., 1977) which is based on two steps:

  - **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

  - **M-step**: it maximizes the expected complete data log-likelihood with respect to the model parameters.

    The parameters in the measurement model are updated in a simple way as:

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

for $u = 1, \ldots, k$,

Maximum likelihood estimation

- **M-step**:

The parameters in the structural model are updated as:

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

- The EM algorithm is initialized with an initial guess based on sample statistics; and different starting values are also generated randomly are employed to check for local maxima

Maximum likelihood estimation

- For model selection we rely on the Bayesian Information Criterion (BIC; Schwarz, 1978) which is based on the following index

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

where $\hat{\ell}_k$ denotes the maximum of the log-likelihood of the model with k states and $\#\mathrm{par}$ denotes the number of free parameters equal to $k[r + r(r+1)/2] + k^2 - 1$ for the heteroschedastic model

- We predict the most likely sequence of hidden states, through the so called local decoding or global decoding

Application

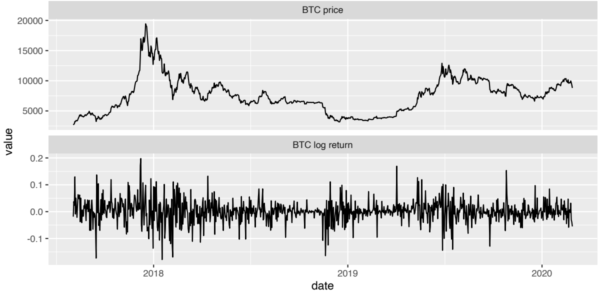
- The selection of the cryptocurrencies for the applicative example are the criteria underlying the Crypto Asset Lab Index (to be published in 2021):
  - more reliable
  - liquid
  - less manipulated crypto-assets in the market

- For the sake of comparability on the liquidity side, we consider a recent time span of three-years: from August 2, 2017, to February, 27, 2020

- Computational tools are implemented by adapting suitable functions of the `R` package `LMest` (Bartolucci *et al.*, 2017)

Application: data description

- We consider: Bitcoin, Ethereum, Ripple, Litecoin, and Bitcoin Cash
- We shows the BTC prices along with the daily log-returns for the whole period of observation



Application: data description

- Observed variance-covariance matrix:

	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

- Observed correlations and partial correlations:

	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

- The BTC dominance does not necessarily results in a unique co-moving driver

Results: model selection

- The order (number of states, k) of the hidden distribution is selected through the BIC
- The model selection strategy accounts for the multimodality of the likelihood function and the best model is the heteroschedastic HMM with $k = 5$ hidden states

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

Results: expected log-returns

- According to the estimated expected log-returns of each state there are tree negative (1,2,3) and two positive regimes (4,5)

	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

- They represent the occurrence of a variety of situations happening on the market

Results: expected log-returns

- States 2 and 3 identify more stable phases of the market, they account for the 45% of the time
- State 1 represents a negative phase of the market featuring negative log-returns
- 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

Results

- Conditional correlations (below the main diagonal), variances (in bold, pink), partial correlations (in italic above the main diagonal)

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**

Results: estimated conditional variances and correlations

- In state 2 the correlation between BTC and XRP is high (0.68) but the partial correlation is low and negative (-0.18).
- In terms of volatility, it is clear that state 3 is the most volatile
- Therefore states 1 and 3 are both marked by negative log-returns, but with very different levels of risk
- State 1 is the one characterized by significant falls of price and by a marked volatility

Results: transition probabilities

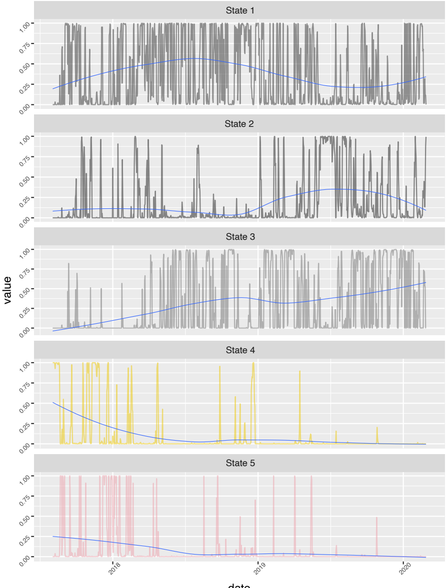
- The estimated matrix of the transition probabilities is the following

	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**

- States 2, 3, and 4 are the most persistent and 1 and 5 are less persistent

- The highest estimated transition from the less persistent state 5 to state 1 can be read as the typical pull back following a substantial price increase

Results: posterior probabilities



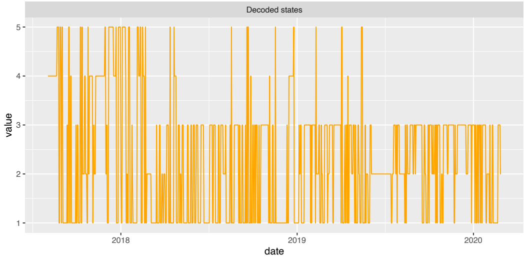
Results: posterior probabilities

- The trend line is overimposed according to a smoothed local regression

- We notice the increasing tendency for state 3 and a decreasing tendency of states 4 and 5 over time

- Apart for few exceptions there are not stable periods

Results: decoded states



- State 1 represents negative phases of the market and is visited the 36.85% of the overall period
- States 2 and 3 represent more stable phases of the market and are visited the 16.19%, and the 31.84% of the overall period
- States 4 and 5 related to phases of a market with textcolorbluerise in prices and are visited the 8.41% and the 6.71% of the overall period

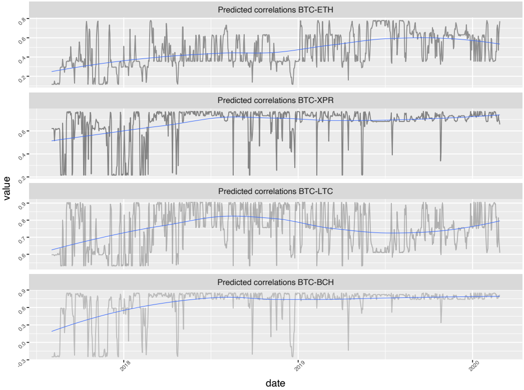
Results: predicted averages and standard deviations

- Observed XPR log-returns (pink), predicted averages (green), and predicted standard deviations (blue) under the HMM with $k = 5$ hidden states
- The model is able to timely detect regimes of high or low returns and volatilities

Results: Predicted averages and s.d.

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

Results: Predicted correlations



- Predicted correlations between BTC and the other cryptocurrencies of the HMM with $k = 5$ hidden states with overimposed smooth trend according to a local regression (blue line)
- Our results confirm a medium term trend of greater correlation relative to BTC with the other cryptocurrencies

Conclusions

- The advantage of employing an HMM traditional regime-switching models is that we estimate state-specific expected log-returns and state volatility
- We show that the model is also able to provide quite remarkable predictions of log-returns and volatility for the future time occasions
- 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

Main References

- 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.

- Dempster, A. P., Laird, N. M., Rubin, D.B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm (with discussion). *Journal of the Royal Statistical Society B*, 39: 1–38.

- 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.